Università degli Studi dell'Insubria
Dipartimento di Scienza e Alta Tecnologia

Dottorato di Ricerca in Informatica e Matematica del Calcolo

Spectral analysis and fast methods for structured matrix sequences and PDE discretizations

Ph.D. thesis of:

Isabella Furci

Advisor: Prof. Stefano Serra–Capizzano
Co-Advisor: Dr. Sven–Erik Ekström

XXXI Cycle

Accademic year 2017/2018

To Rosamaria,
my personal superhero

Abstract

When simulating phenomena in physics, engineering, or applied sciences, often one has to deal with functional equations that do not admit an analytical solution. Describing these real situations is, however, possible, resorting to one of its numerical approximations and treating the resulting mathematical representation. This thesis is placed in this context: Indeed the purpose is that of furnishing several useful tools to deal with some computational problems, stemming from discretization techniques. In most of the cases the numerical methods we analyse are the classical $\mathbb{Q}_p$ Lagrangian FEM and the more recent Galerkin B-spline *Isogeometric Analysis (IgA)* approximation and Staggered *Discontinuous Galerkin (DG)* methods. As our model PDE, we consider classical second-order elliptic differential equations and the Incompressible Navier-Stokes equations. In all these situations the resulting matrix sequences $\{A_n\}_n$ possess a structure, namely they belong to the class of Toeplitz matrix sequences or to the more general class of *Generalized Locally Toeplitz (GLT)* matrix sequences, in the most general block k-level case. Consequently, the spectral analysis of the coefficient matrices plays a crucial role for an efficient and fast resolution. Indeed the convergence properties of iterative methods proposed, like multigrid or preconditioned *Krylov* techniques, are strictly related to the notion of *symbol* of the coefficient matrix sequence. In our setting the symbol is a function which asymptotically provides a reasonable approximation of the eigenvalues [singular values] of A_n by its evaluations of an uniform grid on its domain. These reasons, and many others, make the research of more and more efficient eigensolvers relevant and topical. In this direction, the second goal of this thesis is to provide new tools for computing the spectrum of preconditioned banded symmetric Toeplitz matrices, Toeplitz-like matrices, $n^{-1}K_n^{[p]}$, $nM_n^{[p]}$, $n^{-2}L_n^{[p]}$, coming from the B-spline IgA approximation of $-u'' = \lambda u$, plus its multivariate counterpart for $-\Delta u = \lambda u$, and block and preconditioned block banded symmetric Toeplitz matrices. For all the above cases we propose new algorithms based on the classical concept of symbol, but with an innovative view on the errors of the approximation of eigenvalues by the uniform sampling of the symbol. The algorithms devised are special interpolation-extrapolation procedures performed with a high level of accuracy and only at the cost of computing of the eigenvalues of a moderate number of small sized matrices.

Key words:

Multilevel block GLT algebra, symbol, spectral distribution, asymptotic expansion, interpolation-extrapolation algorithms, multigrid methods, preconditioning Krylov methods, Staggered DG methods, IgA approximation

Contents

Introduction and motivation

The main mission of numerical analysts is to compute quantities that are in general incalculable from an analytical point of view. The pivotal concept in numerical analysis is analyzing and providing algorithms to solve a determined class of the problems of mathematics, whose intrinsic nature can be either continuous or discrete. "Continuous" are most of the real problems which science and engineering are built upon but that, without numerical techniques, would be quickly untreatable. In this thesis the focus is on fast algorithms for the approximation of continuous mathematical equations.

Indeed, when simulating phenomena in physics, engineering, or applied sciences, often one has to deal with functional equations (written, e.g., in differential or integral form) that do not admit an analytical solution.

Describing these real situations is, however, possible, resorting to one of its numerical approximations and treating the resulting mathematical representation. In practice the aim is to construct proper numerical discretization techniques, that "transform" problems from "continuous" to a more manageable "discrete" modelling.

Clearly the task of the numerical analyst does not end once that the approximation is performed. We want to ensure that the solution of the resulting problem, with respect to the original one, is more convenient in terms of resolution speed, resources, and computational cost.

This thesis is placed in this context: it has the purpose to furnish several useful tools to deal with some computational problems, arising from discretization techniques.

In most cases the problems we have in mind come from the linear discretization of partial differential equations (PDEs) of the form

$$\mathscr{A}u = b,$$

where $\mathscr{A}$ is a linear differential operator, taking into account possible initial/boundary conditions. Computing the numerical solution u_n of u, or a part of it, reduces to solving a linear system of the form

$$A_n \mathbf{u}_n = \mathbf{b}_n. \tag{1}$$

Furthermore, if the chosen approximation technique is convergent, the more we increase the number of points of the discretization (n or an increasing function of n) the more the approximation u_n of the analytical solution u will be accurate.

For this reason, one should not consider the specific linear system (1) for a fixed n, but rather

the sequence of linear systems

$$\{A_n \mathbf{u}_n = \mathbf{b}_n\}_n,$$

whose dimensions depend on the number of discretization points, n.

The matrices produced by most types of discretizations possess a structure, namely they are often sparse. Furthermore, depending on the linear differential operator, they can be badly conditioned. Consequently in general (that is without a quite strong structure), direct methods should be avoided, since, not only they may require a high computation cost, but also they often do not take full advantage of the information of the structure.

Iterative solvers (in particular multigrid and preconditioned Krylov techniques) are instead very convenient choices. It is indeed known that iterative methods exploit the spectral information of coefficient matrix and consequently they can be adapted in order to accelerate the convergence and optimize the computational cost.

Hence here the spectral analysis of the matrix A_n (and consequently of the coefficients matrix sequence $\{A_n\}_n$) plays a crucial role for an efficient and fast resolution. Moreover, comparing the spectrum of A_n with that of the differential operator can suggest whether the discretization is appropriate or not to spectrally approximate the operator $\mathscr{A}$.

However, it must be highlighted that the interest in finding eigenvalues is intrinsically important. In fact, on one hand there are problems in which the knowledge of the eigenvalues is indirectly useful in efficiently finding the solution. On the other hand there are situations where they actually have a physical meaning and represent the approximation of the real solution. This is the case, for example, of eigenvalue problems [42, 92].

Among specific applications that are not related to the approximation of differential equations, we can mention structured Markov chains [15], signal and image processing problems with space invariant nature [46, 82], financial applications [110], etc.

The sequences considered in the whole thesis enjoy a very nice structure: they belong to the class of *Toeplitz* matrix sequences or to the more general class of *Generalized Locally Toeplitz (GLT)* matrix sequences.

In general, depending on whether the matrices come from a one-dimensional or a k-dimensional problem, $k > 1$, their structure can be one-level or k-level. That is each matrix has a scheme repeated k times equally in the inner patterns. In such a case the dimension of the matrix is $N(\mathbf{n}) = n_1 n_2 \cdots n_k$ and the matrix is indexed by the multi-index $\mathbf{n} = (n_1, n_2, \ldots, n_k)$. For the multi-index notation, see Section I.2. Depending on the size s of the system of PDEs, we deal with a scalar ($s = 1$) or a block ($s > 1$) matrix sequence. In the latter setting each basic entry in the matrix A_n is in turn an $s \times s$ matrix, so that the global dimension is $sn \times sn$ or $N(\mathbf{n}, s) \times N(\mathbf{n}, s)$, with $N(\mathbf{n}, s) = sN(\mathbf{n}) = sn_1 n_2 \cdots n_k$, $\mathbf{n} = (n_1, n_2, \ldots, n_k)$.

However, even in the case of a scalar PDE, the block structure can be induced by the numerical method, e.g., by classical p-degree finite elements, $p > 1$, or p-degree Discontinuous Galerkin mathods, $p \geq 1$, or p-degree isogeometric analysis of regularity $\Bbbk$ with $p - \Bbbk > 1$.

In all situations the research of spectral informations of the mentioned classes is related to the concept of the *symbol*, that is a function f which, under certain hypotheses, provides a spectral or a singular value description of the associated matrix sequences.

In the simplest scalar, one-level case, where the only requirement on $f : D \subset \mathbb{R} \to \mathbb{C}$ is to be a Lebesgue measurable function on a Lebesgue measurable domain D, with Lebesgue measure

$0 < \mu_1(D) < \infty$, we say that the sequence $\{A_n\}_n$ has an asymptotical spectral [singular value] distribution described by f if it holds that:

$$\lim_{n\to\infty} \frac{1}{n} \sum_{j=1}^{n} F(\lambda_j(A_n)) = \frac{1}{\mu_1(D)} \int_D F(f(\theta))\, \mathrm{d}\theta, \tag{2}$$

$$\left[\lim_{n\to\infty} \frac{1}{n} \sum_{j=1}^{n} F(\sigma_j(A_n)) = \frac{1}{\mu_1(D)} \int_D F(|f(\theta)|)\, \mathrm{d}\theta, \right]$$

for all continuous functions F with bounded support on $\mathbb{C}$, where $\lambda_j(A_n)$, $j = 1, \ldots, n$ [$\sigma_j(A_n)$, $j = 1, \ldots, n$] are the eigenvalues [singular values] of A_n.

The informal (and practical usable) meaning of relation (2) is that for n sufficiently large, a reasonable approximation of the eigenvalues [singular values] of A_n is obtained from an evaluation of $f(\theta)$ [$|f(\theta)|$] over an uniform grid in the domain D. Once the symbol is known we have the "control" of the behaviour of the whole spectrum [singular values], up to a number of outliers which is infinitesimal with respect to the matrix size, and we can exploit the results for designing efficient solvers for the coefficient matrix A_n, for large n.

Along the same lines the k-level block case with blocks of size s can be given by playing with the symbol, which will be k-variate and $s \times s$ matrix-valued. For such general notion see Section I.3 (and Section I.2 for the necessary multi-index notation).

Generally speaking all the concepts, notations and mathematical tools which will be used in the thesis are reported in **Chapter I**.

In the next chapters we face several type of sequence structures: from the most general block k-level setting to the simplest scalar, one-level case. Clearly formula (2) is properly modified for the more general types of treated structured sequences. Indeed, with the obvious changes of notation, the universal role of the symbol is being one of the tool for compactly describing the asymptotic behavior of the eigenvalues [singular values] of A_n, for large n.

In **Chapter II** we consider the GLT sequence arising from the approximation of the incompressible Navier-Stokes equations by semi-implicit Discontinous Galerkin methods on *staggered meshes* (SDG), introduced in [65, 66, 135, 137].

These new schemes have never been analyzed with GLT techniques before. Therefore the first aim is theoretical and concerns the possibility of using and extending the spectral tools mentioned so far to this new numerical framework and of studying its properties. Special attention is given to the structural and spectral analysis of the involved linear systems, in particular: structural properties, in connection with multilevel block Toeplitz-like (and circulant) matrices, distribution spectral analysis in the Weyl sense, conditioning, asymptotic behaviour of the extremal eigenvalues via low rank perturbations and study of outliers. In turn all of them are of interest for numeric and algorithmic purposes: the analysis of the intrinsic difficulty of the problem aimed at designing and analyzing (preconditioned) Krylov methods [5, 11].

First we follow a classical preconditioning strategy, designing a Preconditioned Conjugate Gradient (PCG) method, with circulant Strang preconditioner. But in a multilevel setting, when an asymptotic condition is present, as it is well-known in the literature [98, 124, 129],

the use of any circulant preconditioner permits to reach at most a sub-optimal convergence of the PCG method. So a more original and efficient approach is to design a block multilevel multigrid procedure with two grids (TGM), that will ensure convergence and optimality in terms of iterations. Although, up to our knowledge, many theoretical results on multigrid convergence in block settings are still missing or in preparation [130], we validate them numerically. The choice of the appropriate smoother and prolongation operators are justified by the Laplacian nature of the symbol and supported by the encouraging results in a block context in [48], and in the scalar multilevel cases [2, 3, 69].

Moreover, based on the concept of symbol, it has been possible to design specific fast methods for solving large linear systems with a Toeplitz or Toeplitz-like structure in various settings. When speaking about Toeplitz-like matrices, we are referring to small perturbations of Toeplitz matrices or block Toeplitz matrices, where the precise structure is observed when removing few rows and columns.

Because of their pervasive appearance in any shift-invariant problem [7, 32, 93], there has been a lot of attention on fast methods for solving large linear systems with Toeplitz or Toeplitz-like structure (see the review papers in [32, 93, 94]), including both direct fast and superfast solvers [93, 94] and iterative solvers [3, 32].

Here we consider the problem of computing the spectrum and, for such type of problems, we develope a class of fast methods starting from the results in recent work [62], where Ekström, Garoni, and Serra-Capizzano have conjectured the existence of an asymptotic expansion for the eigenvalues of banded symmetric Toeplitz matrices. Independently Bogoya, Böttcher, Grudsky, and Maximenko [16, 17, 19] have obtained the precise asymptotic expansion for the eigenvalues of a sequence of Toeplitz matrices $\{T_n(f)\}_n$, under suitable assumptions on the associated generating function f.

In [62] the authors provided numerical evidences that some of those assumptions can be relaxed, maintaining only the hypothesis on f of being a real cosine trigonometric polynomial **(RCTP)**, monotone on the domain.

Studying the errors of the approximation of eigenvalues by uniform sampling of the symbol, we devise an extrapolation procedure for computing the eigenvalues of banded symmetric Toeplitz matrices of very large dimension. The algorithm is performed with a high level of accuracy and only at the cost of computing the eigenvalues of a moderate number of small sized matrices.

From a theoretical viewpoint, in **Chapters III, IV, V** the assumptions on the generating function have been relaxed and extended also for the eigenvalues of:

1. preconditioned banded symmetric Toeplitz matrices [1];

2. Toeplitz-like matrices, $n^{-1}K_n^{[p]}$, $nM_n^{[p]}$, $n^{-2}L_n^{[p]}$, coming from the B-spline IgA approximation of $-u'' = \lambda u$, plus its multivariate counterpart for $-\Delta u = \lambda u$ [58];

3. block and preconditioned block banded symmetric Toeplitz matrices [60].

We also prove, for all contexts above, the first order asymptotic term of the expansion and we complement the results of [51, 71, 72, 73, 74, 76, 77], proving several important analytic properties of $e_p(\theta)$, spectral symbol of $\{n^{-2}L_n^{[p]}\}_n$.

For Item 3 we consider the natural extension of the analysis for the case of $\mathbf{f}$ being an $s \times s$ matrix-valued function with $s \geq 1$, and $T_n(\mathbf{f})$ the block Toeplitz matrix generated by $\mathbf{f}$. Hence the natural step is that of deriving the analogous conditions which ensure the existence of an asymptotic expansion for the eigenvalues in block settings. In particular how the assumptions on the scalar symbol f of being a real, monotone, cosine trigonometric polynomial are transformed for the matrix-valued symbol $\mathbf{f}$. Here the eigenvalue functions of $\mathbf{f}$, $\lambda^{(i)}(\mathbf{f})$, $i = 1, \ldots, s$, play an analogous role of f for the scalar cases. Furthermore we deal with the conversion from polynomial **(RCTP)** to Hermitian matrix-valued trigonometric polynomial **(HTP)**.

The hidden idea for the considered asymptotic expansion is based on the right reordering of eigenvalues with respect to the evaluations of f (of $\lambda^{(i)}(\mathbf{f})$, $i = 1, \ldots, s$, in case $s > 1$). Indeed for $s = 1$, the assumption of monotonicity of f is crucial to ensure the correct combination of eigenvalues and evaluations. Analogously, for $s > 1$, the right reordering and the validity of expansion are guaranteed *globally* on the spectrum, requiring the monotonicity of every eigenvalue functions and the empty intersection of the ranges two eigenvalue function $\lambda^{(j)}(\mathbf{f})$ and $\lambda^{(k)}(\mathbf{f})$, for every pair of indices $j, k \in \{1, \ldots, s\}$ such that $j \neq k$. If the global condition is violated, it is, however, possible to recover the asymptotic expansion for the portion of spectrum associated to those eigenvalue functions which verify *locally* both the non-intersection and monotonicity conditions.

The asymptotic spectral expansion becomes a potential tool for the computation of the spectrum of differential operators. In **Chapter IV** we perform a detailed spectral analysis of the matrices $n^{-1}K_n^{[p]}, nM_n^{[p]}, n^{-2}L_n^{[p]}$.

In particular for $p \geq 3$, we provide numerical evidence of a precise asymptotic expansion for the eigenvalues, except for the largest $n_p^{\text{out}} = n - \text{mod}(p, 2)$ outliers, of $n^{-2}L_n^{[p]}$.

In addition, for $p = 1$ and $p = 2$, we compute the exact eigenvalues and eigenvectors of $K_n^{[p]}$, $M_n^{[p]}$, and $L_n^{[p]}$. In both cases of p, the eigenvalues are given respectively by $f_p(\theta_{j,n})$, $g_p(\theta_{j,n})$, and $e_p(\theta_{j,n})$, for $j = 1, \ldots, n+p-2$, $\theta_{j,n} = j\pi/n$, where $f_p(\theta)$, $g_p(\theta)$, and $e_p(\theta)$ are the functions that spectrally describe the sequences $\{n^{-1}K_n^{[p]}\}_n$, $\{nM_n^{[p]}\}_n$, and $\{n^{-2}L_n^{[p]}\}_n$, respectively [77, Section 10.7]. The exact computation is made possible since the matrices $K_n^{[p]}$, $M_n^{[p]}$, $L_n^{[p]}$ belong to the same matrix algebra. By using tensor-product arguments we can also present a detailed extension of the whole analysis to the general k-dimensional setting.

We show indeed that the eigenvalue–eigenvector structure of the matrix arising from the IgA approximation of the 1D problem

$$\begin{cases} -u''(x) = \lambda u(x), & x \in (0, 1), \\ u(0) = u(1) = 0, \end{cases} \tag{3}$$

completely determines the eigenvalue–eigenvector structure of the matrix $L_n^{[p]}$ in the k-dimensional setting.

The exact formulae for the eigenvalues are also presented in **Chapter V** for the scaled matrix sequences, $\{M_n^{(p)}\}_n$, $\{K_n^{(p)}\}_n$ and $\{L_n^{(p)}\}_n = \{(M_n^{(p)})^{-1}K_n^{(p)}\}_n$, coming from order p Lagrangian Finite Element approximations of a second order elliptic differential problem. The algorithm that exactly computes the spectrum of the mass $M_n^{(p)}$, stiffness $K_n^{(p)}$ and $L_n^{(p)}$ is based on a proper evaluation of the spectral symbols $\mathbf{g}$, $\mathbf{f}$ and $\mathbf{r}$ on the correct grid.

Chapters III, IV, V are completed from the computational viewpoint by delivering fast (and parallel) interpolation–extrapolation algorithms for computing the eigenvalues of Items 1, 2, 3.

In all the treated cases the resulting algorithms can be interpretated as eigensolvers that do not need to store either the coefficients of the matrices or perform matrix-vector products, and for this reason they have been recently defined *matrix-less* solvers [57].

We present and critically analyze many numerical examples. On one hand this has the purpose to validate and numerically confirm the proposed theoretical and algorithmic results. On the other hand we show how to manipulate many examples of practical interest. For instance we show how to bypass the monotone condition in few special cases and how to reduce a block problem to few, separate, and simpler scalar problems.

The last sections will be dedicated to illustrate few topics for future research related to the themes of the present thesis. The plan in mind is that of continue providing and analyzing methods in order to deal with the most general classes of structured matrix sequences and PDE discretizations.

Concerning this direction the first step regards a feasible extension of the proposed matrix-less eigensolvers to multilevel contexts, in cases where a tensor product argument cannot be exploited. Here the principal open question concerns the formalization, in both scalar or block case, of the asymptotic spectral expansion for k-level matrices, that in turn depends on the lack of the monotonicity concept for a k variate symbol.

In the following we briefly describe the contents of the upcoming **Chapters I-V** and of the **Chapter VI** of the technical results.

- In **Chapter I** we set the notation used throughout the thesis and we provide the fundamental background that is necessary for understanding the subsequent chapters.

  In particular: the definitions and the main properties of Toeplitz, circulant and GLT sequences in the most general block k-level form, the notion of spectral [singular value] distribution, and the preliminary version of the asymptotic spectral expansion. Moreover we briefly recall the basic ideas which represent the minimal tools for understanding the multigrid and preconditioned conjugate gradient methods.

- In **Chapter II** we are interested in efficiently solving the large linear systems arising from the discretization of the two–dimensional incompressible Navier-Stokes equations by Discontinuous Galerkin methods on staggered meshes. These novel family of high order semi-implicit schemes are analyzed for the first time with GLT techniques. We show that the coefficient matrix sequence has a multilevel block Toeplitz structure plus a low rank corrections. The results are then used for deducing spectral informations on outliers, conditioning and asymptotic behaviour of the extremal eigenvalues. In turn all of them are of interest for numeric and algorithmic purposes: making use of the resulting asymptotic spectral information, we design specific preconditioned Krylov and two grids method for the efficient resolution of the associated linear systems. We obtain that the use of PCG method with circulant Strang preconditioner cannot ensure the superlinear convergence,

as it is known from [98, 124, 129], conversely an efficient approach is represented by the block multilevel two grids procedure, that guarantees convergence and optimality in terms of iterations and global efficiency.

- **Chapter III** is devoted to present the asymptotic spectral expansion for the eigenvalues of preconditioned Toeplitz matrices $\mathcal{P}_n(f,g) = T_n^{-1}(g)T_n(f)$. We consider the case where f is a trigonometric polynomial, g is a nonnegative and not identically zero trigonometric polynomial. We provide numerical evidence that few assumptions of [16, 17, 19] can be relaxed, accompanied by an appropriate error analysis and numerical experiments. Moreover we devise an algorithm that compute an accurate approximation of the eigenvalues of $\mathcal{P}_n(f,g)$ for very large n, having the eigenvalues of $\mathcal{P}_{n_i}(f,g)$, for moderate values of n_i, $i = 1,\dots,\alpha$, where α is a fixed small number.

- in **Chapter IV** we consider the B-spline IgA approximation of the Laplacian eigenvalue problem $-\Delta u = \lambda u$ over the k-dimensional hypercube $(0,1)^k$. We provide the exact eigenvalue–eigenvector structure of the resulting discretization matrices $L_n^{[p]}$, $L_n^{[p]}$, and $L_n^{[p]}$, for $p = 1, 2$.

  For $p \geq 3$, based on the asymptotic spectral expansion, we propose a parallel interpolation–extrapolation algorithm for computing the eigenvalues of $L_n^{[p]}$, excluding the largest $n_p^{\text{out}} = p-2+\text{mod}(p,2)$ outliers. The performance of the algorithm is illustrated through numerical experiments. We end the chapter with a detailed extension of the whole analysis to the general k-dimensional setting. By using tensor-product arguments, we show that the eigenvalue–eigenvector structure of the matrix arising from the IgA approximation in one dimension is enough to cover also the multidimensional case.

- In **Chapter V** we focus on the generalization of the results of **Chapters III-IV** under the assumptions that $\mathbf{f}$ is a $s \times s$ matrix-valued trigonometric polynomial with $s \geq 1$, and $T_n(\mathbf{f})$ is the associated block Toeplitz matrix, whose size is $N(n,s) = sn$.

  First we numerically derive conditions which ensure the existence of an asymptotic expansion for the eigenvalues, generalizing those for the scalar-valued setting $s = 1$. Furthermore, following the proposal for $s = 1$ in the previous chapters, we devise an interpolation–extrapolation algorithm for computing the eigenvalues of banded symmetric block Toeplitz matrices, with a high level of accuracy and a low computational cost, and we present several examples of practical interest. Furthermore we provide exact formulae for the eigenvalues of the matrices coming from the $\mathbb{Q}_p$ Lagrangian Finite Element approximation of a second order elliptic differential problem and the preconditioned block matrices coming from the classical Lagrangian Finite Element approximation of the classical eigenvalue problem for the Laplacian operator in one dimension.

- The **Chapter VI** contains the additional theoretical results and specifications related to the contents in the thesis: it is divided in seven sections where different topics are treated. The choice of collecting them together in the end, instead near the respective chapters, is made in order to make the text more readable, without the interruption, e.g., of the long derivations represented by the proof of the theorems.

The results will be anyway referred through the thesis. Among them we present the proof of the first order asymptotic term of the expansion for the three Items in 1, 2, 3. We show several theoretical results regarding $e_p(\theta)$, the symbol of the normalized sequence $\{n^{-2}L_n^{[p]}\}_n$ of **Chapter IV**, such as the proof of a convergence result and of its monotone increasing behaviour. As observed before, in connection with **Chapters III, IV, V**, the monotonicity of the symbol is crucial for the asymptotic eigenvalue expansions and for the proper efficient behavior of our algorithms.

All our principal findings are summarized in the conclusion chapter.

The results of our research have been published or are in the process of publication in [1, 55, 58, 59, 60, 67].

We stress that the **Chapters II, III, IV, V** faithfully report the contents of the papers [1, 55, 58, 60] respectively. However, in order to avoid possible repetitions and to make the readability of the whole thesis as fluent as possible, in the next chapters some minimal changes are performed with respect to [1, 55, 58, 60]. For example, we use the unified notation stemming from **Chapter I**, the order of the sections is sometimes inverted, and we introduce additional observations and examples, which are not present in the papers, but which help to illustrate and explain better the treated topics.

Chapter I

Definitions and known results

The following chapter is devoted to set the notation and to introduce the definitions and few known results adopted throughout all chapters. In particular, after basic notions of numerical linear algebra, we present the multi-index notation, that will be largely used throughout the whole thesis. Moreover we provide the formulation of the most general multilevel block form of Toeplitz and circulant matrices and their main algebraic, structural, and spectral properties.

We introduce the concept of spectral/singular value distribution, we show key results on the extremal eigenvalues of Hermitian Block Toeplitz sequences, and we briefly describe the main properties of the GLT class, which can be seen as a variable coefficient generalization of the Toeplitz notion.

Starting from basic features of trigonometric polynomials, we focus our attention on the special case of Toeplitz matrices having a trigonometric polynomial as generating function. In particular, a preliminary asymptotic expansion in terms of the fineness parameter related to the matrix size is also presented for the scalar non-preconditioned case.

The chapter ends recalling advanced methods for solving linear systems, including preconditioning strategies to be used in connection, e.g., with Krylov solvers and multigrid methods tailored for Toeplitz structures.

I.1 General notation

- $\mathbb{R}^{m\times n}$ ($\mathbb{C}^{m\times n}$) is the space of real (complex) $m \times n$ matrices.
- If $x = [x_1, \ldots, x_n] \in \mathbb{C}^n$ is a vector,
  - x^T denote the transpose of x;
  - x^* denote the conjugate transpose of x;
- If $A = [a_{ij}]_{i,j=1}^n \in \mathbb{C}^{n\times n}$,
  - A^T denote the transpose of A;
  - A^* denote the conjugate transpose of A;
  - $\text{rank}(A)$ is the rank of A;
  - $\det(A)$ is the determinant of A;

- $\lambda_j(A)$, $j = 1, \dots, n$, ($\sigma_j(A)$, $j = 1, \dots, n$) are the eigenvalues [singular values] of A; If not specified differently, we assume $\sigma_1(A) \leq \sigma_2(A) \leq \cdots \leq \sigma_n(A)$;
- $\Lambda(A) = \{\lambda_1(A), \dots, \lambda_n(A)\}$ is the spectrum of A;
- Given $1 \leq p \leq \infty$, $\|A\|_p$ denotes the Schatten p-norm of A, which is defined as the p-norm of the vector of the singular values $[\sigma_1(A), \dots, \sigma_n(A)]$. The Schatten 1-norm is also called the trace-norm and the Schatten ∞-norm $\|A\|_\infty = \sigma_n(A)$ is the classical induced Euclidean norm (or spectral norm) and it is also denoted by $\|A\|$.
- $\kappa(A)$ is the condition number of an invertible matrix A defined as the quantity

$$\kappa(A) = \|A\|\|A^{-1}\| \quad (\geq \|AA^{-1}\| = 1).$$

  In particular, if in addition A is normal,

$$\kappa(A) = \|A\|\|A^{-1}\| = \frac{\sigma_{\max}(A)}{\sigma_{\min}(A)} = \frac{\max_j |\lambda_j(A)|}{\min_j |\lambda_j(A)|}.$$

- If A, $B \in \mathbb{C}^{n\times n}$,
  - $A \sim B$ means that A is similar to B, that is there exists an invertible matrix P such that $B = P^{-1}AP$;
  - $A \geq B$ if A and B are Hermitian and $A - B$ is Hermitian Positive SemiDefinite (HPSD);
  - $A > B$ if A and B are Hermitian and $A - B$ is Hermitian Positive Definite (HPD);
- if $A \in \mathbb{C}^{n\times n}$ is HPD, $\|\cdot\|_A = \|A^{1/2}\cdot\|_2$ denotes the Euclidean norm weighted by A on $\mathbb{C}^n$ and the associated induced matrix norm.
- O_m and I_m are the $m \times m$ zero matrix and identity matrix, respectively.
- $\mathbf{e}_i$ denote the ith vector of the canonical basis of $\mathbb{R}^k$.
- μ_k denotes the Lebesgue measure in $\mathbb{R}^k$.
- ι is the imaginary unit, that is $\iota^2 = -1$.
- If A, B are matrices of any size, say $A \in \mathbb{C}^{m_1\times m_2}$ and $B \in \mathbb{C}^{l_1\times l_2}$, the tensor (Kronecker) product of A and B is the $m_1l_1 \times m_2l_2$ matrix defined by

$$A \otimes B = [a_{ij}B]_{i=1,\dots,m_1\,,\,j=1,\dots,m_2} = \begin{bmatrix} a_{11}B & \dots & a_{1m_2}B \\ a_{21}B & \dots & a_{2m_2}B \\ \vdots & \ddots & \vdots \\ a_{m_11}B & \dots & a_{m_1m_2}B \end{bmatrix}.$$

- If D is a measurable subset of $\mathbb{R}^k$, we define
  - $L^p(D)$ the space of measurable functions $f : D \to \mathbb{C}$ such that

$$\int_D |f|^p < \infty, \quad 1 \leq p < \infty;$$

– $L^\infty(D)$ the space of measurable functions $f : D \to \mathbb{C}$ such that

$$\operatorname{ess\,sup}_D |f| < \infty.$$

- Given $f \in L^p(D)$, we write $\|f\|_p$ to indicate the L^p-norm of f, that is

$$\|f\|_p = \begin{cases} \left(\int_D |f|^p\right)^{1/p}, & \text{if} \quad 1 \le p < \infty, \\ \operatorname{ess\,sup}_D |f|, & \text{if} \quad p = \infty. \end{cases}$$

- We denote by $\mathcal{I}_k$ the k-dimensional cube $[-\pi, \pi]^k$ and by $L^p(k, s)$ the linear space of k-variate matrix-valued functions $\mathbf{f} : \mathcal{I}_k \to \mathbb{C}^{s\times s}$, $\mathbf{f} \in L^p(\mathcal{I}_k)$. We remark that a matrix-valued function $\mathbf{f}$ belongs to $L^p(D)$ (resp. is measurable, continuous, bounded, etc.) if all its components $f_{ij} : D \to \mathbb{C}$, $i, j = 1, \ldots, s$, belong to $L^p(C)$ (resp. are measurable, continuous, bounded, etc.).

- Given a function $\mathbf{f} \in L^p(k, s)$, we define

$$\|\mathbf{f}\|_p = \begin{cases} \left(\int_{-\pi}^{\pi} \|\mathbf{f}(\mathbf{x})\|_p^p \, d\mathbf{x}\right)^{1/p}, & \text{if} \quad 1 \le p < \infty, \\ \operatorname{ess\,sup}_{\mathbf{x}\in[-\pi,\pi]} \|\mathbf{f}(\mathbf{x})\|_\infty, & \text{if} \quad p = \infty. \end{cases}$$

- Given a function $\mathbf{f} \in L^p(k, s)$, we denote by $\lambda^{(i)}(\mathbf{f})$ [resp. $\sigma^{(i)}(\mathbf{f})$], $i = 1, \ldots, s$, the eigenvalue [resp. singular value] functions of $\mathbf{f}$ and by $\left(\lambda^{(i)}(\mathbf{f})\right)(\theta)$ [resp. $\left(\sigma^{(i)}(\mathbf{f})\right)(\theta)$], $i = 1, \ldots, s$, their evaluation at a point $\theta \in \mathcal{I}_k$.

- If $z \in \mathbb{C}$ and $\epsilon > 0$, we denote by $D(z, \epsilon) = \{\omega \in \mathbb{C} : |\omega - z| < \epsilon\}$ the disk centered at z and with radius ϵ. If $S \subseteq \mathbb{C}$, $D(S, \epsilon) = \cup_{z\in S} D(z, \epsilon)$ denotes the $\epsilon-$expansion of S.

I.2 Multi-index notation

We introduce the multi-index notation that will be systematically used throughout the thesis.

A vector $\mathbf{i} = (i_1, i_2, \ldots, i_k) \in \mathbb{Z}^k$ is called a k-index (or simply a multi-index). For a more detailed description see [76].

- $\mathbf{0}$, $\mathbf{e}$, $\mathbf{2}$,... are respectively the multi-indices of all zeros, all ones, all twos, ... and their size will be clear from the context.

- For all $\mathbf{m} = (m_1, m_2, \ldots, m_k) \in \mathbb{Z}^k$ we set $N(\mathbf{m}) = m_1 m_2 \ldots m_k$ and we write $\mathbf{m} \to \infty$ to indicate that all the components of $\mathbf{m}$ tend to infinity, i.e. $\min_{i=1,\ldots,k} m_i \to \infty$.

- For all $\mathbf{h}$, $\mathbf{m} \in \mathbb{Z}^k$, $\mathbf{h} \le \mathbf{m}$ means $h_i \le m_i$, $\forall\, i = 1, \ldots, k$.

- If $\mathbf{h}$, $\mathbf{m} \in \mathbb{Z}^k$ are such that $\mathbf{h} \le \mathbf{m}$, the multi-index range $\mathbf{h}, \ldots, \mathbf{m}$ is the set

$$\{\mathbf{j} \in \mathbb{Z}^k : \mathbf{h} \le \mathbf{j} \le \mathbf{m}\}.$$

- When a k-index $\mathbf{j}$ varies over a multi-index range $\mathbf{h}, \ldots, \mathbf{m}$ (and we write $\mathbf{j} = \mathbf{h}, \ldots, \mathbf{m}$) it is understood that $\mathbf{j}$ varies from $\mathbf{h}$ to $\mathbf{m}$ following the standard lexicographic ordering. Note that $\mathbf{h}, \ldots, \mathbf{m}$ consists of $N(\mathbf{m}\text{-}\mathbf{h}+\mathbf{e})$ k-indices.

- All the algebraic operations involving k-indices that have no meaning in the space $\mathbb{Z}^k$ must always be interpreted in the componentwise sense: $\mathbf{ij} = (i_1 j_1, \ldots, i_k j_k)$, $\alpha\mathbf{i}/\mathbf{j} = (\alpha i_1/j_1, \ldots, \alpha i_k/j_k)$, for all $\alpha \in \mathbb{C}$ and all $j_1, \ldots, j_k \neq 0$, $\mathbf{i} \mod \mathbf{j} = (i_1 \mod j_1, \ldots, i_k \mod j_k)$, $\max(\mathbf{i}, \mathbf{j}) = (\max(i_1, j_1), \ldots, \max(i_k, j_k))$, and so on.

- Given $\mathbf{h}, \mathbf{m} \in \mathbb{Z}^k$, with $\mathbf{h} \leq \mathbf{m}$, the notation $\sum_{\mathbf{j}=\mathbf{h}}^{\mathbf{m}}$ indicates the summation over all multi-indices $\mathbf{j} = \mathbf{h}, \ldots, \mathbf{m}$.

- If $\mathbf{m} \in \mathbb{N}^k$ then

$$\mathbf{x} = [x_{\mathbf{i}}]_{\mathbf{i}=\mathbf{e}}^{\mathbf{m}}$$

  is a vector of size $N(\mathbf{m})$ whose components $x_{\mathbf{i}}$, $\mathbf{i} = \mathbf{e}, \ldots, \mathbf{m}$ are sorted in accordance with the lexicographic ordering. Similarly

$$\mathbf{X} = [x_{\mathbf{ij}}]_{\mathbf{i},\mathbf{j}=\mathbf{e}}^{\mathbf{m}}$$

  is the $N(\mathbf{m}) \times N(\mathbf{m})$ matrix whose components are indexed by two k-indices, both varying in $\mathbf{e}, \ldots, \mathbf{m}$ according the lexicographic ordering.

  Example

  Let A be the matrix

$$A = \begin{bmatrix} 4 & 4 & 0 & 0 \\ 4 & 4 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 2 & 2 \end{bmatrix}. \tag{I.1}$$

  Instead of using the traditional linear indices $i, j = 1, \ldots, 4$, we can index the entries of A by means of two multi-indices $\mathbf{i}, \mathbf{j} = \mathbf{e}, \ldots, \mathbf{2}$. Thus, instead of $[A_{ij}]_{i,j=1}^{4}$, we have $[A_{\mathbf{ij}}]_{\mathbf{i},\mathbf{j}=\mathbf{e}}^{\mathbf{2}}$.

  The indexing of the entries of A with two multi-indices $\mathbf{i}, \mathbf{j}$ reflects the fact that we are thinking at the matrix A as a block matrix as (I.1): for all $\mathbf{i}, \mathbf{j} = \mathbf{e}, \ldots, \mathbf{2}$ the entry $A_{\mathbf{ij}}$ is the (i_2, j_2) entry of the (i_1, j_1) block of A.

Throughout the thesis we indicate by $\{A_{\mathbf{n}}\}_{\mathbf{n}\in\mathbb{N}^k}$, or simply $\{A_{\mathbf{n}}\}_{\mathbf{n}}$, the matrix sequence whose elements are the matrices $A_{\mathbf{n}}$ of dimensions $N(\mathbf{n}, s) \times N(\mathbf{n}, s)$, with $N(\mathbf{n}, s) = sN(\mathbf{n}) = sn_1 n_2 \ldots n_k$, $\mathbf{n} = (n_1, n_2, \ldots, n_k)$.

I.3 Spectral distribution of matrix sequences

In this section we first introduce the concept of spectral/singular value distribution of generic matrix sequence $\{A_{\mathbf{n}}\}_{\mathbf{n}\in\mathbb{N}^v}$, $v \geq 1$ (whose dimension, $N \equiv N(\mathbf{n}, s)$, has to be a monotonic function with respect to every single variable n_i, $i = 1, \ldots, v$). Secondly we provide the formulation of Toeplitz and circulant matrix sequences in the most general block k-level form, recalling their main algebraic and spectral properties. In particular special attention is dedicated to the localization results and to the asymptotic behaviour of the extremal eigenvalues of the Hermitian Block Toeplitz sequences.

Definition I.3.1 (**clustering of a matrix-sequence**). *Let $S \subseteq \mathbb{C}$ be a nonempty subset of $\mathbb{C}$. Let $\{A_{\mathbf{n}}\}_{\mathbf{n}\in\mathbb{N}^v}$, $v \geq 1$, be a sequence of matrices with eigenvalues $\lambda_j(A_{\mathbf{n}})$, $j = 1,\ldots,N$ and singular values $\sigma_j(A_{\mathbf{n}})$, $j = 1,\ldots,N$.*

- *We say that $\{A_{\mathbf{n}}\}_{\mathbf{n}\in\mathbb{N}^v}$ is strongly clustered at S (in the sense of the eigenvalues), or equivalently that the eigenvalues of $\{A_{\mathbf{n}}\}_{\mathbf{n}\in\mathbb{N}^v}$ are strongly clustered at S, if, for every $\epsilon > 0$, the number of eigenvalues of $A_{\mathbf{n}}$ outside $D(S,\epsilon)$ is bounded by a constant C_ϵ independent of $\boldsymbol{n}$. In other words, for every $\epsilon > 0$ we have*

$$\#\{j \in \{1,\ldots,N\} :\ \lambda_j(A_{\mathbf{n}}) \notin D(S,\epsilon)\} = O(1) \quad as \quad \boldsymbol{n} \to \infty. \tag{I.2}$$

- *We say that $\{A_{\mathbf{n}}\}_{\mathbf{n}\in\mathbb{N}^v}$ is weakly clustered at S (in the sense of the eigenvalues), or equivalently that the eigenvalues of $\{A_{\mathbf{n}}\}_{\mathbf{n}\in\mathbb{N}^v}$ are weakly clustered at S, if, for every $\epsilon > 0$,*

$$\#\{j \in \{1,\ldots,N\} :\ \lambda_j(A_{\mathbf{n}}) \notin D(S,\epsilon)\} = o(N) \quad as \quad \boldsymbol{n} \to \infty. \tag{I.3}$$

By replacing "eigenvalues" with "singular values" and $\lambda_j(A_{\mathbf{n}})$ with $\sigma_j(A_{\mathbf{n}})$ in (I.2)–(I.3)*, we obtain the definitions of a matrix-sequence strongly or weakly clustered at S in the sense of the singular values.*

When we write strong/weak cluster, matrix-sequence strongly/weakly clustered, etc., without further specifications, it is understood "in the sense of the eigenvalues".

Definition I.3.2. ***[essential range of a complex-valued function]****. Let $f : D \subset \mathbb{R}^\ell \to \mathbb{C}$, $\ell \geq 1$, be a measurable complex-valued function defined on a measurable set with $0 < \mu_\ell(D) < \infty$. The essential range of f is denoted by $\mathcal{ER}(f)$ and is defined as the set of points $z \in \mathbb{C}$ such that, for every $\epsilon > 0$, the measure of the set $\{f(\boldsymbol{\theta}) \in D(z,\epsilon)\}$ is positive. In other words,*

$$\mathcal{ER}(f) = \{z \in \mathbb{C} :\ \mu_\ell\{f(\boldsymbol{\theta}) \in D(z,\epsilon)\} > 0 \quad for \quad all \quad \epsilon > 0\}.$$

Definition I.3.3. ***[essential range of a matrix-valued function]****. Let $\boldsymbol{f} : D \subset \mathbb{R}^\ell \to \mathbb{C}^{s\times s}$, $\ell \geq 1$, be a measurable matrix-valued function defined on a measurable set with $0 < \mu_\ell(D) < \infty$. The essential range of $\boldsymbol{f}$ is denoted by $\mathcal{ER}(\boldsymbol{f})$ and is defined as the union of the essential ranges of the eigenvalue functions of $\boldsymbol{f}$, $\lambda^{(i)}(\boldsymbol{f}) : D \to \mathbb{C}$, $i = 1,\ldots,s$. In other words,*

$$\mathcal{ER}(\boldsymbol{f}) = \cup_{i=1}^s \mathcal{ER}(\lambda^{(i)}(\boldsymbol{f})).$$

Definition I.3.4. ***[spectral/singular value distribution]****. Let $\boldsymbol{f} : G \to \mathbb{C}^{s\times s}$ be a measurable function, defined on a measurable set $G \subset \mathbb{R}^\ell$ with $\ell \geq 1$, $0 < \mu_\ell(G) < \infty$. Let $\mathcal{C}_0(\mathbb{K})$ be the set of continuous functions with compact support over $\mathbb{K} \in \{\mathbb{C}, \mathbb{R}_0^+\}$ and let $\{A_{\mathbf{n}}\}_{\mathbf{n}\in\mathbb{N}^v}$, $v \geq 1$, be a sequence of matrices with eigenvalues $\lambda_j(A_{\mathbf{n}})$, $j = 1,\ldots,N$ and singular values $\sigma_j(A_{\mathbf{n}})$, $j = 1,\ldots,N$.*

- *$\{A_{\mathbf{n}}\}_{\mathbf{n}\in\mathbb{N}^v}$ is* distributed as the pair $(\mathbf{f}, G)$ in the sense of the eigenvalues, *in symbols*

$$\{A_{\mathbf{n}}\}_{\mathbf{n}\in\mathbb{N}^v} \sim_\lambda (\boldsymbol{f}, G),$$

if the following limit relation holds for all $F \in \mathcal{C}_0(\mathbb{C})$:

$$\lim_{\mathbf{n}\to\infty} \frac{1}{N}\sum_{j=1}^N F(\lambda_j(A_{\mathbf{n}})) = \frac{1}{\mu_\ell(G)}\int_G \frac{\sum_{i=1}^s F\Big(\big(\lambda^{(i)}(\boldsymbol{f})\big)(\boldsymbol{\theta})\Big)}{s}\mathrm{d}\boldsymbol{\theta}. \tag{I.4}$$

- $\{A_{\mathbf{n}}\}_{\mathbf{n}\in\mathbb{N}^v}$ *is* distributed as the pair $(\mathbf{f}, G)$ in the sense of the singular values, *in symbols*

$$\{A_{\mathbf{n}}\}_{\mathbf{n}\in\mathbb{N}^v} \sim_\sigma (\boldsymbol{f}, G),$$

if the following limit relation holds for all $F \in \mathcal{C}_0(\mathbb{R}_0^+)$*:*

$$\lim_{\mathbf{n}\to\infty} \frac{1}{N}\sum_{j=1}^{N} F(\sigma_j(A_{\mathbf{n}})) = \frac{1}{\mu_\ell(G)}\int_G \frac{\sum_{i=1}^{s} F\Big(\left(\sigma^{(i)}(\boldsymbol{f})\right)(\boldsymbol{\theta})\Big)}{s}\mathrm{d}\boldsymbol{\theta}. \tag{I.5}$$

In this setting the expression $\mathbf{n}\to\infty$ *means that every component of the vector* $\mathbf{n}$ *tends to infinity, that is,* $\min_{i=1,\ldots,v} n_i \to \infty$.

Remark 1. *Denote by* $\lambda^{(1)}(\boldsymbol{f}),\ldots,\lambda^{(s)}(\boldsymbol{f})$ *and by* $\sigma^{(1)}(\boldsymbol{f}),\ldots,\sigma^{(s)}(\boldsymbol{f})$ *the eigenvalues and the singular values of a* $s\times s$ *matrix-valued function* $\boldsymbol{f}$*, respectively. If* $\boldsymbol{f}$ *is smooth enough, an informal interpretation of the limit relation* (I.4) *(resp.* (I.5)*) is that when the matrix-size of* $A_{\mathbf{n}}$ *is sufficiently large, then* N/s *eigenvalues (resp. singular values) of* $A_{\mathbf{n}}$ *can be approximated by a sampling of* $\lambda^{(1)}(\boldsymbol{f})$ *(resp.* $\sigma^{(1)}(\boldsymbol{f})$*) on a uniform equispaced grid of the domain* G*, and so on until the last* N/s *eigenvalues (resp. singular values) which can be approximated by an equispaced sampling of* $\lambda^{(s)}(\boldsymbol{f})$ *(resp.* $\sigma^{(s)}(\boldsymbol{f})$*) in the domain.*

For example, take G any domain as in Definition I.3.4 and let $F = \chi_{[a,b]}(\cdot)$ for a fixed real interval $[a,b]$ such that

$$\mu_\ell\left\{\boldsymbol{\theta}\in G : \left(\lambda^{(r)}(\mathbf{f})\right)(\boldsymbol{\theta}) = a\right\} = \mu_\ell\left\{\boldsymbol{\theta}\in G : \left(\lambda^{(r)}(\mathbf{f})\right)(\boldsymbol{\theta}) = b\right\} = 0 \tag{I.6}$$

for every $r = 1,\ldots,s$. Note that $F = \chi_{[a,b]}(\cdot)$ is a discontinuous function, but, under the assumptions in (I.6), the limit relation (I.4) still holds. The argument of the proof relies in choosing two families of continuous approximations $\{F_\delta^-\}_\delta$, $\{F_\delta^+\}_\delta$ of $\chi_{[a,b]}$ such that $F_\delta^+ < \chi_{[a,b]} < F_\delta^+$ (see [117] for more details). If we define

$$m_r = \mathrm{essinf}_G\left(\lambda^{(r)}(\mathbf{f})\right)(\boldsymbol{\theta}), \qquad M_r = \mathrm{esssup}_G\left(\lambda^{(r)}(\mathbf{f})\right)(\boldsymbol{\theta}), \qquad r = 1,\ldots,s,$$

when $F = \chi_{[m_r,M_r]}(\cdot)$, then equation (I.4) becomes

$$\lim_{n\to\infty}\frac{1}{N}\sum_{j=1}^{N}\chi_{[m_r,M_r]}\left(\lambda_j(A_{\mathbf{n}})\right) = \frac{1}{s\mu_\ell(G)}\int_G\sum_{i=1}^{s}\chi_{[m_r,M_r]}\left(\left(\lambda^{(i)}(\mathbf{f})\right)(\boldsymbol{\theta})\right)\mathrm{d}\boldsymbol{\theta}, \tag{I.7}$$

and hence

$$\begin{aligned}\lim_{N\to\infty}\frac{1}{N}\#\left\{j : \lambda_j(A_{\mathbf{n}})\in[m_r,M_r]\right\} = \\ \frac{1}{s\mu_\ell(G)}\sum_{i=1}^{s}\mu_\ell\left\{\boldsymbol{\theta}\in G : \left(\lambda^{(i)}(\mathbf{f})\right)(\boldsymbol{\theta})\in[m_r,M_r]\right\}.\end{aligned} \tag{I.8}$$

Moreover, if

$$\mathrm{esssup}_G\left(\lambda^{(r)}(\mathbf{f})\right)(\boldsymbol{\theta}) \le \mathrm{essinf}_G\left(\lambda^{(r+1)}(\mathbf{f})\right)(\boldsymbol{\theta}), \quad r = 1,\ldots,s-1,$$

and

$$\mu_\ell \left\{ \boldsymbol{\theta} \in G \,:\, \left(\lambda^{(r)}(\mathbf{f})\right)(\boldsymbol{\theta}) = c \right\} = 0,\ \forall c \in \mathbb{R},\ r = 1, \ldots, s-1,$$

then equation (I.8) in turn becomes

$$\lim_{n\to\infty} \frac{1}{N} \#\left\{ j \,:\, \lambda_j(A_{\mathbf{n}}) \in [m_r, M_r] \right\} = \frac{1}{s\mu_\ell(G)} \mu_\ell \left\{ \boldsymbol{\theta} \in G \,:\, \left(\lambda^{(r)}(\mathbf{f})\right)(\boldsymbol{\theta}) \in [m_r, M_r] \right\} = \frac{1}{s}$$

which means that

$$\#\left\{ j \,:\, \lambda_j(A_{\mathbf{n}}) \in [m_r, M_r] \right\} = \frac{N}{s} + o(N).$$

I.4 Toeplitz structures

Toeplitz matrices represent an important and very active topic introduced more than one hundred years ago in the original papers by O. Toeplitz [143, 144]. The treated Toeplitz matrices derive mostly from the approximation of differential equations, but, in general, they can be found in many applications: they arise, for example, also from structured Markov chains [15], signal and image processing problems with space invariant nature [46, 82] and financial applications [110].

In the following, we provide the formulation of the most general multilevel block form of Toeplitz matrix. We start from the simplest concept in the scalar setting and we generalize the definitions, achieving the case of multilevel block Toeplitz matrix generated by a matrix-valued function $\mathbf{f}$.

I.4.1 Scalar Toeplitz matrices

A matrix of order n, having a fixed entry along each diagonal, is called Toeplitz and enjoys the expression

$$A_n = [a_{i-j}]_{i,j=1}^n = \begin{bmatrix} a_0 & a_{-1} & a_{-2} & \cdots & \cdots & a_{-(n-1)} \\ a_1 & \ddots & \ddots & \ddots & & \vdots \\ a_2 & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & a_{-2} \\ \vdots & & \ddots & \ddots & \ddots & a_{-1} \\ a_{n-1} & \cdots & \cdots & a_2 & a_1 & a_0 \end{bmatrix}.$$

An interesting case of Toeplitz matrix is given by $T_n(f) \in \mathbb{C}^{n\times n}$, that is associated with a scalar valued function $f \in L^1(1,1)$, defined on $[-\pi, \pi]$ and periodically extended on the whole real line. Such a matrix $T_n(f)$ is defined via the Fourier series of f

$$f(\theta) = \sum_{k=-\infty}^{\infty} \hat{f}_k e^{\iota k\theta},$$

and has the following expression,

$$T_n(f) = \left[\hat{f}_{i-j}\right]_{i,j=1}^n,$$

where the quantities $\hat{f}_k$

$$\hat{f}_k = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(\theta)\,\mathrm{e}^{-\iota k\theta}\mathrm{d}\theta, \qquad k \in \mathbb{Z},$$

are the Fourier coefficients of f.

If n is varying in $\mathbb{N}$, we obtain a matrix sequence $\{T_n(f)\}_n$, consisting of Toeplitz matrices of increasing size.

We refer to $\{T_n(f)\}_n$ as the Toeplitz sequence generated by f, which in turn is called the generating function of $\{T_n(f)\}_n$.

There are many properties of $T_n(f)$ that follow by direct computation from assumptions on f. In the following, we report those that will be used in next chapters [31, 96].

1. If f is complex-valued, then $T_n(f)$ is non-Hermitian for all sufficiently large n. Conversely, if f is real-valued, then $T_n(f)$ is Hermitian for all n.

2. If f is real-valued, nonnegative and not identically zero almost everywhere, then $T_n(f)$ is HPD for all n.

3. If f is even, $f(\theta) = f(-\theta)$, $T_n(f)$ is symmetric for all n. Thus, from property one, if f is real-valued and even, $T_n(f)$ is real and symmetric for all n.

I.4.2 Block and multilevel block Toeplitz matrices

The following Subsection is devoted to the generalization of the concept of scalar Toeplitz matrix. In general the entries a_k of the matrix $A_n = [a_{i-j}]_{i,j=1}^n$ can be matrices themselves. If the dimensions of the blocks are $s \times s$, $s > 1$, the resulting matrix is the block Toeplitz matrix $\mathbf{A}_n$, where the bold points out the following block structure of the matrix

$$\mathbf{A}_n = [A_{i-j}]_{i,j=1}^n = \begin{bmatrix} A_0 & A_{-1} & A_{-2} & \cdots & \cdots & A_{-(n-1)} \\ A_1 & \ddots & \ddots & \ddots & & \vdots \\ A_2 & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & A_{-2} \\ \vdots & & \ddots & \ddots & \ddots & A_{-1} \\ A_{n-1} & \cdots & \cdots & A_2 & A_1 & A_0 \end{bmatrix},$$

where $A_{-(n-1)}, \ldots, A_{n-1} \in \mathbb{C}^{s\times s}$ are the "block" generalization of $a_{-(n-1)}, \ldots, a_{n-1}$ of the setting $s = 1$. Note that now the size of $\mathbf{A}_n$ is $N = N(n, s) = sn$.

Following the scalar case, we can define particular block (resp. k-level block) Toeplitz matrices $T_{\mathbf{n}}(\mathbf{f})$ starting from matrix-valued (resp. k-variate matrix-valued) function $\mathbf{f} \in L^1(1, s)$ (resp. $\mathbf{f} \in L^1(k, s)$). For the block settings we will write the function $\mathbf{f}$ (and corresponding Fourier coefficients) in bold.

Definition I.4.1. *Let the Fourier coefficients of a given function $\boldsymbol{f} \in L^1(k, s)$ be defined as*

$$\hat{\boldsymbol{f}}_{\mathbf{j}} := \frac{1}{(2\pi)^k}\int_{\mathcal{I}_k} \boldsymbol{f}(\boldsymbol{\theta})\mathrm{e}^{-\iota\langle \mathbf{j},\boldsymbol{\theta}\rangle}\,\mathrm{d}\boldsymbol{\theta} \in \mathbb{C}^{s\times s}, \qquad \mathbf{j} = (j_1, \ldots, j_k) \in \mathbb{Z}^k, \quad \iota^2 = -1, \tag{I.9}$$

where $\langle \mathbf{j}, \boldsymbol{\theta} \rangle = \sum_{t=1}^{k} j_t \theta_t$ *and the integrals in* (I.9) *are computed componentwise.*
Then, the $\mathbf{n}$*th Toeplitz matrix associated with* $\boldsymbol{f}$ *is the matrix of order* $N(\boldsymbol{n}, s) = sn_1 n_2 \dots n_k$ *given by*

$$T_{\mathbf{n}}(\boldsymbol{f}) = \sum_{\mathbf{j}=-(\mathbf{n}-\mathbf{e})}^{\mathbf{n}-\mathbf{e}} J_{n_1}^{j_1} \otimes \cdots \otimes J_{n_k}^{j_k} \otimes \hat{\boldsymbol{f}}_{\mathbf{j}}. \tag{I.10}$$

where $\mathbf{e} = (1, \dots, 1) \in \mathbb{N}^k$, $\mathbf{j} = (j_1, \dots, j_k) \in \mathbb{N}^k$ *and* $J_{n_\xi}^{j_\xi}$ *is the* $n_\xi \times n_\xi$ *matrix whose* (i,h)*th entry equals 1 if* $(i-h) = j_\xi$ *and* 0 *otherwise.*
The set $\{T_{\mathbf{n}}(\boldsymbol{f})\}_{\mathbf{n}}$ *(with* $\mathbf{n} \in \mathbb{N}^k$*) is called the* family of k-level Toeplitz matrices generated by $\mathbf{f}$*, that in turn is referred to as the* generating function or the symbol of $\{T_{\mathbf{n}}(\mathbf{f})\}_{\mathbf{n}}$.

I.4.3 Spectral analysis of Hermitian block Toeplitz sequences: distribution results

The singular value and spectral distribution of Toeplitz matrix sequences has been of interest over the past few decades.

The representation of the spectral distribution of Toeplitz sequences in terms of a function (i.e. the symbol) was performed by Szegő, Tyrtyshnikov and Zamarashkin, Tilli see, e.g., [81, 140, 146].

The earliest result on the eigenvalue distribution of Toeplitz matrices was established by Szegő in [81], proving that the eigenvalues of the Toeplitz matrix $T_n(f)$ generated by a real-valued $f \in L^\infty([-\pi, \pi])$ are asymptotically distributed as f.

Zamarashkin and Tyrtyshnikov [152], and Tilli [140] further weakened the requirement on f and showed that the same result holds for $f \in L^1([-\pi, \pi])$.

The work of Tilli [141] produced a key contribution, by allowing the concept of smoothly varying diagonals and so allowing to treat the approximation of one-variable differential operators with variable coefficients.

Based on an approximation class sequence approach, Garoni, Serra-Capizzano, and Vassalos [79] provided the same theorem for $f \in L^1([-\pi, \pi])$ in the framework of the newly developed theory of Generalized Locally Toeplitz (GLT) sequences [77].

In the following we illustrate the result concerning the spectral distribution of Toeplitz sequences under the hypothesis of f being a real-valued function.

Theorem I.4.1 ([81]). *Let* $f \in L^1(k, 1)$ *be a real-valued function with* $k \geq 1$. *Then,*

$$\{T_{\mathbf{n}}(f)\}_{\mathbf{n} \in \mathbb{N}^k} \sim_\lambda (f, \mathcal{I}_k).$$

In the case where $\mathbf{f}$ is a Hermitian matrix-valued function, according to Tilli [140], the previous theorem can be extended as follows:

Theorem I.4.2 ([140]). *Let* $\boldsymbol{f} \in L^1(k, s)$ *be a Hermitian matrix-valued function with* $k \geq 1, s \geq 2$. *Then,*

$$\{T_{\mathbf{n}}(\boldsymbol{f})\}_{\mathbf{n} \in \mathbb{N}^k} \sim_\lambda (\boldsymbol{f}, \mathcal{I}_k).$$

Remark 2. *If $\{T_{\mathbf{n}}(\boldsymbol{f})\}_{\mathbf{n}\in\mathbb{N}^k}$ is such that each $T_{\mathbf{n}}(\boldsymbol{f})$ is symmetric with symmetric and real blocks, then the symbol has the additional property that $\boldsymbol{f}(\pm\theta_1,\ldots,\pm\theta_k)\equiv\boldsymbol{f}(\theta_1,\ldots,\theta_k)$, $\forall(\theta_1,\ldots,\theta_k)\in\mathcal{I}_k^+=[0,\pi]^k$ and therefore Theorem I.4.2 can be rephrased as*

$$\{T_{\mathbf{n}}(\boldsymbol{f})\}_{\mathbf{n}\in\mathbb{N}^k}\sim_\lambda\ (\boldsymbol{f},\mathcal{I}_k^+).$$

In the Toeplitz setting, when $\mathbf{f}$ is a k-variate polynomial, the quantity $o(N)$ of Remark 1 becomes proportional to $N^{1-\frac{1}{k}}$, with constant proportional to s and to the degree of the polynomial.

I.4.4 Spectral analysis of Hermitian block Toeplitz sequences: extremal eigenvalues

Concerning the localization and the extremal behaviour of the spectra of Toeplitz sequences there is a lot of work in the last 80 years culminated with the works of Böttcher, Grudsky, and Serra-Capizzano [18, 115, 117, 119, 121]. More precisely, if f is a real-valued function, then we have the following result.

Theorem I.4.3 ([18, 119]). *Let $f\in L^1(k,1)$ be a real-valued function with $k\geq 1$. Let m be the essential infimum of f and M be the essential supremum of f.*

1. *If $m=M$ then $f=m$ a.e. and $T_{\mathbf{n}}(f)$ coincides with m times the identity of order $N(\boldsymbol{n})$.*
2. *If $m<M$ then all the eigenvalues of $T_{\mathbf{n}}(f)$ belong to the open set (m,M) for every $\mathbf{n}\in\mathbb{N}^k$.*
3. *If $m=\boldsymbol{0}$ and $\tilde{\boldsymbol{\theta}}$ is the unique zero of f such that there exist positive constants c,C,α for which*
$$c\|\boldsymbol{\theta}-\tilde{\boldsymbol{\theta}}\|^\alpha\leq f(\boldsymbol{\theta})\leq C\|\boldsymbol{\theta}-\tilde{\boldsymbol{\theta}}\|^\alpha,$$
then the minimal eigenvalue of $T_{\mathbf{n}}(f)$ goes to zero as $(N(\boldsymbol{n}))^{-\alpha/k}$.

In the case where $\mathbf{f}$ is a Hermitian matrix-valued function, according to the analysis in [117, 121], the previous theorem can be extended as follows:

Theorem I.4.4 ([117, 121]). *Let $\boldsymbol{f}\in L^1(k,s)$ be a Hermitian matrix-valued function with $k\geq 1, s\geq 2$. Let m_1 be the essential infimum of the minimal eigenvalue of $\boldsymbol{f}$, M_1 be the essential supremum of the minimal eigenvalue of $\boldsymbol{f}$, m_s be the essential infimum of the maximal eigenvalue of $\boldsymbol{f}$, and M_s be the essential supremum of the maximal eigenvalue of $\boldsymbol{f}$.*

1. *If $m_1=M_s$ then $\boldsymbol{f}$ is the constant m_1I_s a.e. and $T_{\mathbf{n}}(\boldsymbol{f})$ coincides with m_1 times the identity of size $N(\boldsymbol{n},s)=sn_1n_2\ldots n_k$.*
2. *If $m_1<M_1$ then all the eigenvalues of $T_{\mathbf{n}}(\boldsymbol{f})$ belong to the open set $(m_1,M_s]$ for every $\mathbf{n}\in\mathbb{N}^k$. If $m_s<M_s$ then all the eigenvalues of $T_{\mathbf{n}}(\boldsymbol{f})$ belong to the open set $[m_1,M_s)$ for every $\mathbf{n}\in\mathbb{N}^k$.*
3. *If $m_1=\boldsymbol{0}$ and $\tilde{\boldsymbol{\theta}}$ is the unique zero of $\lambda^{(\min)}(\boldsymbol{f})$ such that there exist positive constants c,C,α for which*
$$c\|\boldsymbol{\theta}-\tilde{\boldsymbol{\theta}}\|^\alpha\leq\lambda^{(\min)}(\boldsymbol{f}(\boldsymbol{\theta}))\leq C\|\boldsymbol{\theta}-\tilde{\boldsymbol{\theta}}\|^\alpha,$$
then the minimal eigenvalue of $T_{\mathbf{n}}(\boldsymbol{f})$ goes to zero as $(N(\boldsymbol{n}))^{-\alpha/k}$.

I.5 Trigonometric polynomials and banded Toeplitz matrices

In the current section we recall the definitions and the principal properties of a k-variate matrix-valued trigonometric polynomial and we concentrate on the special case of Toeplitz sequences, having a trigonometric polynomial as generating function.

In the next chapters we deal with generating functions of various nature. We recall that

- depending on whether the dimension of the domain $\mathcal{I}_k$ is $k = 1$ or $k > 1$ we deal with an univariate or a multivariate trigonometric polynomial, respectively;
- depending on whether $s = 1$ or $s > 1$ in the codomain $\mathbb{C}^{s\times s}$ we deal with a scalar valued or a matrix-valued trigonometric polynomial, respectively.

The polynomials treated in the thesis will be multivariate and matrix-valued (in **Chapter II**), univariate and scalar (in **Chapters III**, **IV)**, and univariate and matrix-valued (in **Chapter V**). Thus we recall the definitions of all the possible four configurations, that are:

- univariate and scalar;
- univariate and matrix-valued;
- multivariate and scalar;
- multivariate and matrix-valued.

Definition I.5.1. *[**univariate and scalar**] A scalar univariate trigonometric polynomial is a function $f : \mathcal{I}_1 \to \mathbb{C}$ that can be written as a finite linear combination of the Fourier frequencies $\{\mathrm{e}^{\iota j\theta} : j \in \mathbb{Z}\}$. Note that $f(\theta)$ has a finite number of nonzero Fourier coefficients $\hat{f}_j$. The degree of f is a positive integer r defined as*

$$r = \max\{|j| : \hat{f}_j \neq 0,\, j \in \mathbb{Z}\}.$$

Hence f can be written as the Fourier sum

$$f(\theta) = \sum_{j=-r}^{r} \hat{f}_j \mathrm{e}^{\iota j\theta}.$$

We say that f is a real-valued cosine trigonometric polynomial (RCTP), if f is the following scalar univariate trigonometric polynomial of degree r.

$$f(\theta) = \hat{f}_0 + 2\sum_{l=1}^{r} \hat{f}_l \cos(l\theta), \qquad \hat{f}_0, \hat{f}_1, \ldots, \hat{f}_r \in \mathbb{R}.$$

In the following, every time we deal with cosine trigonometric polynomial, we can replace, using Remark 2, the interval $\mathcal{I}_1$ with $\mathcal{I}_1^+$.

The properties of Toeplitz matrices in Section I.4.1 imply that the nth Toeplitz matrix generated by f is the banded real symmetric matrix given by

$$T_n(f) = \begin{bmatrix} \hat{f}_0 & \hat{f}_1 & \cdots & \hat{f}_r & & & & & & & \\ \hat{f}_1 & \ddots & \ddots & & \ddots & & & & & & \\ \vdots & \ddots & \ddots & \ddots & & \ddots & & & & & \\ \hat{f}_r & & \ddots & \ddots & \ddots & & \ddots & & & & \\ & \ddots & & \ddots & \ddots & \ddots & & \ddots & & & \\ & & \boxed{\hat{f}_r} & \cdots & \boxed{\hat{f}_1} & \boxed{\hat{f}_0} & \boxed{\hat{f}_1} & \cdots & \boxed{\hat{f}_r} & & \\ & & & \ddots & & \ddots & \ddots & \ddots & & \ddots & \\ & & & & \ddots & & \ddots & \ddots & \ddots & & \hat{f}_r \\ & & & & & \ddots & & \ddots & \ddots & \ddots & \vdots \\ & & & & & & \ddots & & \ddots & \ddots & \hat{f}_1 \\ & & & & & & & \hat{f}_r & \cdots & \hat{f}_1 & \hat{f}_0 \end{bmatrix}.$$

Furthermore $T_n(f)$ can be written as

$$T_n(f) = \tau_n(f) + H_n(f), \tag{I.11}$$

where, for ψ RCTP of degree r and orthogonal $Q = \left(\sqrt{\frac{2}{n+1}} \sin\left(\frac{ij\pi}{n+1}\right)\right)_{i,j=1}^{n}$, $\tau_n(\psi)$ is the following τ matrix [14] of size n generated by ψ

$$\tau_n(\psi) = Q \operatorname*{diag}_{1\le j\le n}\left(\psi\left(\frac{j\pi}{n+1}\right)\right)Q, \quad Q = Q^T = Q^{-1},$$

and $H_n(\psi)$ is the Hankel matrix

$$H_n(\phi) = \begin{bmatrix} \hat{\psi}_2 & \hat{\psi}_3 & \cdots & \hat{\psi}_r & & & & & \\ \hat{\psi}_3 & & \iddots & & & & & & \\ \vdots & \iddots & & & & & & & \\ \hat{\psi}_r & & & & & & & & \\ & & & & & & & \hat{\psi}_r & \\ & & & & & & \iddots & \vdots & \\ & & & & & \iddots & & \hat{\psi}_3 & \\ & & & & \hat{\psi}_r & \cdots & \hat{\psi}_3 & \hat{\psi}_2 & \end{bmatrix},$$

with $\operatorname{rank}(H_n(\psi)) \le 2(r-1)$.

Definition I.5.2. *[**univariate and matrix-valued**] A matrix-valued univariate trigonometric polynomial is a function $\boldsymbol{f} : \mathcal{I}_1 \to \mathbb{C}^{s\times s}$, $s > 1$, written as a finite linear combination of the*

Fourier frequencies $\{\mathrm{e}^{-\iota j\theta} : j \in \mathbb{Z}\}$ *or, equivalently, for all* $l, m = 1, \ldots, s$, *its* (l, m)*th component* $f_{lm} : \mathcal{I}_1 \to \mathbb{C}$ *is a scalar univariate trigonometric polynomial of degree* r_{lm}. *The degree of* $\boldsymbol{f}$ *is a positive integer* r *defined as*

$$\mathrm{r} = \max_{\substack{l=1,\ldots,s \\ m=1,\ldots,s}} r_{lm}.$$

Thus $\boldsymbol{f}$ *can be written as the Fourier sum*

$$\boldsymbol{f}(\theta) = \sum_{j=-\mathrm{r}}^{\mathrm{r}} \hat{\boldsymbol{f}}_j \mathrm{e}^{\iota j\theta}.$$

We say that $\mathbf{f}$ is an Hermitian matrix-valued trigonometric polynomial (HTP) with Fourier coefficients $\hat{\mathbf{f}}_0, \hat{\mathbf{f}}_1, \ldots, \hat{\mathbf{f}}_{\mathrm{r}} \in \mathbb{R}^{s\times s}$, if $\mathbf{f}$ is of the form

$$\mathbf{f}(\theta) = \sum_{l=-\mathrm{r}}^{\mathrm{r}} \hat{\mathbf{f}}_l \mathrm{e}^{\iota l\theta} = \hat{\mathbf{f}}_0 + \sum_{l=1}^{\mathrm{r}} \left(\hat{\mathbf{f}}_l \mathrm{e}^{\iota l\theta} + \hat{\mathbf{f}}_l^T \mathrm{e}^{-\iota l\theta} \right), \qquad \mathrm{r} = \deg\left(\mathbf{f}(\theta)\right),$$

where we set

$$\hat{\mathbf{f}}_{-l} = \hat{\mathbf{f}}_l^T, \quad l = 0, \ldots, \mathrm{r}. \tag{I.12}$$

The assumptions on $\mathbf{f}(\theta)$ imply that $T_n(\mathbf{f})$ is the $N(n,s) \times N(n,s)$ block real banded symmetric Toeplitz matrix with "block bandwidth" $2\mathrm{r}+1$, of the form

$$T_n(\mathbf{f}) = \begin{bmatrix}
\hat{\mathbf{f}}_0 & \hat{\mathbf{f}}_1^T & \cdots & \hat{\mathbf{f}}_{\mathrm{r}}^T & & & & & & & \\
\hat{\mathbf{f}}_1 & \ddots & \ddots & & \ddots & & & & & & \\
\vdots & \ddots & \ddots & \ddots & & \ddots & & & & & \\
\hat{\mathbf{f}}_{\mathrm{r}} & & \ddots & \ddots & \ddots & & \ddots & & & & \\
& \ddots & & \ddots & \ddots & \ddots & & \ddots & & & \\
& & \boxed{\hat{\mathbf{f}}_{\mathrm{r}}} & \cdots & \boxed{\hat{\mathbf{f}}_1} & \boxed{\hat{\mathbf{f}}_0} & \boxed{\hat{\mathbf{f}}_1^T} & \cdots & \boxed{\hat{\mathbf{f}}_{\mathrm{r}}^T} & & \\
& & & \ddots & & \ddots & \ddots & \ddots & & \ddots & \\
& & & & \ddots & & \ddots & \ddots & \ddots & & \hat{\mathbf{f}}_{\mathrm{r}}^T \\
& & & & & \ddots & & \ddots & \ddots & \ddots & \vdots \\
& & & & & & \ddots & & \ddots & \ddots & \hat{\mathbf{f}}_1^T \\
& & & & & & & \hat{\mathbf{f}}_{\mathrm{r}} & \cdots & \hat{\mathbf{f}}_1 & \hat{\mathbf{f}}_0
\end{bmatrix}.$$

We remark that

$$T_n(\mathbf{f}) = \tau_{N(n,s)}(\mathbf{f}) + H_{N(n,s)}(\mathbf{f}), \tag{I.13}$$

where, for $\boldsymbol{\psi}$ (HTP) of degree r and orthogonal $Q = \left(\sqrt{\frac{2}{n+1}} \sin\left(\frac{ij\pi}{n+1}\right)\right)_{i,j=1}^{n}$, $\tau_{N(n,s)}(\boldsymbol{\psi})$ is the following τ matrix [14] of size $N(n,s)$ generated by $\boldsymbol{\psi}$

$$\tau_N(\boldsymbol{\psi}) = (Q \otimes I_s) \operatorname*{diag}_{1\le j\le n} \left(\boldsymbol{\psi}\left(\frac{j\pi}{n+1}\right)\right)(Q \otimes I_s), \quad Q = Q^T = Q^{-1},$$

where $H_{N(n,s)}(\boldsymbol{\psi})$ is the Hankel matrix

$$H_{N(n,s)}(\boldsymbol{\psi}) = \begin{bmatrix} \hat{\boldsymbol{\psi}}_2 & \hat{\boldsymbol{\psi}}_3 & \cdots & \hat{\boldsymbol{\psi}}_{\mathrm{r}} & & & & \\ \hat{\boldsymbol{\psi}}_3 & & \iddots & & & & & \\ \vdots & \iddots & & & & & & \\ \hat{\boldsymbol{\psi}}_{\mathrm{r}} & & & & & & & \\ & & & & & & & \hat{\boldsymbol{\psi}}_{\mathrm{r}} \\ & & & & & & \iddots & \vdots \\ & & & & & \iddots & & \hat{\boldsymbol{\psi}}_3 \\ & & & & \hat{\boldsymbol{\psi}}_{\mathrm{r}} & \cdots & \hat{\boldsymbol{\psi}}_3 & \hat{\boldsymbol{\psi}}_2 \end{bmatrix},$$

with $\nu := \nu(s, \mathrm{r}) = \mathrm{rank}(H_{N(n,s)}(\boldsymbol{\psi})) \leq 2s(\mathrm{r}-1)$.

Definition I.5.3. *[**multivariate and scalar**] A scalar multivariate trigonometric polynomial is a function $f : \mathcal{I}_k \to \mathbb{C}$, $k > 1$, written as a finite linear combination of the Fourier frequencies $\{\mathrm{e}^{-\iota\langle \mathbf{j}, \boldsymbol{\theta}\rangle} : \boldsymbol{j} \in \mathbb{Z}^k\}$. The degree of f is a positive k-index $\boldsymbol{r}$ defined as*

$$\boldsymbol{r} = \max\{|\boldsymbol{j}| : \hat{f}_{\boldsymbol{j}} \neq 0, \boldsymbol{j} \in \mathbb{Z}^k\} = \max_{\boldsymbol{j}=(j_1,\ldots,j_k)\in\mathbb{Z}^k} \{(|j_1|, \ldots, |j_k|) : \hat{f}_{\boldsymbol{j}} \neq 0\}.$$

Thus f can be written as the Fourier sum

$$f(\boldsymbol{\theta}) = \sum_{\boldsymbol{j}=-\boldsymbol{r}}^{\boldsymbol{r}} \hat{f}_{\boldsymbol{j}} \mathrm{e}^{\langle \mathbf{j}, \boldsymbol{\theta}\rangle}.$$

We say that f is a multivariate real-valued cosine trigonometric polynomial (RCTP), if f is the following scalar multivariate trigonometric polynomial of degree $\boldsymbol{r}$.

$$f(\boldsymbol{\theta}) = \hat{f}_{\boldsymbol{0}} + 2\sum_{\boldsymbol{l}=\boldsymbol{e}}^{\boldsymbol{r}} \hat{f}_{\boldsymbol{l}} \cos(\langle \boldsymbol{l}, \boldsymbol{\theta}\rangle), \qquad \hat{f}_{\boldsymbol{0}}, \ldots, \hat{f}_{\boldsymbol{r}} \in \mathbb{R}.$$

The generalization of the properties in Section I.4.1 to the multilevel context imply that the $N(n,1) \times N(n,1)$ Toeplitz matrix generated by f is the real banded symmetric matrix given by

$$T_{\mathbf{n}}(f) = \sum_{\boldsymbol{l}=-(\boldsymbol{r}-\mathbf{e})}^{\boldsymbol{r}-\mathbf{e}} (J_{n_1}^{l_1} \otimes \cdots \otimes J_{n_k}^{l_k}) \hat{f}_{\boldsymbol{l}}. \tag{I.14}$$

where $J_{n_\xi}^{l_\xi}$ is defined as in formula (I.10).

Definition I.5.4. *[**multivariate and matrix-valued**] A matrix-valued multivariate trigonometric polynomial is a function $\boldsymbol{f} : \mathcal{I}_k \to \mathbb{C}^{s\times s}$, $k, s > 1$, written as a finite linear combination of the Fourier frequencies $\{\mathrm{e}^{-\iota\langle \mathbf{j}, \boldsymbol{\theta}\rangle} : \boldsymbol{j} \in \mathbb{Z}^k\}$ or, equivalently, for all $l, m = 1, \ldots, s$, its (l,m)th component $f_{lm} : \mathcal{I}_k \to \mathbb{C}$ is a scalar multivariate trigonometric polynomial of degree $\boldsymbol{r}_{lm}$. The degree of $\boldsymbol{f}$ is a positive k-index $\mathbf{r}$ defined as*

$$\mathbf{r} = \max_{\substack{l=1,\ldots,s \\ m=1,\ldots,s}} \boldsymbol{r}_{lm}.$$

Thus $\boldsymbol{f}$ can be written as the Fourier sum

$$\boldsymbol{f}(\boldsymbol{\theta}) = \sum_{\boldsymbol{j}=-\mathbf{r}}^{\mathbf{r}} \hat{\boldsymbol{f}}_{\boldsymbol{j}} \mathrm{e}^{\langle \mathbf{j}, \boldsymbol{\theta} \rangle}.$$

We say that $\mathbf{f}$ is an Hermitian matrix-valued multivariate trigonometric polynomial (HTP) with Fourier coefficients $\hat{\mathbf{f}}_0, \ldots, \hat{\mathbf{f}}_{\mathbf{r}} \in \mathbb{R}^{s \times s}$, if $\mathbf{f}$ is of the form

$$\mathbf{f}(\boldsymbol{\theta}) = \hat{\mathbf{f}}_{\mathbf{0}} + \sum_{\mathbf{l}=\mathbf{e}}^{\mathbf{r}} \left(\hat{\mathbf{f}}_{\mathbf{l}} \mathrm{e}^{\iota \langle \mathbf{l}, \boldsymbol{\theta} \rangle} + \hat{\mathbf{f}}_{\mathbf{l}}^{T} \mathrm{e}^{-\iota \langle \mathbf{l}, \boldsymbol{\theta} \rangle} \right), \qquad \mathbf{r} = \deg\left(\mathbf{f}(\boldsymbol{\theta})\right),$$

where we set

$$\hat{\mathbf{f}}_{-\mathbf{l}} = \hat{\mathbf{f}}_{\mathbf{l}}^{T}, \quad \mathbf{l} = \mathbf{0}, \ldots, \mathbf{r}. \tag{I.15}$$

The assumptions on $\mathbf{f}(\boldsymbol{\theta})$ imply that the $N(\mathbf{n}, s) \times N(\mathbf{n}, s)$ Toeplitz matrix $T_{\mathbf{n}}(\mathbf{f})$ generated by $\mathbf{f}$ is the multilevel block banded real and symmetric matrix given by

$$T_{\mathbf{n}}(\mathbf{f}) = \sum_{\mathbf{l}=-(\mathbf{r}-\mathbf{e})}^{\mathbf{r}-\mathbf{e}} (J_{n_1}^{l_1} \otimes \cdots \otimes J_{n_k}^{l_k}) \otimes \hat{\mathbf{f}}_{\mathbf{l}}. \tag{I.16}$$

where $J_{n_\xi}^{l_\xi}$ is defined as in formula (I.10).

I.6 Spectral analysis and computational features of block circulant matrices

In this section we report key features of the (block) circulant matrices, also in connection with the generating function.

Definition I.6.1. *Let the Fourier coefficients of a given function $\boldsymbol{f} \in L^1(k, s)$ be defined as in formula (I.9).*

Then, the $\mathbf{n}$th circulant matrix associated with $\boldsymbol{f}$ is the matrix of order $N(\boldsymbol{n}, s) = s n_1 n_2 \ldots n_k$ given by

$$C_{\mathbf{n}}(\boldsymbol{f}) = \sum_{\mathbf{j}=-(\mathbf{n}-\mathbf{e})}^{\mathbf{n}-\mathbf{e}} Z_{n_1}^{j_1} \otimes \cdots \otimes Z_{n_k}^{j_k} \otimes \hat{\boldsymbol{f}}_{\mathbf{j}}, \tag{I.17}$$

where $\mathbf{e} = (1, \ldots, 1) \in \mathbb{N}^k$, $\mathbf{j} = (j_1, \ldots, j_k) \in \mathbb{N}^k$ and $Z_{n_\xi}^{j_\xi}$ is the $n_\xi \times n_\xi$ matrix whose (i, h)th entry equals 1 if $(i - h)$ mod $n_\xi = j_\xi$ and 0 otherwise.

Theorem I.6.1 ([44])**.** *Let $f \in L^1(k, 1)$ be a complex-valued function with $k \geq 1$. Then, the following Schur decomposition of $C_{\mathbf{n}}(f)$ is valid:*

$$C_{\mathbf{n}}(f) = F_{\mathbf{n}} D_{\mathbf{n}}(f) F_{\mathbf{n}}^{*}, \tag{I.18}$$

where

$$D_{\mathbf{n}}(f) = \operatorname*{diag}_{\mathbf{0} \leq \mathbf{r} \leq \mathbf{n}-\mathbf{e}} \left(S_{\mathbf{n}}(f)\right)\left(\theta_{\mathbf{r}}^{(\mathbf{n})}\right), \quad \theta_{\mathbf{r}}^{(\mathbf{n})} = 2\pi \frac{\mathbf{r}}{\mathbf{n}}, \quad F_{\mathbf{n}} = \frac{1}{\sqrt{N(\boldsymbol{n})}} \left(\mathrm{e}^{-\iota \left\langle \mathbf{j}, \theta_{\mathbf{r}}^{(\mathbf{n})} \right\rangle} \right)_{\mathbf{j}, \mathbf{r}=\mathbf{0}}^{\mathbf{n}-\mathbf{e}}, \tag{I.19}$$

with $\left\langle \mathbf{j}, \theta_{\mathbf{r}}^{(\mathbf{n})} \right\rangle = \sum_{t=1}^{k} 2\pi \frac{j_t r_t}{n_t}$. *Here* $S_{\mathbf{n}}(f)(\cdot)$ *is the* $\mathbf{n}$*th Fourier sum of* f *given by*

$$\left(S_{\mathbf{n}}(f)\right)(\boldsymbol{\theta}) = \sum_{j_1=1-n_1}^{n_1-1} \cdots \sum_{j_k=1-n_k}^{n_k-1} \hat{f}_{\mathbf{j}} \mathrm{e}^{\iota \langle \mathbf{j}, \boldsymbol{\theta} \rangle}, \qquad \langle \mathbf{j}, \boldsymbol{\theta} \rangle = \sum_{t=1}^{k} j_t \theta_t. \tag{I.20}$$

Here $F_{\mathbf{n}}$ *is the* k*-level Fourier matrix,* $F_{\mathbf{n}} = F_{n_1} \otimes \cdots \otimes F_{n_k}$*, and its columns are the eigenvectors of* $C_{\mathbf{n}}(f)$ *with eigenvalues given by the evaluations of the* $\mathbf{n}$*th Fourier sum* $S_{\mathbf{n}}(f)(\cdot)$ *at the grid points*

$$\theta_{\mathbf{r}}^{(\mathbf{n})} = 2\pi \frac{\mathbf{r}}{\mathbf{n}}.$$

In the case where $\mathbf{f}$ is a Hermitian matrix-valued function, the previous theorem can be extended as follows:

Theorem I.6.2 ([78])**.** *Let* $\boldsymbol{f} \in L^1(k, s)$ *be a matrix-valued function with* $k \geq 1, s \geq 2$*. Then, the following block-Schur decomposition of* $C_{\mathbf{n}}(\boldsymbol{f})$ *is valid:*

$$C_{\mathbf{n}}(\boldsymbol{f}) = (F_{\mathbf{n}} \otimes I_s) D_{\mathbf{n}}(\boldsymbol{f}) (F_{\mathbf{n}} \otimes I_s)^*, \tag{I.21}$$

where

$$D_{\mathbf{n}}(\boldsymbol{f}) = \underset{\mathbf{0} \leq \mathbf{r} \leq \mathbf{n}-\mathbf{e}}{\operatorname{diag}} \left(S_{\mathbf{n}}(\boldsymbol{f})\right)\left(\theta_{\mathbf{r}}^{(\mathbf{n})}\right), \quad \theta_{\mathbf{r}}^{(\mathbf{n})} = 2\pi \frac{\mathbf{r}}{\mathbf{n}}, \quad F_{\mathbf{n}} = \frac{1}{\sqrt{N(\boldsymbol{n})}} \left(\mathrm{e}^{-\iota \left\langle \mathbf{j}, \theta_{\mathbf{r}}^{(\mathbf{n})} \right\rangle}\right)_{\mathbf{j},\mathbf{r}=\mathbf{0}}^{\mathbf{n}-\mathbf{e}}, \tag{I.22}$$

with $\left\langle \mathbf{j}, \theta_{\mathbf{r}}^{(\mathbf{n})} \right\rangle = \sum_{t=1}^{k} 2\pi \frac{j_t r_t}{n_t}$ *and* I_s *the* $s \times s$ *identity matrix. Here* $S_{\mathbf{n}}(\boldsymbol{f})(\cdot)$ *is the* $\mathbf{n}$*th Fourier sum of* $\boldsymbol{f}$ *given by*

$$\left(S_{\mathbf{n}}(\boldsymbol{f})\right)(\boldsymbol{\theta}) = \sum_{j_1=1-n_1}^{n_1-1} \cdots \sum_{j_k=1-n_k}^{n_k-1} \hat{\boldsymbol{f}}_{\mathbf{j}} \mathrm{e}^{\iota \langle \mathbf{j}, \boldsymbol{\theta} \rangle}, \qquad \langle \mathbf{j}, \boldsymbol{\theta} \rangle = \sum_{t=1}^{k} j_t \theta_t. \tag{I.23}$$

Here the eigenvalues of $C_{\mathbf{n}}(\boldsymbol{f})$ *are given by the evaluations of* $\lambda^{(t)}(S_{\mathbf{n}}(\boldsymbol{f}))$*,* $t = 1, \ldots, s$*, at the grid points*

$$\theta_{\mathbf{r}}^{(\mathbf{n})} = 2\pi \frac{\mathbf{r}}{\mathbf{n}}.$$

Definition I.6.2. *We say that a continuous* 2π*-periodic function* $\boldsymbol{f}$ *belongs to the Dini-Lipschitz class if its modulus of continuity evaluated at* δ *goes to zero faster than* $1/|\log(\delta)|$*, that is*

$$\lim_{\delta \to 0^+} \log(\delta) \omega_{\boldsymbol{f}}(\delta) = 0.$$

Remark 3. *If* $\boldsymbol{f}$ *is a trigonometric polynomial of fixed degree (with respect to* $\mathbf{n}$*), then it is worth noticing that* $S_{\mathbf{n}}(\boldsymbol{f})(\cdot) = \boldsymbol{f}(\cdot)$ *for* $\mathbf{n}$ *large enough: more precisely, every* n_j *should be larger than the double of the degree with respect to the* j*th variable. Therefore, in such a setting, the eigenvalues of* $C_{\mathbf{n}}(\boldsymbol{f})$ *are either the evaluations of* f *at the grid points if* $s = 1$ *or the quantity* $\lambda_t(\boldsymbol{f}(\cdot))$*,* $t = 1, \ldots, s$*, evaluated at the very same grid points. It is worth noting that we write* $\lambda_t(\boldsymbol{f}(\cdot))$*, instead of* $\left(\lambda^{(t)}(\boldsymbol{f})\right)(\cdot)$*, pointing out that we first calculate the matrices*

$$\boldsymbol{f}\left(\theta_{\mathbf{r}}^{(\mathbf{n})}\right), \quad \mathbf{0} \leq \mathbf{r} \leq \mathbf{n} - \mathbf{e},$$

and then their eigenvalues

$$\lambda_t\left(\boldsymbol{f}\left(\theta_{\mathbf{r}}^{(\mathbf{n})}\right)\right), \quad t = 1, \ldots, s.$$

A more detailed discussion on the evaluation of the eigenvalue functions of a matrix-valued symbol will treated in Subsections II.2.2.1, II.2.5, for $k > 1$*, and Section V.3, for* $k = 1$*.*

Remark 4. *Though the eigenvalues of any $C_{\mathbf{n}}(\boldsymbol{f})$ are explicitly known, results like Theorem I.4.1 and Theorem I.4.2 do not hold for sequences $\{C_{\mathbf{n}}(\boldsymbol{f})\}_{\mathbf{n}\in\mathbb{N}^k}$ in full generality: this is due to the fact that the Fourier sum of $\boldsymbol{f}$ converges to $\boldsymbol{f}$ under quite restrictive assumptions (see [154]). In fact if $\boldsymbol{f}$, belongs to the Dini-Lipschitz class, then $\{C_{\mathbf{n}}(\boldsymbol{f})\}_{\mathbf{n}\in\mathbb{N}^k} \sim_\lambda (f, \mathcal{I}_k)$, simply because $S_{\mathbf{n}}(\boldsymbol{f})(\cdot)$ uniformly converges to $\boldsymbol{f}$ (see [64] for more relations between circulant sequences and spectral distribution results).*

We end this subsection by recalling the computational properties of (block) circulants. Every matrix/vector operation with circulants has cost $O(N(\mathbf{n})\log N(\mathbf{n}))$ with moderate multiplicative constants: in particular, this is true for the matrix-vector product, for the solution of a linear system, for the computation of the blocks $(S_{\mathbf{n}}(\mathbf{f}))\left(\theta_{\mathbf{r}}^{(\mathbf{n})}\right)$ and consequently of the eigenvalues (see, e.g., [148]).

I.7 GLT sequences: operative features

We briefly present the class of Generalized Locally Toeplitz (GLT) sequences and its operative features, which will be the pivotal tools used in the next chapters (see the pioneering work [141] by Tilli for describing the spectrum of one-dimensional differential operators and the generalizations in [125, 126] by Serra-Capizzano for multivariate differential operators).

Going through the details of GLT class requires rather technical tools and is not within the aims of this thesis, hence here we list some properties of the GLT class in their multilevel block form (see [126]), which are sufficient to our purposes.

GLT1 Each GLT sequence has a singular value symbol $\mathbf{f}(\mathbf{x},\boldsymbol{\theta})$ for $(\mathbf{x},\boldsymbol{\theta}) \in [0,1]^k \times [-\pi,\pi]^k$ according to the second Item in Definition I.3.4 with $\ell = 2k$. If the sequence is Hermitian, then the distribution also holds in the eigenvalue sense. If $\{G_{\mathbf{n}}\}_{\mathbf{n}}$ has a GLT symbol $\mathbf{f}(\mathbf{x},\boldsymbol{\theta})$ we will write $\{G_{\mathbf{n}}\}_{\mathbf{n}} \sim_{\text{GLT}} \mathbf{f}(\mathbf{x},\boldsymbol{\theta})$.

GLT2 The set of GLT sequences form a $*$-algebra, i.e., it is closed under linear combinations, products, inversion (whenever the symbol is singular, at most, in a set of zero Lebesgue measure), conjugation. Hence, the sequence obtained via algebraic operations on a finite set of given GLT sequences is still a GLT sequence and its symbol is obtained by performing the same algebraic manipulations on the corresponding symbols of the input GLT sequences.

GLT3 Every Toeplitz sequence generated by an $L^1(k,s)$ function $\mathbf{f} = \mathbf{f}(\boldsymbol{\theta})$ is a GLT sequences and its symbol is $\mathbf{f}$, with the specifications reported in Item **GLT1**. We note that the function $\mathbf{f}$ does not depend on the space variables $\mathbf{x} \in [0,1]^k$.

GLT4 Every sequence which is distributed as the constant zero in the singular value sense is a GLT sequence with symbol 0 (in particular every sequence in which the rank divided by the size tends to zero as the matrix size tends to infinity).

In short, GLT sequences form an algebra containing sequences of matrices including the Toeplitz sequences with Lebesgue integrable symbols and virtually any sequence of matrices coming from "reasonable" approximations by local discretization methods (Finite Differences, Finite Elements, Isogeometric Analysis, etc.) of Partial Differential Equations.

I.8 Preconditioning and multigrid methods for Toeplitz matrices

We recall that often the approximation of a problem in an infinite dimensional space produces a sequence of large linear system

$$\{A_{\mathbf{n}}\mathbf{u_n} = \mathbf{b_n}\}_{\mathbf{n}},$$

of size $N \equiv N(\mathbf{n}, s)$, where $A_{\mathbf{n}}$ are structured matrices. If high precision is required, then we have to compute the numerical solution $\mathbf{u_n}$ for a large value of $\mathbf{n}$. Indeed the larger the dimension $\mathbf{n}$ is, the more accurate the solution will be.

In these cases the direct methods can be unstable and often are too costly since they do not exploit the structure of the coefficient matrices, conversely the use of iterative methods is recommended because of the memory and accuracy requirements.

The goal is to choose the resolution methods which are *optimal*. Here we give a definition of optimality for iterative methods applied to sequences of linear systems [6, 123].

Definition I.8.1. ***[Optimality]*** *Given a sequence*

$$\{A_{\boldsymbol{n}}\mathbf{u}_{\boldsymbol{n}} = \mathbf{b}_{\boldsymbol{n}}\}_{\boldsymbol{n}} \tag{I.24}$$

of linear systems of size $N \equiv N(\mathbf{n}, s)$, an iterative method is said to be optimal if

1) *its cost for computing the solution is proportional to that of the matrix–vector product;*

2) *the number of iterations required for computing $\mathbf{u}_{\boldsymbol{n}}$ within a preassigned accuracy ϵ is bounded by a constant independent of $\boldsymbol{n}$ and possibly depending on ϵ*

In case of Toeplitz structures the most popular iterative solvers (Conjugate Gradient (CG), Conjugate Gradient for Least Squares (CGLS), Generalized Minimal Residual (GMRES), etc.) satisfy the first requirement.

Conversely the second Item is the critical point. In most cases, the condition number of $A_{\mathbf{n}}$ diverges to infinity quickly as $\mathbf{n}$ increases (for example, if the minimal eigenvalue $\lambda_1(A_{\mathbf{n}})$ tends to zero as $\mathbf{n}$ tends to infinity). In such situations the classical iterative methods can be very slow. Indeed it is well known that their convergence rate depends on the condition number of the coefficient matrix and on how the spectrum of $A_{\mathbf{n}}$ is clustered.

Preconditioning With regard to this feature, (when $A_{\mathbf{n}}$ is an HPD matrix) one of the most successful iterative solvers is the preconditioned conjugate gradient (PCG) method. The use of a preconditioner $P_{\mathbf{n}}$ can accelerate the convergence by reducing the number of steps required for the convergence.

Hence, instead of solving the problems (I.24), we deal with the preconditioned systems

$$\{P_{\mathbf{n}}^{-1}A_{\mathbf{n}}\mathbf{u_n} = P_{\mathbf{n}}^{-1}\mathbf{b_n}\}_{\mathbf{n}}. \tag{I.25}$$

In particular when the eigenvalues/singular values of $P_{\mathbf{n}}^{-1}A_{\mathbf{n}} - I_{\mathbf{n}}$ are strongly clustered at zero or when the sequence of the spectral condition numbers $\kappa(P_{\mathbf{n}}^{-1}A_{\mathbf{n}})$ of $\{P_{\mathbf{n}}^{-1}A_{\mathbf{n}}\}_{\mathbf{n}}$ is upper-bounded by a constant independent of $\mathbf{n}$, we know [5] that a constant number of iterations are

required for the convergence of the PCG method. In particular, if $\{P_{\mathbf{n}}^{-1}A_{\mathbf{n}} - I_{\mathbf{n}}\}_{\mathbf{n}}$ is strongly clustered at zero and $\{P_{\mathbf{n}}^{-1}A_{\mathbf{n}}\}_{\mathbf{n}}$ is spectrally bounded, then the PCG method with preconditioner $P_{\mathbf{n}}$ is optimal and the convergence is superlinear. Consequently the preconditioner $P_{\mathbf{n}}$ should be chosen in order to balance the following two requirements.

a) The solution of a generic system $P_{\mathbf{n}}\mathbf{y}_{\mathbf{n}} = \mathbf{c}_{\mathbf{n}}$ has computational cost bounded by vector-product with matrix $A_{\mathbf{n}}$;

b) $\kappa(P_{\mathbf{n}}^{-1}A_{\mathbf{n}})$ is upperbounded by a constant independent of $\mathbf{n}$ (that is $\{P_{\mathbf{n}}\}_{\mathbf{n}}$ is "close" to $\{A_{\mathbf{n}}\}_{\mathbf{n}}$ in spectral norm) or $\{P_{\mathbf{n}}^{-1}A_{\mathbf{n}} - I_{\mathbf{n}}\}_{\mathbf{n}}$ is strongly clustered at 0 (that is $\{P_{\mathbf{n}}\}_{\mathbf{n}}$ is "close" to $\{A_{\mathbf{n}}\}_{\mathbf{n}}$ in the clustering sense).

The issues a) and b) are often conflicting since when a preconditioner $P_{\mathbf{n}}$ is too close to $A_{\mathbf{n}}$ (the requirement in b)) it also requires the same computational cost as to invert (hence contradicting the requirement in a)).

However, in the context of structured matrices of Toeplitz type many satisfactory solutions can be found (see [32, 96, 120] and references therein). One of the possibilities is to look for preconditioners within matrix algebras such as the circulant class.

This choice automatically satisfies requirement a). Indeed the computational cost of a matrix-vector product when a circulant matrix is involved is proportional to $O(n\log(n))$ (or $O\left(N(\mathbf{n},s)\log(N(\mathbf{n},s))\right)$ in the block multilevel case) and can be achieved by using the Fast Fourier Transform (FFT) [148].

In applications, the condition in b) may not be easily verifiable. However, in the scalar setting the classical one level circulant preconditioners proposed by Strang [133] and T. Chan [36] gives a superlinear convergence under suitable assumptions on the generating function. In [33] authors proved that if $T_n(f) = [\hat{f}_{i-j}]_{i,j=1}^{n}$ is a $n \times n$ Toeplitz matrix and it is associated with an absolutely convergent Fourier series for a positive generating function, $f(\theta) = \sum_{l=-\infty}^{\infty} \hat{f}_l \mathrm{e}^{\imath l\theta}$, then the following circulant preconditioners [36, 133] provide a superlinear convergence. More precisely:

- the Strang preconditioner is the matrix S_n obtained by copying the central diagonals of $T_n(f)$ and by incorporating the remainders in order to complete the circulant structure of S_n. Specifically, the diagonals s_i of S_n are given by

$$s_i = \begin{cases} \hat{f}_i, & 0 < i \leq \lfloor \frac{n}{2} \rfloor; \\ \hat{f}_{i-n}, & \lfloor \frac{n}{2} \rfloor < i < n; \\ s_{n+i}, & 0 < -i < n. \end{cases}$$

  The superlinear convergence of PCG with the Strang preconditioner is guaranteed whenever f belongs to the Dini–Lipschitz class [122].

- The Chan preconditioner is the matrix C_n defined as

$$\underset{C_n \text{ circulant}}{\arg\min} \|T_n(f) - C_n\|_F = \underset{C_n = F_n D_n F_n^* \text{ circulant}}{\arg\min} \|F_n^* T_n(f) F_n - D_n\|_F,$$

where D_n is diagonal and F_n is the Fourier matrix of size n.

Specifically, the entries c_i of C_n are given by

$$c_i = \begin{cases} \frac{i\hat{f}_{-(n-i)}+(n-i)\hat{f}_i}{n}, & 0 \le i < n; \\ c_{n+i}, & 0 < -i < n. \end{cases}$$

If f is a positive continuous generating function then the superlinear convergence of PCG with the Tony Chan preconditioner is ensured [35, 120, 122].

Unfortunately in the k-level context, the construction of circulant preconditioners similar to those of the unilevel case leads only to sublinear preconditioners, even for the well-conditioned matrices.

The performances of multilevel circulant preconditioners indeed deteriorate as k increases and in fact it is proved in [98, 124, 129] that the use of any multilevel circulant preconditioner permits us to reach at most a sub-optimal convergence of the PCG method.

Multigrid methods The first important remark is that the requirement in b) can be overpassed by using a multigrid technique, also in the multilevel setting.

Under suitable assumptions, these methods are optimal and their excellent features are identical in the multilevel setting. Furthermore they are optimal also for polynomially ill-conditioned multidimensional problems and can be extended to the case of low-rank corrections of the considered structured matrices, so that the computational barriers holding in the preconditioning setting [2, 3, 127] do not hold.

Here we briefly sketch the basic ideas for defining the classical multigrid methods (MGM). Firstly we focus on the general scalar unilevel matrices, then we give the main MGM convergence results for scalar Toeplitz matrices with a scalar valued symbol.

In Section II.3.4 the following strategy is generalized to deal with the block multilevel matrix sequence associated with a multivariate matrix-valued generating function.

The basic idea of a multigrid method is to create a sequence of linear systems of decreasing dimensions by consecutive projections. In this way the computational cost is reduced at each level and the convergence speed can be improved. Here, for MGM algorithm, we mean the simplest and less expensive version of the large family of multigrid methods and which named V-cycle procedures. In particular, first we consider the method with only two levels, known as the Two Grid Method (TGM). Once that the TGM is introduced, the V-cycle algorithm is obtained recursively applying a projection strategy.

An important choice for the TGM concerns the prolongation/restriction operators. When deriving convergence estimates, usually the restriction is chosen to be the adjoint of the prolongation. These conditions are known in the related literature as Galerkin conditions and the resulting method is the so-called algebraic multigrid (AMG).

We start from a linear system

$$A_n x_n = b_n, \tag{I.26}$$

where x_n, $b_n \in \mathbb{C}^n$, $A_n = \mathcal{W}_n - \mathcal{B}_n \in \mathbb{C}^{n\times n}$, $\mathcal{W}_n$ is a non singular matrix. Let

$$x^{(j+1)} = V_n(x^{(j)}, b_1) = V_n x^{(j)} + b_1 \tag{I.27}$$

be an iterative method for the solution of system (I.26), where $b_1 = \mathcal{W}_n^{-1} b \in \mathbb{C}^n$ and $V_n = I_n - \mathcal{W}_n^{-1} A_n \in \mathbb{C}^{n\times n}$. Given a full-rank matrix $p_n^m \in \mathbb{C}^{n\times m}$, with $m < n$, a Two-Grid Method (TGM) is defined by the following algorithm [145]

1. $d_n = A_n x_n^{(j)} - b_n$
2. $d_m = (p_n^m)^* d_n$
3. $A_m = (p_n^m)^* A_n (p_n^m)$
4. Solve $A_m y = d_m$
5. $\hat{x}^{(j)} = x^{(j)} - p_n^m y$
6. $x^{(j+1)} = V_n^\nu(\hat{x}^{(j)}, b_1)$

Step 6 consists in applying the "smoothing iteration" (I.27) ν times while steps 1-5 define the "coarse grid correction", that depends only on the prolongation operator p_n^m.

The global iteration matrix of the TGM is given by

$$TGM(V_n, p_n^m) = V_n^\nu \left[I - p_n^m \left((p_n^m)^* A_n p_n^m \right)^{-1} (p_n^m)^* A_n \right],$$

which implies that the TGM can be seen as a classical stationary iteration technique.

A possible "pre-smoothing iteration" can be performed before step 1. If step 4 is replaced by a recursive call to the same algorithm, then the scheme given before defines a V-cycle procedure. Note that in the AMG the coarse grid matrix $A_{n_{i+1}}$ at the level $(i+1)$ is chosen as $(P_{n_i})^* A_{n_i} P_{n_i}$, where P_{n_i} is the prolongation operator at level i.

In the following we illustrate the result concerning the convergence and the optimality of the TGM [106].

Theorem I.8.1. *Let A_n be a positive definite matrix of size n and let V_n be defined as in the* TGM *algorithm. Assume*

(a) $\exists \alpha_{\text{post}} > 0 : \|V_n x_n\|_{A_n}^2 \le \|x_n\|_{A_n}^2 - \alpha_{\text{post}} \|x_n\|_{A_n^2}^2, \qquad \forall x_n \in \mathbb{C}^n,$

(b) $\exists \gamma > 0 : \min_{y\in\mathbb{C}^m} \|x_n - p_n^m y\|_2^2 \le \gamma \|x_n\|_{A_n}^2, \qquad \forall x_n \in \mathbb{C}^n.$

Then $\gamma \ge \alpha_{\text{post}}$ and

$$\|\text{TGM}(I, V_n^{\nu_{\text{post}}}, p_n^m)\|_{A_n} \le \sqrt{1 - \alpha_{\text{post}}/\gamma}.$$

Note that, since α_{post} and γ are independent from n, if the assumptions of Theorem I.8.1 are satisfied, the number of iterations, in order to reach a given accuracy ϵ, can be bounded from above by a constant independent of n.

We refer to (a) and (b) as the "smoothing" and the "approximation" properties, respectively. Indeed the condition (a) is related only to the smoothers, conversely the assumption (b) depends only on the choice of the projector. Hence the two requirements can be treated separately and the latter represents a substantial simplification for studying the convergence analysis.

We are interested in the case where $T_{\mathbf{n}}(\mathbf{f})$ is a multilevel block Toeplitz matrix associated with a matrix-valued trigonometric polynomial.

Many theoretical results have been provided on the validation of the smoothing property in the one (and multi) level scalar case [2, 3, 128]. Furthermore the theory can be easily extended to the block context. Conversely, up to our knowledge, the generalization in block settings of the theorems concerning the approximation property for scalar matrices [2, 3] are still missing or under investigation [130]. Here the principal open question concerns the lack of commutative property for matrix-valued symbols.

In what follows we report the results for the validation of both the smoothing and the approximation conditions in the scalar multilevel case. We postpone the discussion on the block case to Section II.3.4, where we construct the appropriate smoother and prolongation operator, exploiting the Laplacian nature of the problem.

We assume that the matrices $T_{\mathbf{n}}(f)$ are such that $\mathbf{n} = (2^t - 1)\mathbf{e} \in \mathbb{N}^k$, with t positive integer (the case $\mathbf{n} = 2^t\mathbf{e} \in \mathbb{N}^k$ is analogous [3, 69]).

Theorem I.8.2. *[128] Let $T_{\boldsymbol{n}}(f)$ be a multilevel Toeplitz matrix associate to a k-variate generating function $f : \mathcal{I}_k \to \mathbb{C}$ nonnegative and not identically zero and define $V_{\boldsymbol{n}} = I_{\boldsymbol{n}} - \omega T_{\boldsymbol{n}}(f)$. If we choose α_{post} such that $0 \le \alpha_{\text{post}} \le \frac{2}{\|f\|_\infty}$, then relation (a) in Theorem I.8.1 holds.*

This theorem can be possibly generalized when considering both pre-smoothing and post-smoothing as in [2].

The other important choice for the TGM convergence concerns the operator $p_{\mathbf{n}}^{\mathbf{m}} \in \mathbb{C}^{N(\mathbf{n})\times N(\mathbf{m})}$, with $\mathbf{m} < \mathbf{n}$.

In AMG the procedure starts with $\mathbf{n}_0 = \mathbf{n}$ and the indices in the coarse levels are defined as $\mathbf{n}_i = (2^{t-i} - 1)\mathbf{e}$, such that $\mathbf{n}_i > \mathbf{n}_{i+1}$.

The matrix $p_{\mathbf{n}}^{\mathbf{m}}$ has a double role. On one hand it projects the problem into a coarse one, "cutting" the matrix $T_{\mathbf{n}}(f)$, on the other hand it should maintain the same structure and the properties of $T_{\mathbf{n}}(f)$ in the "cut" and projected matrix $(p_{\mathbf{n}}^{\mathbf{m}})^* T_{\mathbf{n}}(f) p_{\mathbf{n}}^{\mathbf{m}}$. Therefore it is chosen as the product between $T_{\mathbf{n}}(p)$, with p non negative trigonometric polynomial, and a cutting matrix $Z_{\mathbf{n}}^{\mathbf{m}} \in \mathbb{C}^{N(\mathbf{n})\times N(\mathbf{m})}$. That is the projector has the form

$$p_{\mathbf{n}}^{\mathbf{m}} = T_{\mathbf{n}}(p) Z_{\mathbf{n}}^{\mathbf{m}}, \tag{I.28}$$

where

$$Z_{\mathbf{n}}^{\mathbf{m}} = Z_{n_1}^{m_1} \otimes Z_{n_2}^{m_2} \otimes \cdots \otimes Z_{n_k}^{m_k},$$

and $Z_{n_l}^{m_l}$ is the $n_l \times m_l$ matrix given by

$$(Z_{n_l}^{m_l})_{i,j} = \begin{cases} 1 & \text{for } i = 2j \\ 0 & \text{otherwise} \end{cases} \qquad i = 1, \ldots, n_l, \quad j = 1, \ldots, m_l. \tag{I.29}$$

Then the matrix at the coarse level $T_{\mathbf{m}}(\tilde{f}) = (p_{\mathbf{n}}^{\mathbf{m}})^* T_{\mathbf{n}}(f) p_{\mathbf{n}}^{\mathbf{m}}$ is still a Toeplitz matrix, up to a lower rank correction, where

$$\tilde{f}(\boldsymbol{\theta}) = \frac{1}{2}\left[p^2 f(\boldsymbol{\theta}/2) + p^2 f(\boldsymbol{\theta}/2 + \pi)\right].$$

In our setting the correction is not present.

In the following we show how the polynomial p should be chosen in order to ensure the validity of the approximation property (b). For a fixed $\boldsymbol{\theta} \in \mathbb{R}^k$, we define the sets $\Omega(\boldsymbol{\theta})$ and $\mathcal{M}(\boldsymbol{\theta})$ of the all *corner* and *mirror* points respectively, that are

$$\Omega(\boldsymbol{\theta}) = \{\boldsymbol{\eta} = (\eta_1, \ldots, \eta_k) \,:\, \eta_j \in \{\theta_j, \theta_j + \pi\}\}, \qquad \mathcal{M}(\boldsymbol{\theta}) = \Omega(\boldsymbol{\theta}) \setminus \{\boldsymbol{\theta}\}.$$

Theorem I.8.3. *Let $A_{\boldsymbol{n}} = T_{\boldsymbol{n}}(f)$ with $\boldsymbol{n} = (2^t - 1)\boldsymbol{e}$, f a k variate nonnegative trigonometric polynomial and let $\boldsymbol{m} = (m_1, \ldots, m_k) < \boldsymbol{n} = (n_1, \ldots, n_k)$. Let $\boldsymbol{\theta}^0 = (\theta_1^0, \ldots, \theta_k^0)$ be the unique zero of f in $\mathcal{I}_k$ of order at most 2, and let $p_{\boldsymbol{n}}^{\boldsymbol{m}}$ be defined as in (I.28) with p of the form*

$$p(\boldsymbol{\theta}) = c \prod_{j=1}^{k} (1 + \cos(\theta_j - \theta_j^0)),$$

with c constant. Then the approximation property (b) holds if, for all $\boldsymbol{\theta} \in \mathcal{I}_k$, p verifies

$$\begin{gathered} \limsup_{\boldsymbol{\theta} \to \boldsymbol{\theta}^0} \frac{|p(\boldsymbol{\eta})|^2}{f(\boldsymbol{\theta})} < \infty, \quad \boldsymbol{\eta} \in \mathcal{M}(\boldsymbol{\theta}), \\ \sum_{\boldsymbol{\eta} \in \Omega(\boldsymbol{\theta})} p^2(\boldsymbol{\eta}) > 0. \end{gathered} \tag{I.30}$$

Remark 5 ([128]). *Let A_n and B_n be two Hermitian positive definite matrices, with*

$$A_n \leq \theta B_n,$$

for some positive θ independent of n.
If a TGM is optimal for A_n then the same algorithm is optimal also for B_n.

Hence if a proposed TGM is optimally convergent for a positive definite sequence $\{T_{\mathbf{n}}(f)\}_{\mathbf{n}}$, then the same smoother and projector provide optimality when considering the definite positive sequence $\{K_N = T_{\mathbf{n}}(f) + E_{\mathbf{n}}\}_{\mathbf{n}}$, with $E_{\mathbf{n}}$ nonnegative definite matrix. In the context of **Chapter II** we will see that $E_{\mathbf{n}}$ is a nonnegative definite small rank correction of $T_{\mathbf{n}}(f)$.

I.9 Asymptotic Expansion: idea of the approximation errors

In this section we consider the problem of computing the spectrum of banded symmetric Toeplitz matrices. In the next chapters, for such type of problems, we develop a class of fast methods starting from the results in the recent work [62], where Ekström, Garoni, and Serra-Capizzano conjecture the existence of an asymptotic spectral expansion for this class of matrices.

We illustrate the preliminary version of the proposed expansion, which will be generalized and used for computing the eigenvalues of large Toeplitz-like matrices in various contexts.

It must be highlighted that, independently from [62], Bogoya, Böttcher, Grudsky, and Maximenko in [16, 17, 19] have obtained the precise asymptotic expansion for the eigenvalues of a sequence of Toeplitz matrices $\{T_n(f)\}_n$, under suitable assumptions on the associated generating function f.

However, in [62] the authors provided numerical evidences that some of those assumptions can be relaxed, maintaining only the hypothesis on f of being a real, cosine trigonometric polynomial **(RCTP)**, monotone on the domain.

In [62] it was conjectured and numerically confirmed that, if f is a monotone RCTP on $\mathcal{I}_1^+$, then, for every integer $\alpha \geq 0$, every n and every $j = 1, \ldots, n$, the following asymptotic expansion holds:

$$\lambda_j(T_n(f)) = f(\theta_{j,n}) + \sum_{k=1}^{\alpha} c_k(\theta_{j,n})h^k + E_{j,n,\alpha}, \tag{I.31}$$

where:

- the eigenvalues of $T_n(f)$ are arranged in non decreasing or non increasing order, depending on whether f is increasing or decreasing;
- $\{c_k\}_{k=1,2,\ldots}$ is a sequence of functions from $[0, \pi]$ to $\mathbb{R}$ which depends only on f;
- $h = \frac{1}{n+1}$ and $\theta_{j,n} = \frac{j\pi}{n+1} = j\pi h$;
- $E_{j,n,\alpha} = O(h^{\alpha+1})$ is the remainder (the error), which satisfies the inequality $|E_{j,n,\alpha}| \leq C_\alpha h^{\alpha+1}$ for some constant C_α depending only on α and f.

The idea under the expansion I.31 is based on the study of the approximation errors

$$E_{j,n} = \lambda_j(A_n) - f(\theta_{j,n}), \tag{I.32}$$

that arise when $\{A_n\}_n$ is a (special) matrix sequence such that $\{A_n\}_n \sim_{\text{GLT},\sigma,\lambda} f$ and $\theta_{j,n}$ is a suitable uniform grid.

Assume we want to study the spectral properties of the matrix sequences $\{B_n\}_n$ and $\{T_n(g)\}_n$, where for a fixed n, $T_n(g)$ is the Toeplitz matrix generated by

$$g(\theta) = f(\theta)^2 = (2 - 2\cos(\theta))^2 = 6 - 8\cos(\theta) + 2\cos(2\theta).$$

and $B_n = (T_n(f))^2$. The matrix B_n is the discretized bi-Laplacian, in the sense that we apply twice the discretized Laplace operator by second order finite differences, and was proved [125] that $g(\theta)$ is its GLT symbol, that is $\{B_n\}_n \sim_{\text{GLT}} g$. Since B_n is in addition Hermitian we also have that $\{B_n\}_n \sim_{\sigma,\lambda} g$.

The sequence $\{T_n(g)\}_n$ is a banded Toeplitz matrix sequence generated by the trigonometric polynomial $g(\theta)$, hence $\{T_n(g)\}_n \sim_{\text{GLT}\sigma,\lambda} g$.

The matrix B_n is equal to $T_n(g)$ except for a *low-rank correction* R_n (in this case of rank 2),

$$R_n = B_n - T_n(g) = -\mathbf{e}_1\mathbf{e}_1{}^T - \mathbf{e}_n\mathbf{e}_n{}^T.$$

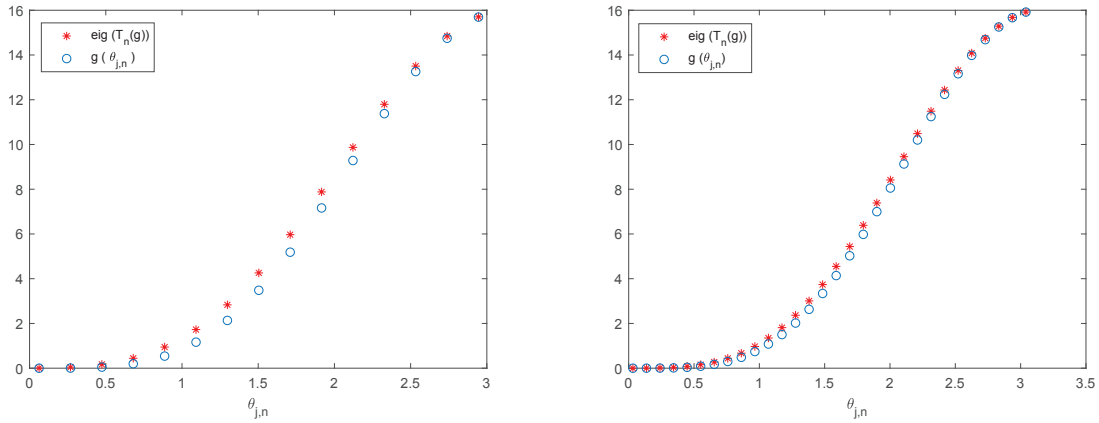
The low-rank correction sequence $\{R_n\}_n$ is zero-distributed, $\{R_n\}_n \sim_\lambda 0$, according to the definition in Section I.3. This means that as $n \to \infty$ the eigenvalues of B_n and $T_n(g)$ will coincide. However, the finite-dimensional matrices have different eigenvalues.

In particular sampling $g(\theta)$ with the grid of the τ algebra

$$\theta_{j,n} = \frac{j\pi}{n+1}, \quad j = 1, \ldots, n, \tag{I.33}$$

returns the exact eigenvalues of B_n, that is $\lambda_j(B_n) = g(\theta_{j,n})$. Conversely the eigenvalues of $T_n(g)$ are not exactly given by the sampling of g on grid (I.33). The Figure I.1 indeed

confirms that the uniform sampling of g (in blue circles $\circ$) provides just an approximation of the eigenvalues of $T_n(g)$ (in red stars $*$). Furthermore, as n grows the more accurate the approximation will be. For example, we take $n = 15$ in Figure I.1(a) and $n = 30$ in Figure I.1(b). Hence, for $n \to \infty$ the eigenvalues $T_n(g)$ will coincide with the evaluations of g, but for finite n we do not know the explicit grid $\theta_{j,n}$ which yields $E_{j,n} = \lambda_j(T_n(g)) - g(\theta_{j,n}) = 0$ for all j.



(a) $n = 15$ (b) $n = 30$

Figure I.1: Comparison between the eigenvalues of $T_n(g)$ (in red stars $*$) and the uniform sampling of g (in blue circles $\circ$) on the grid in (I.33). The parameter n equals 15 in subplots (a) and it is doubled, $n = 30$, in (b).

Thus we have

$$\lambda_j(T_n(g)) = g(\theta_{j,n}) + E_{j,n}, \quad j = 1, \ldots, n.$$

However, in Figures in I.2 we can observe that the errors $E_{j,n}$ have an interesting property. Indeed, when using the τ-grid for three different $n \in \{100, 200, 400\}$, the shape of the "error curve", is retained as n increases, see Figure I.2(a).

In addition the errors $E_{j,n}$ behave as expected, that is, they decrease linearly in n as n increases, equivalently they are of order $\mathcal{O}(h)$, where $h = 1/(n+1)$ for each n. Furthermore we can observe in Figure I.2(b) that the curves of the scaled errors $E_{j,n}/h = (\lambda_j(T_n(g)) - g(\theta_{j,n}))/h$ for $n = 20, 40, 80, 100$ overlap perfectly.

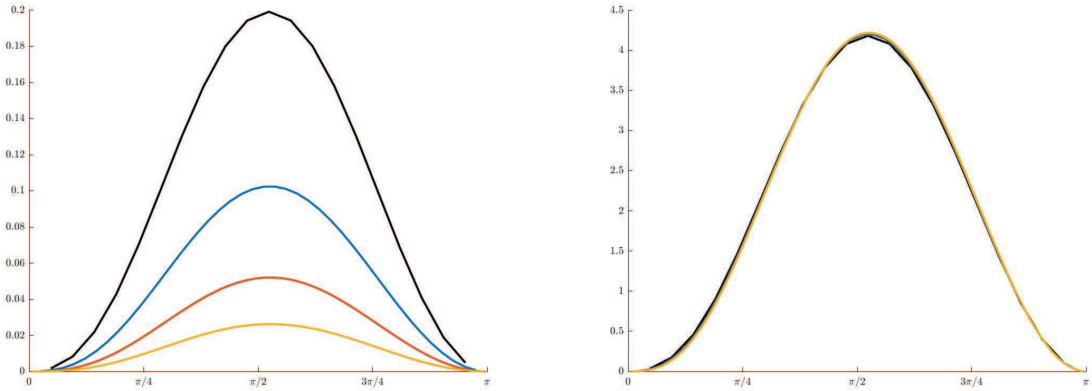
This behavior of the curves $E_{j,n}/h$ suggests that for g and other types of symbols there exists an asymptotic expansion of the error in (I.32) of the form

$$\lambda_j(T_n(g)) - g(\theta_{j,n}) = \sum_{k=1}^{\alpha} c_k(\theta_{j,n}) h^k + E_{j,n,\alpha}.$$

Note that, if $\alpha = 0$, then $E_{j,n} = \lambda_j(T_n(g)) - g(\theta_{j,n}) = E_{j,n,0}$. If $\alpha = 1$, then $E_{j,n}/h = c_1(\theta_{j,n}) E_{j,n,1}/h$, for each value of n. Indeed in Figure I.2(b) the four curves coincide since the scaled remainder $E_{j,n,1}/h$ is small and the function c_1 does not depend on n.

In the Sections VI.3, VI.4, VI.6 of the **Chapter VI** we present the proof of the first order asymptotic term of the expansion for

1. preconditioned banded symmetric Toeplitz matrices [1];



(a) $E_{j,n} = \lambda_j(T_n(g)) - g(\theta_{j,n}) = \mathcal{O}(h)$ (b) $E_{j,n}/h = (\lambda_j(T_n(g)) - g(\theta_{j,n}))/h = c_1$

Figure I.2: The errors $E_{j,n}$ (left) and the scaled errors $E_{j,n}/h$ (right) when approximating $\lambda_j(T_n(g))$ with the sampling $g(\theta_{j,n})$, $j = 1, \ldots, n$ for $n \in \{20, 40, 80, 160\}$

2. Toeplitz-like matrices, $n^{-2}L_n^{[p]}$, coming from the B-spline IgA approximation of $-u'' = \lambda u$ [58];

3. block and preconditioned block banded symmetric Toeplitz matrices [60].

In all the contexts the proof is based on the following common facts. If $\tau_n(f)$ is the τ matrix of size n generated by a monotone RCTP f of degree m (the case **f** monotone HTP is analogous), then

- $\tau_n(f)$ is a real symmetric matrix with eigenvalues given by $f(\theta_{j,n})$, $j = 1, \ldots, n$;
- the matrix $T_n(f)$ can be written as $T_n(f) = \tau_n(f) + H_n(f)$, where $H_n(f)$ is symmetric real Hankel matrix generated by f with $\nu = \text{rank}(H_n) \leq 2(m-1)$;
- from the classical Interlacing theorem for the eigenvalues (see [13] or [77]), it holds

$$f(\theta_{j-\nu,n}) \leq \lambda_j(T_n(f)) \leq f(\theta_{j+\nu,n}),\ j = \nu+1, \ldots, n-\nu;$$

- $\lambda_j(T_n(f)) \in (m_f, M_f)$, $j = 1, \ldots, n$, where $m_f = \min f < M_f = \max f$; see [20, 77].

In the Chapters III, IV, V, studying the errors of the approximation of eigenvalues by uniform sampling of the symbol, it is possible to devise an extrapolation–interpolation procedure for computing the eigenvalues of Toeplitz-like matrices of very large dimension. The resulting algorithm can be performed with a high level of accuracy and at the cost of the computation of the eigenvalues of a moderate number of small sized matrices.

We remark that in [7, 16, 17] it has been prove that if the symbol $f(\theta)$ does not comply with the simple-loop conditions, that is the requirment that $f'(\theta) \neq 0$ for $\theta \in (0, \pi)$ and $f''(\theta) \neq 0$ for $\theta \in \{0, \pi\}$, the expansion (I.31) will not be true point-wise for all eigenvalues. In practice when using standard double precision computations, in the next chapters we demonstrate why this is not a problem when using the proposed algorithm.

Chapter II

Spectral analysis on SDG methods for the incompressible Navier-Stokes equations

In this chapter we consider the incompressible Navier-Stokes equations approximated by a novel family of high order semi-implicit Discontinous Galerkin methods on *staggered meshes* (SDG) introduced in [65, 66, 135, 137]. These new schemes are analysed for the first time by means of GLT techniques and therefore the aim is to use and extend the spectral tools mentioned so far to the present numerical framework and to study its properties.

We recall that computational fluid dynamics (CFD) represents a vast sector of ongoing research in engineering and applied mathematics, which has also a wide applicability to real world problems, such as aerodynamics of airplanes and cars, geophysical flows in oceans, lakes and rivers, Tsunami wave propagation, blood flow in the human cardiovascular system, weather forecasting and many others. The governing equations for incompressible fluids are given by the incompressible Navier-Stokes equations that consist in a divergence-free condition for the velocity

$$\nabla \cdot \mathbf{v} = 0, \tag{II.1}$$

and a momentum equation that involves nonlinear convection, the pressure gradient and viscosity effects:

$$\frac{\partial \mathbf{v}}{\partial t} + \nabla \cdot \mathbf{F} + \nabla p = \nabla \cdot (\nu \nabla \mathbf{v}) \,. \tag{II.2}$$

Here, $\mathbf{v}$ is the velocity field; p is the pressure; ν is the kinematic viscosity coefficient and $\mathbf{F} = \mathbf{v} \otimes \mathbf{v}$ is the tensor containing the nonlinear convective term. The dynamics induced by equations (II.1)-(II.2) can be rather complex and have been observed in various experiments, see [4, 109, 150]. In the last decades a lot of effort was made to numerically solve the incompressible Navier-Stokes equations using finite difference schemes (see [83, 99, 100, 147]), continuous finite elements (see [24, 70, 85, 86, 91, 139, 149]) and more recently high order DG methods, see, e.g., [9, 10, 43, 68, 95, 97, 104, 105, 131]. The main difficulty in the numerical solution of the incompressible Navier-Stokes equations (II.1)-(II.2) lies in the elliptic pressure Poisson

equation and the associated linear equation system that needs to be solved. On the discrete level the pressure system is obtained by substitution of the discrete momentum equation (II.2) into the discrete form of the divergence-free condition (II.1). Since the solution of the incompressible Navier-Stokes equations requires necessarily the solution of large systems of algebraic equations, it is indeed very important to have a scheme that uses a stencil that is as small as possible, in order to improve the sparsity pattern of the resulting system matrix. It is also desirable to use methods that lead to reasonably well conditioned systems that can be solved with iterative solvers, like the conjugate gradient (CG) method [84] or the generalized minimal residual (GMRES) algorithm [108]. Very recently, a new class of arbitrary high order accurate semi-implicit DG schemes for the solution of the incompressible Navier-Stokes equations on structured, adaptive Cartesian and unstructured edge-based *staggered* grids was proposed in [65, 66, 135, 136, 137], following a philosophy that had been first introduced in finite difference schemes, see [25, 26, 27, 28, 29, 30, 83, 87, 99, 100, 147]. All those approaches have in common that the pressure is defined on a main grid, while the velocity field is defined on an appropriate edge-based staggered grid. The nonlinear convective terms are discretized explicitly by using a standard DG scheme based on the upwind flux or a local Lax-Friedrichs (Rusanov) flux [107]. Then, the discrete momentum equation is inserted into the discrete continuity equation in order to obtain the discrete form of the pressure Poisson equation. The advantage in using staggered grids is that they allow to improve significantly the sparsity pattern of the final linear system that has to be solved for the pressure. For the structured case the resulting main linear system is a sparse block penta-diagonal and hepta-diagonal one in two and three space dimensions, respectively. Furthermore, several desirable properties, such as the symmetry and the positive definiteness can be achieved see, e.g., [65, 137].

The main advantage of using an edge-based staggered grid is that the resulting matrix involves only the direct neighbors. For instance the total stencil in the three-dimensional Cartesian grid case is 13 for a collocated grid, 27 for a vertex-based staggering and it is only 7 for an edge-based staggered mesh. The edge-based staggered semi-implicit DG scheme therefore allows the use of the most compact stencil together with the minimum number of unknowns (only the scalar pressure). If one wants to achieve the same compact stencil on a collocated grid, a four times larger system needs to be solved, including the scalar pressure and the three components of the velocity vector.

Compared to classical continuous finite elements the discontinuous Galerkin method is known to handle dynamic adaptive mesh refinement (AMR) with hanging nodes [66, 153] as well as p-refinement very easily. It is also possible to deal with flow discontinuities in the boundary conditions, see, e.g., [65, 137], since boundary conditions are only imposed weakly.

The DG framework has also been very successfully applied in the past to high Mach number flows with shock waves, see, e.g., [39, 40, 56] for some examples and an overview of recent developments. The new class of staggered semi-implicit DG schemes analyzed in this chapter has very recently also successfully been extended to the fully compressible case [138], allowing to deal with *all Mach number flows*, ranging from nearly incompressible low Mach number flows to supersonic flows with shock waves.

The regular shape of Cartesian grids allows to further describe the structure of the main linear system for the pressure in the framework of multilevel block Toeplitz matrices: in this

setting we can deliver spectral and computational properties, including specific preconditioners for the original coefficient matrices and specific multigrid methods both for the preconditioning matrices and the coefficient matrices.

Main contributions

The main contributions of this Chapter can be summarized as follows.

1. We study the linear systems stemming from the considered approximations in a setting of structured linear algebra. These new schemes have never been analyzed with GLT techniques before and therefore our aim is to use and extend the spectral tools mentioned so far to this new numerical framework and to study its properties.

2. One of the main goal is the proof that these matrix sequences can be viewed as perturbations of matrices known in the literature, such as Toeplitz, and for which spectral studies already exist.

3. We detect the symbol associated to the coefficient matrix. This allows us to study the nonsingularity of the associated Toeplitz matrix sequence $\{T_{\mathbf{n}}(\mathbf{f})\}_{\mathbf{n}}$, together with information on the conditioning, the distribution of the spectrum, the behavior of the extremal eigenvalues and of the outliers.

4. The study is extended to the case of the global matrix sequence $\{K_N\}_{\mathbf{n}} = \{T_{\mathbf{n}}(\mathbf{f}) + \mathbf{E}_{\mathbf{n}}\}_{\mathbf{n}}$, by making a careful analysis of the low rank matrix $E_{\mathbf{n}}$. In particular, we show that $E_{\mathbf{n}}$ affects the number of outliers of K_N, but it does not influence the behaviour of the minimum eigenvalue of K_N with respect to that of $T_{\mathbf{n}}(\mathbf{f})$.

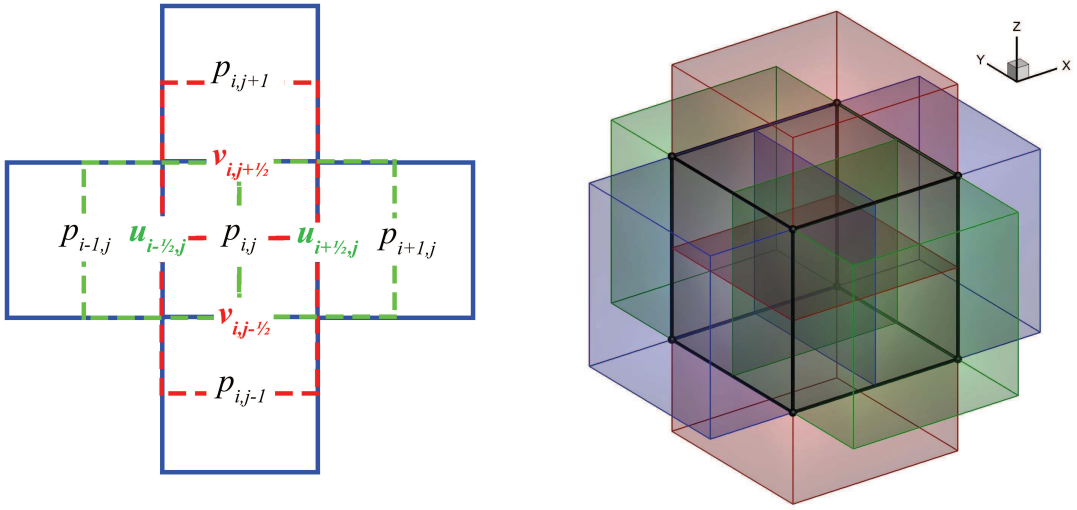
5. The spectral features are used for proposing specific (preconditioned) Krylov methods, with a study of the complexity and of the convergence speed, and for sketching a multigrid strategy, again based on the spectral information contained in the symbol.

The Chapter is organized as follows. Section II.1 is devoted to a brief overview of the numerical methods used in this chapter for the solution of the incompressible Navier-Stokes equations. Section II.2 studies the linear systems stemming from the considered approximations in a setting of structured linear algebra. In Section II.3 the spectral features are used for proposing specific (preconditioned) Krylov methods, with a study of the complexity and of the convergence speed, and for sketching a multigrid strategy, again based on the spectral information contained in the symbol. In the above two directions, several numerical experiments are reported and critically discussed. Finally, Section II.3.4 deals with conclusions, open problems, and future lines of research.

II.1 Overview

In the framework of high order semi-implicit *staggered* discontinuous Galerkin schemes for the incompressible Navier-Stokes equations, the numerical solution for the velocity $\mathbf{v} = (u, v, w)$ and the pressure p is represented by piecewise polynomials on overlapping staggered grids. The numerical solution can be written as a linear combination of polynomial basis functions,

i.e. $p_h(\mathbf{x},t) = \sum_l \phi_l(\mathbf{x})\,\hat{p}_l(t)$ and $\mathbf{v}_h(\mathbf{x},t) = \sum_l \psi_l(\mathbf{x})\,\hat{\mathbf{v}}_l(t)$. Here, ϕ_l represents the vector of piecewise polynomial basis functions computed in $\mathbf{x}$ on the *main grid*, while ψ_l are the basis functions on the edge-based staggered dual grid; the $\hat{\mathbf{v}}_l$ and $\hat{p}_l$ are the vectors of the so called *degrees of freedom* associated with the discrete solution $\mathbf{v}_h$ and p_h, respectively. The chosen staggered grid is an *edge based* staggering, corresponding to the one used in [54]. The staggering of the flow quantities is briefly depicted in Figure *II*.1, where also the main indexing used for the numerical solution is reported, together with fractional indices referring to staggered grids.



(a) Two dimensional case

(b) Three dimensional case

Figure II.1: Mesh-staggering for the two dimensional case (left) and for the three-dimensional case (right).

The discrete form of the incompressible Navier-Stokes equations after a high-order DG discretization on Cartesian staggered grids as proposed in [65] reads as

$$\mathbf{M}_{xyz}\left(\widehat{\mathbf{u}}^{\tau+\delta\tau}_{i+\frac{1}{2},j,r} - \widehat{\mathbf{F}\mathbf{u}}^{\tau}_{i+\frac{1}{2},j,r}\right) + \frac{\delta\tau}{\Delta x}\mathbf{M}_{yz}\left(\mathcal{R}_x\widehat{\mathbf{p}}^{\tau+\delta\tau}_{i+1,j,r} - \mathcal{L}_x\widehat{\mathbf{p}}^{\tau+\delta\tau}_{i,j,r}\right) = 0, \tag{II.3}$$

$$\mathbf{M}_{xyz}\left(\widehat{\mathbf{v}}^{\tau+\delta\tau}_{i,j+\frac{1}{2},r} - \widehat{\mathbf{F}\mathbf{v}}^{\tau}_{i,j+\frac{1}{2},r}\right) + \frac{\delta\tau}{\Delta y}\mathbf{M}_{zx}\left(\mathcal{R}_y\widehat{\mathbf{p}}^{\tau+\delta\tau}_{i,j+1,r} - \mathcal{L}_y\widehat{\mathbf{p}}^{\tau+\delta\tau}_{i,j,r}\right) = 0, \tag{II.4}$$

$$\mathbf{M}_{xyz}\left(\widehat{\mathbf{w}}^{\tau+\delta\tau}_{i,j,r+\frac{1}{2}} - \widehat{\mathbf{F}\mathbf{w}}^{\tau}_{i,j,r+\frac{1}{2}}\right) + \frac{\delta\tau}{\Delta z}\mathbf{M}_{xy}\left(\mathcal{R}_z\widehat{\mathbf{p}}^{\tau+\delta\tau}_{i,j,r+1} - \mathcal{L}_z\widehat{\mathbf{p}}^{\tau+\delta\tau}_{i,j,r}\right) = 0, \tag{II.5}$$

$$\frac{\mathbf{M}_{yz}\left(\mathcal{L}_x^{\top}\widehat{\mathbf{u}}^{\tau+\delta\tau}_{i+\frac{1}{2},j,r} - \mathcal{R}_x^{\top}\widehat{\mathbf{u}}^{\tau+\delta\tau}_{i-\frac{1}{2},j,r}\right)}{\Delta x} + \frac{\mathbf{M}_{zx}\left(\mathcal{L}_y^{\top}\widehat{\mathbf{v}}^{\tau+\delta\tau}_{i,j+\frac{1}{2},r} - \mathcal{R}_y^{\top}\widehat{\mathbf{v}}^{\tau+\delta\tau}_{i,j-\frac{1}{2},r}\right)}{\Delta y} \tag{II.6}$$

$$+\frac{\mathbf{M}_{xy}\left(\mathcal{L}_z^{\top}\widehat{\mathbf{w}}^{\tau+\delta\tau}_{i,j,r+\frac{1}{2}} - \mathcal{R}_z^{\top}\widehat{\mathbf{w}}^{\tau+\delta\tau}_{i,j,r-\frac{1}{2}}\right)}{\Delta z} = 0, \tag{II.7}$$

where (II.3-II.5) are the discrete momentum equations and (II.7) is the discrete divergence-free condition of the velocity.

$\mathbf{M}_{\xi_1\xi_2\xi_3}$ and $\mathbf{M}_{\xi_1\xi_2}$ for $\xi_1, \xi_2, \xi_3 \in [x, y, z]$ are the *mass matrices* defined in the standard way as the tensor product of the one dimensional mass matrix given by

$$\mathbf{M} \equiv \{M_{q\tilde{q}}\}_{q,\tilde{q}=0,...,p} \equiv \left\{ \int_0^1 \varphi_q(\xi)\, \varphi_{\tilde{q}}(\xi)\, \mathrm{d}\xi \right\}_{q,\tilde{q}=0,...,p}.$$

$\mathcal{R}_\xi$ and $\mathcal{L}_\xi$ are real-valued matrices related to the discrete form of the gradient operator in the $\xi-$direction. Their definitions are strongly related to the used staggering-framework and, in the one dimensional case, they have the following expression:

$$\mathcal{R} \equiv \{R_{q\tilde{q}}\}_{q,\tilde{q}=0,...,p} \equiv \left\{ \varphi_q(\frac{1}{2})\varphi_{\tilde{q}}(0) + \frac{1}{2}\int_0^1 \varphi_q\left(\frac{1}{2}+\frac{\xi}{2}\right)\varphi'_{\tilde{q}}\left(\frac{\xi}{2}\right)\mathrm{d}\xi \right\}_{q,\tilde{q}=0,...,p}$$

$$\mathcal{L} \equiv \{L_{q\tilde{q}}\}_{q,\tilde{q}=0,...,p} \equiv \left\{ \varphi_q(\frac{1}{2})\varphi_{\tilde{q}}(1) - \frac{1}{2}\int_0^1 \varphi_q\left(\frac{\xi}{2}\right)\varphi'_{\tilde{q}}\left(\frac{1}{2}+\frac{\xi}{2}\right)\mathrm{d}\xi \right\}_{q,\tilde{q}=0,...,p}$$

where p is the polynomial degree of the DG discretization, Δx, Δy, Δz, and $\delta\tau$ are the space and time step size. Note that the basis functions $\phi(\mathbf{x})$ and $\psi(\mathbf{x})$ on the main and dual grid in physical space can be generated after appropriate shifting by *tensor products* of the one-dimensional basis functions $\varphi(\xi)$ in a reference coordinate system with $0 \leq \xi \leq 1$. In this chapter, we consider a *nodal basis* based on the Lagrange interpolation polynomials passing through a predefined set of distinct nodes on the unit interval $[0, 1]$.

An exhaustive derivation of system (II.3-II.7) is available in [65]. The adopted discretization on staggered grids allows to link the definition of the gradient and the divergence operator at the discrete level, that are indeed both described by the same matrices $\mathcal{R}$ and $\mathcal{L}$ and their transpose.

Formal substitution of the implicit velocities $[\widehat{\mathbf{u}}_{\cdot}^{\tau+\delta\tau}, \widehat{\mathbf{v}}_{\cdot}^{\tau+\delta\tau}, \widehat{\mathbf{w}}_{\cdot}^{\tau+\delta\tau}]$ given in equations (II.3)-(II.5) into (II.7) leads to a linear system for the new pressure $\widehat{\mathbf{p}}_{\cdot}^{\tau+\delta\tau}$ that reads

$$\begin{aligned}
&\frac{\delta\tau}{\Delta x^2}\left(\mathbf{M}_{yz}\mathbb{V}^x\right)\widehat{\mathbf{p}}_{i+1,j,r}^{\tau+\delta\tau} + \frac{\delta\tau}{\Delta y^2}\left(\mathbf{M}_{zx}\mathbb{V}^y\right)\widehat{\mathbf{p}}_{i,j+1,r}^{\tau+\delta\tau} + \frac{\delta\tau}{\Delta z^2}\left(\mathbf{M}_{xy}\mathbb{V}^z\right)\widehat{\mathbf{p}}_{i,j,r+1}^{\tau+\delta\tau} \\
&+\left(\frac{\delta\tau}{\Delta x^2}\mathbf{M}_{yz}\mathbb{W}^x + \frac{\delta\tau}{\Delta y^2}\mathbf{M}_{zx}\mathbb{W}^y + \frac{\delta\tau}{\Delta z^2}\mathbf{M}_{xy}\mathbb{W}^z\right)\widehat{\mathbf{p}}_{i,j,r}^{\tau+\delta\tau} \\
&+\frac{\delta\tau}{\Delta x^2}\left(\mathbf{M}_{yz}\mathbb{L}^x\right)\widehat{\mathbf{p}}_{i-1,j,r}^{\tau+\delta\tau} + \frac{\delta\tau}{\Delta y^2}\left(\mathbf{M}_{zx}\mathbb{L}^y\right)\widehat{\mathbf{p}}_{i,j-1,r}^{\tau+\delta\tau} + \frac{\delta\tau}{\Delta z^2}\left(\mathbf{M}_{xy}\mathbb{L}^z\right)\widehat{\mathbf{p}}_{i,j,r-1}^{\tau+\delta\tau} \\
&=\widehat{\mathbf{b}}_{i,j,r}^{\tau}, \qquad\qquad (II.8)\\
&\quad \text{for} \quad i = 2, ..., n_1 - 1; \quad j = 2, ..., n_2 - 1; \quad r = 2, ..., n_3 - 1
\end{aligned}$$

where

$$\mathbb{V} = -\left(\mathcal{L}^\top\mathbf{M}^{-1}\mathcal{R}\right), \quad \mathbb{L} = -\left(\mathcal{R}^\top\mathbf{M}^{-1}\mathcal{L}\right), \quad \mathbb{W} = \left(\mathcal{L}^\top\mathbf{M}^{-1}\mathcal{L}\right) + \left(\mathcal{R}^\top\mathbf{M}^{-1}\mathcal{R}\right). \qquad (II.9)$$

and n_1, n_2, n_3 are the total number of elements in the x, y, and z direction, respectively. System $(II.8)$ is then written in compact form as $K_N\mathbf{p}^{\tau+\delta\tau} = b^\tau$. Here $\mathbf{p}^{\tau+\delta\tau}$ collects all the unknown pressure degrees of freedom at the new time step $\tau + \delta\tau$ and b^τ contains all the terms known at

the time step τ, see again [65] for more details. In particular, in [65] it has been shown that the resulting linear system is *symmetric*. Furthermore, it is clear from system (II.8) that the stencil involves only the direct neighbors, and hence it is a symmetric 7 block diagonal system for the $3D$ case and a 5 block-diagonal system for the $2D$ case.

Once the new pressure $\mathbf{p}^{\tau+\delta\tau}$ is known, we can readily compute the new velocity field $[\widehat{\mathbf{u}}_{\cdot}^{\tau+\delta\tau}, \widehat{\mathbf{v}}_{\cdot}^{\tau+\delta\tau}, \widehat{\mathbf{w}}_{\cdot}^{\tau+\delta\tau}]$ from equations (II.3)-(II.5).

II.2 Spectral analysis

This section is devoted to the structural and spectral analysis of the linear systems arising from the staggered semi-implicit DG approximation of incompressible two-dimensional incompressible Navier-Stokes equations, with special attention to the following Items:

- structural properties, in connection with multilevel block Toeplitz (and circulant) matrices,
- distribution spectral analysis in the Weyl sense,
- conditioning and asymptotic behaviour of the extremal eigenvalues.

In particular, the first Item is used for the second two, which in turn are of interest in the analysis of the intrinsic difficulty of the problem and in the design and convergence analysis of (preconditioned) Krylov methods [5, 11].

Our aim is to efficiently solve large linear systems arising from the staggered DG approximation of incompressible two-dimensional Navier-Stokes equations taking advantage of the structure of the coefficient matrix and especially of its spectral features. More precisely, when discretizing the problem of interest for a sequence of discretization parameters h_N we obtain a sequence of linear systems, in which the Nth component is of the form

$$K_N x = b, \quad K_N \in \mathbb{R}^{N\times N}, \quad x, b \in \mathbb{R}^N, \tag{II.10}$$

whose approximation error tends to zero as the coefficient matrix size N grows to infinity. In order to analyze standard methods and for designing new efficient solvers for the considered linear systems, it is of crucial importance to have a spectral analysis of the matrix-sequence $\{K_N\}_N$. As we will show in the next sections, the coefficient matrix K_N is, up to low-rank perturbations, a 2-level block Toeplitz matrix: however, when considering variable coefficients or for the study of the preconditioning, standard Toeplitz structures are not sufficient. For this reason, we need to introduce the notion of multilevel block-Toeplitz sequences associated with a matrix-valued symbol and of Generalized Locally Toeplitz (GLT) algebra.

II.2.1 Analysis of the spectral symbol

Using Definition I.4.1, we can now explicitly express the symbol of the matrix K_N in (II.10). Let $\mathbf{n} = (n_1, n_2)$ be a 2-index and consequently $N(\mathbf{n}) = n_1 n_2$. If p is the degree of the basis functions used for the staggered DG, we obtain the following Hermitian matrix

$$K_N = T_{\mathbf{n}}(\mathbf{f}) + E_{\mathbf{n}}, \quad N = (p+1)^2 N(\mathbf{n}), \tag{II.11}$$

where

$$T_{\mathbf{n}}(\mathbf{f}) = \left[\hat{\mathbf{f}}_{\mathbf{i}-\mathbf{j}}\right]_{\mathbf{i},\mathbf{j}=\mathbf{e}}^{\mathbf{n}} \in \mathcal{M}_N$$

and $\mathbf{f} : \mathcal{I}_2 \to \mathbb{C}^{s\times s}$, $s = (p+1)^2$, while $E_{\mathbf{n}}$ is a low-rank perturbation whose rank grows at most proportionally to $\sqrt{N(\mathbf{n})}$ and with constant depending on the bandwidths of K_N. The nonzero coefficients of $T_{\mathbf{n}}(\mathbf{f}) = [\hat{\mathbf{f}}_{\mathbf{i}-\mathbf{j}}]_{\mathbf{i},\mathbf{j}=\mathbf{e}}^{\mathbf{n}}$ correspond to the indices $\mathbf{i} = (i_1, i_2), \mathbf{j} = (j_1, j_2)$ such that

$$|i_1 - j_1| + |i_2 - j_2| \leq 1.$$

For example, for $\mathbf{n} = (3, 3)$,

$$T_{\mathbf{n}}(\mathbf{f}) = \left(\begin{array}{ccc|ccc|ccc}
\hat{\mathbf{f}}_{(0,0)} & \hat{\mathbf{f}}_{(0,-1)} & 0 & \hat{\mathbf{f}}_{(-1,0)} & 0 & 0 & 0 & 0 & 0 \\
\hat{\mathbf{f}}_{(0,1)} & \hat{\mathbf{f}}_{(0,0)} & \hat{\mathbf{f}}_{(0,-1)} & 0 & \hat{\mathbf{f}}_{(-1,0)} & 0 & 0 & 0 & 0 \\
0 & \hat{\mathbf{f}}_{(0,1)} & \hat{\mathbf{f}}_{(0,0)} & 0 & 0 & \hat{\mathbf{f}}_{(-1,0)} & 0 & 0 & 0 \\
\hline
\hat{\mathbf{f}}_{(1,0)} & 0 & 0 & \hat{\mathbf{f}}_{(0,0)} & \hat{\mathbf{f}}_{(0,-1)} & 0 & \hat{\mathbf{f}}_{(-1,0)} & 0 & 0 \\
0 & \hat{\mathbf{f}}_{(1,0)} & 0 & \hat{\mathbf{f}}_{(0,1)} & \hat{\mathbf{f}}_{(0,0)} & \hat{\mathbf{f}}_{(0,-1)} & 0 & \hat{\mathbf{f}}_{(-1,0)} & 0 \\
0 & 0 & \hat{\mathbf{f}}_{(1,0)} & 0 & \hat{\mathbf{f}}_{(0,1)} & \hat{\mathbf{f}}_{(0,0)} & 0 & 0 & \hat{\mathbf{f}}_{(-1,0)} \\
\hline
0 & 0 & 0 & \hat{\mathbf{f}}_{(1,0)} & 0 & 0 & \hat{\mathbf{f}}_{(0,0)} & \hat{\mathbf{f}}_{(0,-1)} & 0 \\
0 & 0 & 0 & 0 & \hat{\mathbf{f}}_{(1,0)} & 0 & \hat{\mathbf{f}}_{(0,1)} & \hat{\mathbf{f}}_{(0,0)} & \hat{\mathbf{f}}_{(0,-1)} \\
0 & 0 & 0 & 0 & 0 & \hat{\mathbf{f}}_{(1,0)} & 0 & \hat{\mathbf{f}}_{(0,1)} & \hat{\mathbf{f}}_{(0,0)}
\end{array}\right). \tag{II.12}$$

Therefore, in the two-dimensional case ($k = 2$) the symbol $\mathbf{f}$ is given by

$$\mathbf{f}(\theta_1, \theta_2) = \hat{\mathbf{f}}_{(0,0)} + \hat{\mathbf{f}}_{(-1,0)} e^{-\mathbf{i}\theta_1} + \hat{\mathbf{f}}_{(0,-1)} e^{-\mathbf{i}\theta_2} + \hat{\mathbf{f}}_{(1,0)} e^{\mathbf{i}\theta_1} + \hat{\mathbf{f}}_{(0,1)} e^{\mathbf{i}\theta_2}, \tag{II.13}$$

where $\hat{\mathbf{f}}_{(0,0)}, \hat{\mathbf{f}}_{(-1,0)}, \hat{\mathbf{f}}_{(0,-1)}, \hat{\mathbf{f}}_{(1,0)}, \hat{\mathbf{f}}_{(0,1)} \in \mathbb{R}^{(p+1)^2\times(p+1)^2}$, that is $\mathbf{f}$ is a linear trigonometric polynomial in the variables θ_1 and θ_2. For detailed expressions of these matrices in the particular case $k = 2$ and $p = 3$, see VI.1. Furthermore, the coefficients of $T_{\mathbf{n}}(\mathbf{f})$ satisfy the following relations

$$\hat{\mathbf{f}}_{(0,0)}^T = \hat{\mathbf{f}}_{(0,0)}, \qquad \hat{\mathbf{f}}_{(-1,0)}^T = \hat{\mathbf{f}}_{(1,0)}, \qquad \hat{\mathbf{f}}_{(0,-1)}^T = \hat{\mathbf{f}}_{(0,1)}.$$

As a consequence,

$$\mathbf{f}^*(\theta_1, \theta_2) = \mathbf{f}(\theta_1, \theta_2),$$

that is $\mathbf{f}$ is a Hermitian matrix-valued function which implies that $T_{\mathbf{n}}(\mathbf{f})$ is a Hermitian matrix. Using Theorem I.4.2, we can conclude that

$$\{T_{\mathbf{n}}(\mathbf{f})\}_{\mathbf{n}\in\mathbb{N}^2} \sim_\lambda (\mathbf{f}, \mathcal{I}_2). \tag{II.14}$$

From **GLT3**, we know that $\{T_{\mathbf{n}}(\mathbf{f})\}_{\mathbf{n}\in\mathbb{N}^2}$ is a GLT sequence with symbol $\mathbf{f}$. Moreover, let us observe that $\{E_{\mathbf{n}}\}_{\mathbf{n}\in\mathbb{N}^2} \sim_\sigma 0$ and hence, by the property **GLT4**, the sequence $\{E_{\mathbf{n}}\}_{\mathbf{n}\in\mathbb{N}^2}$ is a GLT sequence with symbol identically zero. Therefore, by **GLT2** and by relation (II.14), the sequence $\{T_{\mathbf{n}}(\mathbf{f}) + E_{\mathbf{n}}\}_{\mathbf{n}\in\mathbb{N}^2}$ is a GLT sequence with symbol $\mathbf{f}$. Consequently, by recalling that $T_{\mathbf{n}}(\mathbf{f}) + E_{\mathbf{n}}$ is real symmetric for every $\mathbf{n}$ and using **GLT1**, we deduce

$$\{K_N\}_N \sim_\lambda (\mathbf{f}, \mathcal{I}_2). \tag{II.15}$$

Furthermore, since each K_N is symmetric and its blocks are symmetric and real, from Remark 2 with $k = 2$, we have

$$\{K_N\}_N \sim_\lambda (\mathbf{f}, \mathcal{I}_2^+). \tag{II.16}$$

Let

$$\lambda_1(K_N) \leq \lambda_2(K_N) \leq \cdots \leq \lambda_N(K_N).$$

be the eigenvalues of K_N. Recalling Remark 1, from equation (II.16), we know that for N sufficiently large, $N/(p+1)^2$ eigenvalues of K_N, up to outliers, can be approximated by a sampling of $\lambda^{(1)}(\mathbf{f})$ on a uniform equispaced grid of the domain $\mathcal{I}_2^+$, and so on until the last $N/(p+1)^2$ eigenvalues which can be approximated by an equispaced sampling of $\lambda^{((p+1)^2)}(\mathbf{f})$ in the domain. In the following section we give numerical evidence of this result.

II.2.2 Numerical tests

Let us fix $\mathbf{n} = (n_1, n_2)$, with $n_1, n_2 = n$, and let $p = 2$. Within these choices, the matrix-size of K_N defined as in (II.11) is $N = 9n^2$. This section is devoted to the comparison of the eigenvalues of K_N with a sampling of the eigenvalue functions $\lambda^{(1)}(\mathbf{f}), \ldots, \lambda^{(9)}(\mathbf{f})$. Actually, we do not analytically compute the eigenvalue functions, but, according to Theorem I.6.2 and Remark 3, we are able to provide an "exact" evaluation of them on an equispaced grid on $\mathcal{I}_2^+$ (see Subsection II.2.2.1) and this is sufficient for our aims.

II.2.2.1 Evaluation of the eigenvalue functions of the symbol

Let us define the following equispaced grid on $\mathcal{I}_2^+$

$$G_n = \left\{ (\theta_1^{(j)}, \theta_2^{(k)}) = \left(\frac{j\pi}{n}, \frac{k\pi}{n} \right), \qquad j,k = 0, \ldots, n-1 \right\}$$

and let us consider the following n^2 Hermitian matrices of size 9×9

$$A_{j,k} := \mathbf{f}(\theta_1^{(j)}, \theta_2^{(k)}), \quad j,k = 0, \ldots, n-1. \tag{II.17}$$

Ordering in non decreasing way the eigenvalues of $A_{j,k}$

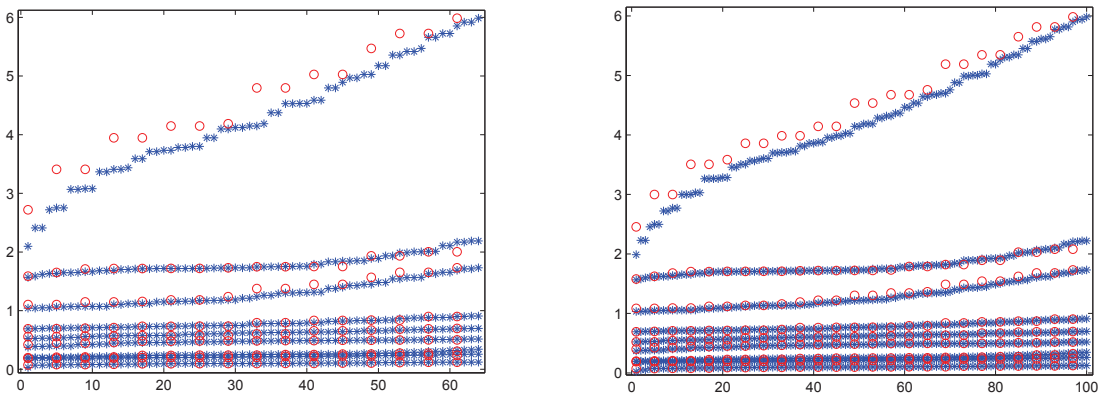
$$\lambda_1(A_{j,k}) \leq \lambda_2(A_{j,k}) \leq \cdots \leq \lambda_9(A_{j,k}), \quad j,k = 0, \ldots, n-1,$$

for a fixed $l = 1, \ldots, 9$, an evaluation of $\lambda^{(l)}(\mathbf{f})$ at $(\theta_1^{(j)}, \theta_2^{(k)})$ is given by $\lambda_l(A_{j,k})$, $j,k = 0, \ldots, n-1$. From now onwards, fixed l, we will denote by $P_l^{(n)}$ the vector of all eigenvalues $\lambda_l(A_{j,k})$, $j,k = 0, \ldots, n-1$, that is

$$P_l^{(n)} := [\lambda_l(A_{0,0}), \lambda_l(A_{0,1}), \ldots, \lambda_l(A_{n-1,n-1})],$$

and by $P^{(n)}$ the vector of all eigenvalues $\lambda_l(A_{j,k})$, $j,k = 0,\dots,n-1$ varying l

$$P^{(n)} := [\lambda_1(A_{0,0}),\dots,\lambda_1(A_{n-1,n-1}),\dots,\lambda_9(A_{0,0}),\dots,\lambda_9(A_{n-1,n-1})]\,.$$



(a) $n = 4$ (b) $n = 5$

Figure II.2: Comparison between the approximation of the eigenvalue functions $\lambda^{(l)}(\mathbf{f})$, $l = 1,\dots,9$ on the grid G_n contained in $P_l^{(n)}$ ($\circ$) and the corresponding approximation on the grid twice as fine G_{2n} contained in $P_l^{(2n)}$ ($*$). Each "curve" refers to a different value of l. The parameter n equals 4 and 5 in subplots (a) and (b), respectively.

Refining the grid G_n by increasing n, we can provide the evaluation of the eigenvalue functions of $\mathbf{f}$ in a larger number of grid points: convincing numerical evidences of the latter claim are reported in Figure II.2. More specifically, in Figures II.2(a), II.2(b) we compare the approximation of $\lambda^{(l)}(\mathbf{f})$ on G_n, $n = 4,5$ contained in $P_l^{(n)}$ (ordered in non decreasing way) with the approximation of the same eigenvalue function on a grid that is twice as fine G_{2n}, $n = 4,5$ contained in $P_l^{(2n)}$ (ordered in non decreasing way as well) for every $l = 1,\dots,9$.

Therefore, for n sufficiently large, a feasible approximation of $\lambda^{(l)}(\mathbf{f})$, $l = 1,\dots,9$, can be obtained by displaying $P_l^{(n)}$ as a mesh on G_n (see Figure II.3, for $n = 40$).

II.2.2.2 Spectral distribution of $\{K_N\}_N$

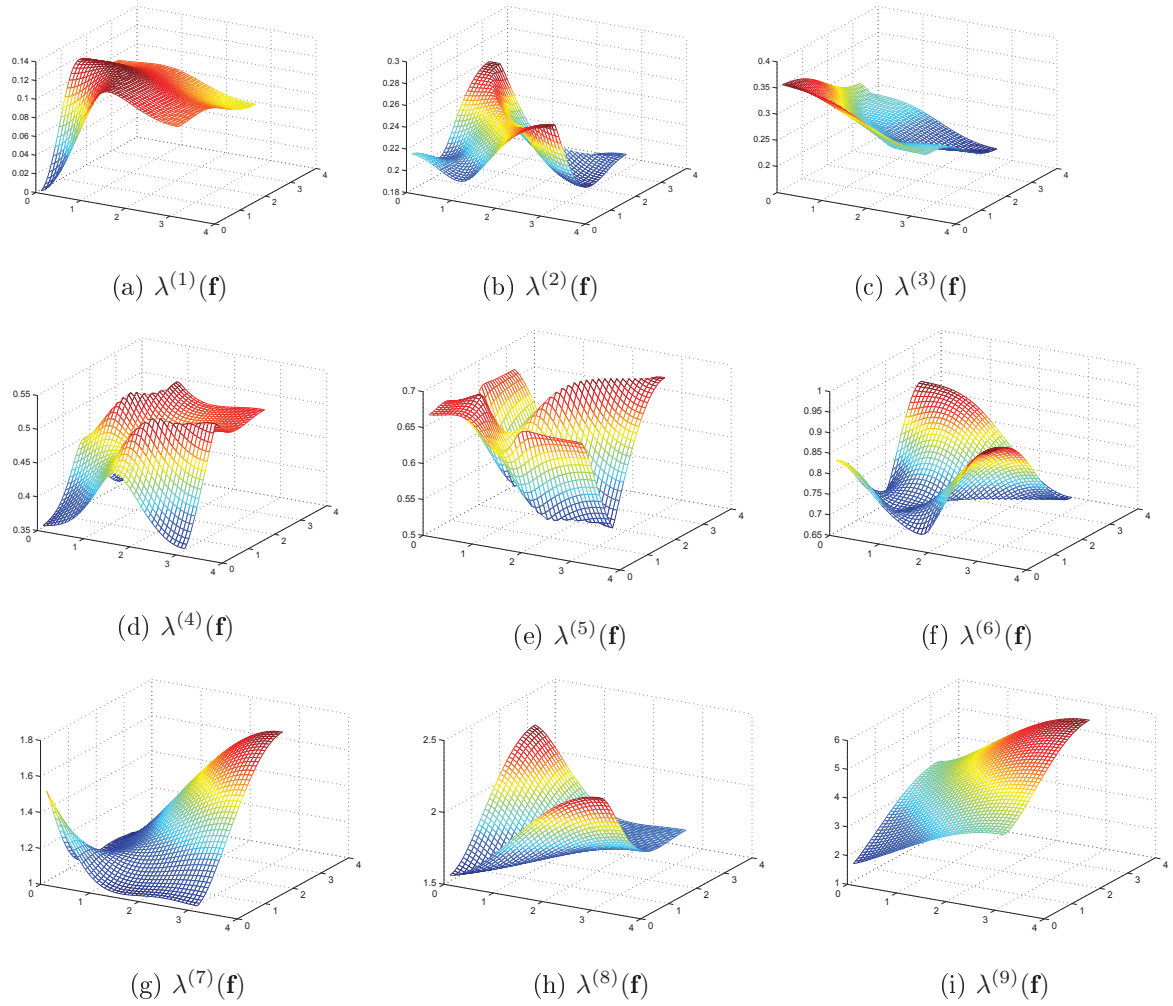
In this subsection we provide numerical evidences of the distribution result (II.16), making use of the strategy for computing an approximation of $\lambda^{(l)}(\mathbf{f})$ on an equispaced grid showed in Subsection II.2.2.1.

As a first evidence, we compare the eigenvalues of K_N with the evaluation of $\lambda^{(l)}(\mathbf{f})$ $l = 1,\dots,9$ at G_n given by a proper ordering of $P^{(n)}$. As shown in Figure II.4 in which we fixed $n = 40$, the eigenvalues of K_N mimic, up to outliers, the sampling of the eigenvalue functions. This agrees with relation (II.16).

Aside from such a global comparison, if

$$\text{esssup}_{\mathcal{I}_2^+}\left(\lambda^{(l)}(\mathbf{f})\right)(\boldsymbol{\theta}) \le \text{essinf}_{\mathcal{I}_2^+}\left(\lambda^{(l+1)}(\mathbf{f})\right)(\boldsymbol{\theta}),$$

for some $l = 1,\dots,8$, exploiting Remark 1, we can provide a more accurate analysis of the spectrum of K_N determining how many blocks it is made up of and how many eigenvalues



(a) $\lambda^{(1)}(\mathbf{f})$ (b) $\lambda^{(2)}(\mathbf{f})$ (c) $\lambda^{(3)}(\mathbf{f})$

(d) $\lambda^{(4)}(\mathbf{f})$ (e) $\lambda^{(5)}(\mathbf{f})$ (f) $\lambda^{(6)}(\mathbf{f})$

(g) $\lambda^{(7)}(\mathbf{f})$ (h) $\lambda^{(8)}(\mathbf{f})$ (i) $\lambda^{(9)}(\mathbf{f})$

Figure II.3: Approximation of the eigenvalues functions $\lambda^{(l)}(\mathbf{f})$, $l = 1, \ldots, 9$ as a mesh on G_n, when $n = 40$

contains each block. With this aim, let us observe that, for a sufficiently large n, if we order in non decreasing way $P_l^{(n)}$, the first and the last element in $P_l^{(n)}$ satisfy the following relations:

$$(P_l^{(n)})_1 \approx m_l, \qquad (P_l^{(n)})_{n^2} \approx M_l, \quad l = 1, \ldots, 9.$$

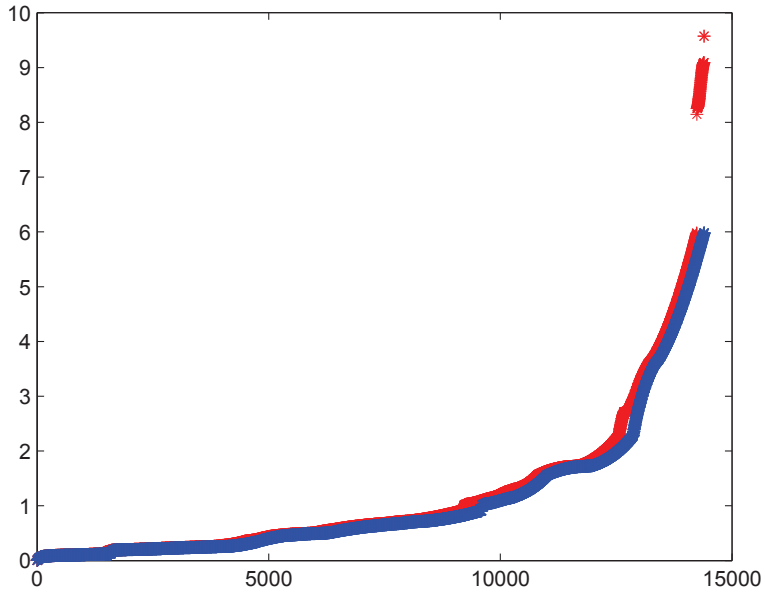


Figure II.4: Comparison of the eigenvalues of K_N ($*$) with the approximation of $\lambda^{(l)}(\mathbf{f})$ $l = 1, \ldots, 9$ on G_n given by a proper ordering of $P^{(n)}$ ($*$), for $n = 40$.

A satisfactory approximation of $[m_l, M_l]$ can be numerically computed by setting $n = 500$; as a result we obtain the following approximations

$$
\begin{aligned}
[m_1, M_1] &\approx [0.000000000, 0.123775621],\\
[m_2, M_2] &\approx [0.186715287, 0.260786617],\\
[m_3, M_3] &\approx [0.197732806, 0.355965321],\\
[m_4, M_4] &\approx [0.355965321, 0.524158720],\\
[m_5, M_5] &\approx [0.520903995, 0.696882517],\\
[m_6, M_6] &\approx [0.677870643, 0.910001758],\\
[m_7, M_7] &\approx [1.015599697, 1.731431133],\\
[m_8, M_8] &\approx [1.560701345, 2.284336270],\\
[m_9, M_9] &\approx [1.651355307, 5.985129348].
\end{aligned}
$$

However, looking at

$$\mathbf{f}(0,0) = \begin{bmatrix} \frac{19}{45} & \frac{1}{60} & \frac{-7}{40} & \frac{1}{60} & \frac{-4}{45} & \frac{-7}{180} & \frac{-7}{40} & \frac{-7}{180} & \frac{11}{180} \\ \frac{1}{60} & \frac{46}{45} & \frac{1}{60} & \frac{-4}{45} & \frac{-4}{15} & \frac{-4}{45} & \frac{-7}{180} & \frac{-8}{15} & \frac{-7}{180} \\ \frac{-7}{40} & \frac{1}{60} & \frac{19}{45} & \frac{-7}{180} & \frac{-4}{45} & \frac{1}{60} & \frac{11}{180} & \frac{-7}{180} & \frac{-7}{40} \\ \frac{1}{60} & \frac{-4}{45} & \frac{-7}{180} & \frac{46}{45} & \frac{-4}{15} & \frac{-8}{15} & \frac{1}{60} & \frac{-4}{45} & \frac{-7}{180} \\ \frac{-4}{45} & \frac{-4}{15} & \frac{-4}{45} & \frac{-4}{15} & \frac{64}{45} & \frac{-4}{15} & \frac{-4}{45} & \frac{-4}{15} & \frac{-4}{45} \\ \frac{-7}{180} & \frac{-4}{45} & \frac{1}{60} & \frac{-8}{15} & \frac{-4}{15} & \frac{46}{45} & \frac{-7}{180} & \frac{-4}{45} & \frac{1}{60} \\ \frac{-7}{40} & \frac{-7}{180} & \frac{11}{180} & \frac{1}{60} & \frac{-4}{45} & \frac{-7}{180} & \frac{19}{45} & \frac{1}{60} & \frac{-7}{40} \\ \frac{-7}{180} & \frac{-8}{15} & \frac{-7}{180} & \frac{-4}{45} & \frac{-4}{15} & \frac{-4}{45} & \frac{1}{60} & \frac{46}{45} & \frac{1}{60} \\ \frac{11}{180} & \frac{-7}{180} & \frac{-7}{40} & \frac{-7}{180} & \frac{-4}{45} & \frac{1}{60} & \frac{-7}{40} & \frac{1}{60} & \frac{19}{45} \end{bmatrix}$$

we observe that the matrix has row sum equal to zero for every row.
This means that $\mathbf{f}(0,0)e = 0$ where $e \in \mathbb{R}^9$ is the vector of all ones. Therefore $\mathbf{f}(0,0)$ is analytically singular and $m_1 = 0$, since the symbol is theoretically nonnegative definite because of the Galerkin approach. Now, recalling the second Item of Theorem I.4.4 and observing that $\mathbf{f}(\pi,\pi)$ is positive definite, we deduce that $\left(\lambda^{(1)}(\mathbf{f})\right)(\theta_1,\theta_2)$ has positive maximum and therefore the interval $[m_1, M_1]$ can be replaced by $(0, M_1]$.

From now onwards, we assume $(0, M_1]$, (m_l, M_l), $l = 2, \ldots, 9$, to be equal to its estimate. Let us observe that the following relations hold

$$\begin{aligned} &M_1 < m_2, \\ &M_3 = m_4, \\ &M_6 < m_7. \end{aligned} \tag{II.18}$$

In other words, according to relations (II.16), (II.18), and Remark 1, we expect the eigenvalues of K_N to satisfy

$$\begin{aligned} \#\{i : \lambda_i(K_N) \in (0, M_1]\} &= \frac{9n^2}{9} + o(9n^2), \\ \#\{i : \lambda_i(K_N) \in [m_2, M_3]\} &= 2\frac{9n^2}{9} + o(9n^2), \\ \#\{i : \lambda_i(K_N) \in [m_4, M_6]\} &= 3\frac{9n^2}{9} + o(9n^2), \\ \#\{i : \lambda_i(K_N) \in [m_7, M_9]\} &= 3\frac{9n^2}{9} + o(9n^2), \end{aligned} \tag{II.19}$$

and then to identify 4 blocks

$$\begin{aligned}
\mathrm{Bl}_1 &= [\lambda_1(K_N), \dots, \lambda_{n^2}(K_N)]\,, \\
\mathrm{Bl}_2 &= [\lambda_{n^2+1}(K_N), \dots, \lambda_{3n^2}(K_N)]\,, \\
\mathrm{Bl}_3 &= [\lambda_{3n^2+1}(K_N), \dots, \lambda_{6n^2}(K_N)]\,, \\
\mathrm{Bl}_4 &= [\lambda_{6n^2+1}(K_N), \dots, \lambda_{9n^2}(K_N)]\,.
\end{aligned}$$

Correspondingly, we can split the vector $P^{(n)}$ containing the sampling of the eigenvalue functions on G_n as follows

$$\begin{aligned}
\mathrm{Eval}_1 &= [(P^{(n)})_1, \dots, (P^{(n)})_{n^2}], \\
\mathrm{Eval}_2 &= [(P^{(n)})_{n^2+1}, \dots, (P^{(n)})_{3n^2}], \\
\mathrm{Eval}_3 &= [(P^{(n)})_{3n^2+1}, \dots, (P^{(n)})_{6n^2}], \\
\mathrm{Eval}_4 &= [(P^{(n)})_{6n^2+1}, \dots, (P^{(n)})_{9n^2}].
\end{aligned}$$

Note that because of (II.19), a number of outliers infinitesimal in the dimension N is allowed. For instance, when $n = 40$ ($N = 14400$), we find

$$\frac{9n^2}{9} = 1600, \qquad 2\frac{9n^2}{9} = 3200, \qquad 3\frac{9n^2}{9} = 4800,$$

and

$$\begin{aligned}
\#\{i \,:\, \lambda_i(K_N) \in (0, M_1]\} &= 1444, \\
\#\{i \,:\, \lambda_i(K_N) \in [m_2, M_3]\} &= 2911, \\
\#\{i \,:\, \lambda_i(K_N) \in [m_4, M_6]\} &= 4670, \\
\#\{i \,:\, \lambda_i(K_N) \in [m_7, M_9]\} &= 5016.
\end{aligned} \tag{II.20}$$

Therefore, from relations (II.20), we expect a number of eigenvalues of K_N which are in none of the blocks or which are in the "wrong" block (5016 effective against 4800 expected eigenvalues in the last block). This is confirmed by Figure II.5 in which we represent in black the whole spectrum of K_N and highlight by means of different colours the eigenvalues belonging to different blocks. On the other hand, such a phenomenon is in line with relations (II.19) and the order of what is missing/exceeding is infinitesimal in the dimension N. As an example, in Table II.1 we compare the actual number of eigenvalues of K_N contained in the first interval $(0, M_1]$ with the expected number $9n^2/9$. In such way, we succeed in counting the outliers of K_N in $(0, M_1]$, whose cardinality behaves as $O(\sqrt{9n^2})$.

A further evidence of relation (II.16) can be obtained by comparing block by block the eigenvalues of K_N with the sampling of the eigenvalue functions of $\mathbf{f}$, that is comparing Bl_1, Bl_2, Bl_3, Bl_4 with Eval_1, Eval_2, Eval_3, Eval_4, respectively. Two possibilities are available.

- On one hand, we can order Eval_t in a non decreasing way and compare it with Bl_t.

  As an example, in Figure II.6 we compare Bl_1 with Eval_1 fixed $n = 40$. Note that a certain number of eigenvalues of K_N seems not to behave as the corresponding sampling of $\lambda^{(1)}(\mathbf{f})$. Nevertheless, a direct computation showed that such a number agrees with the one reported in Table II.1. Similar results can be obtained in the comparison between Bl_2 with Eval_2, Bl_3 with Eval_3, Bl_4 with Eval_4.

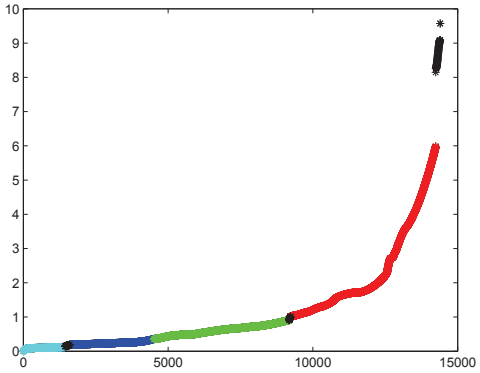


Figure II.5: Eigenvalues of K_N for $n = 40$ $(*)$ together with the eigenvalues of K_N satisfying (II.19) $(*)(*)(*)(*)$

n	eigs in $(0, M_1]$	$9n^2/9$	Out.	Out./$\sqrt{9n^2}$
10	64	100	36	1.20
15	169	225	56	1.24
20	324	400	76	1.26
25	529	625	96	1.28
30	784	900	116	1.29
35	1089	1225	136	1.29
40	1444	1600	156	1.30

Table II.1: Comparison of the effective number of eigenvalues of K_N contained in the first interval $(0, M_1]$ with the expected number $9n^2/9$

- On the other hand, we can compare the elements of Eval_t with the elements of Bl_t by means of the following matching algorithm

  - for a fixed $\lambda \in \mathrm{Bl}_t$ find $\tilde{\eta} \in \mathrm{Eval}_t$ such that

$$\|\lambda - \tilde{\eta}\| = \min_{\eta \in \mathrm{Eval}_t} \|\lambda - \eta\|;$$

  - associate λ to the couple in G_n corresponding to $\tilde{\eta}$.

  Making use of the previous algorithm, in Figure II.7, we compare the eigenvalues of K_N with $\lambda^{(l)}(\mathbf{f})$, $l = 1, \ldots, 9$ displayed as a mesh on G_n, for $n = 40$. Once again, the eigenvalues of K_N mimic, up to outliers, the sampling of the eigenvalue functions.

  Moreover, looking at Figure II.7(a), we computed the eigenvalues of K_N which do not behave as the corresponding sampling of $\lambda^{(1)}(\mathbf{f})$ and, as expected, their order is $O(\sqrt{9n^2})$ (see again Table II.1). As an additional confirmation of such a behaviour, in Table II.2 we show the number of outliers of K_N with respect to the sampling of $\lambda^{(9)}(\mathbf{f})$ (see Figure II.7(i)).

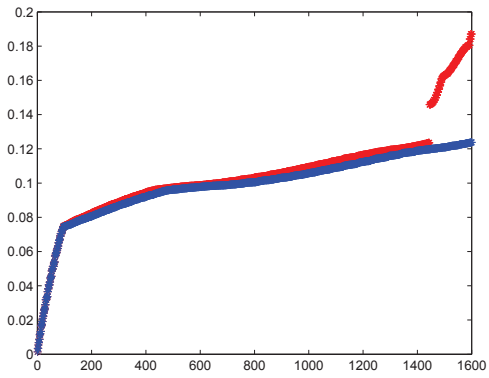


Figure II.6: Comparison between Bl_1 ($*$) and Eval_1 ($*$), for $n = 40$

n	Out.	Out./$\sqrt{9n^2}$
10	40	1.33
15	60	1.33
20	80	1.33
25	100	1.33
30	120	1.33
35	140	1.33
40	160	1.33

Table II.2: Number of eigenvalues of K_N which do not behave as the corresponding sampling of $\lambda^{(9)}(\mathbf{f})$.

II.2.3 A focus on the eigenvalue functions in a neighborhood of the origin

In this subsection we study in more detail the behaviour of the eigenvalues $\lambda^{(l)}(\mathbf{f})$, $l = 1, \ldots, 9$ at $(0,0)$. Such information is crucial when studying the convergence of a preconditioned Krylov or of a multigrid method. Since

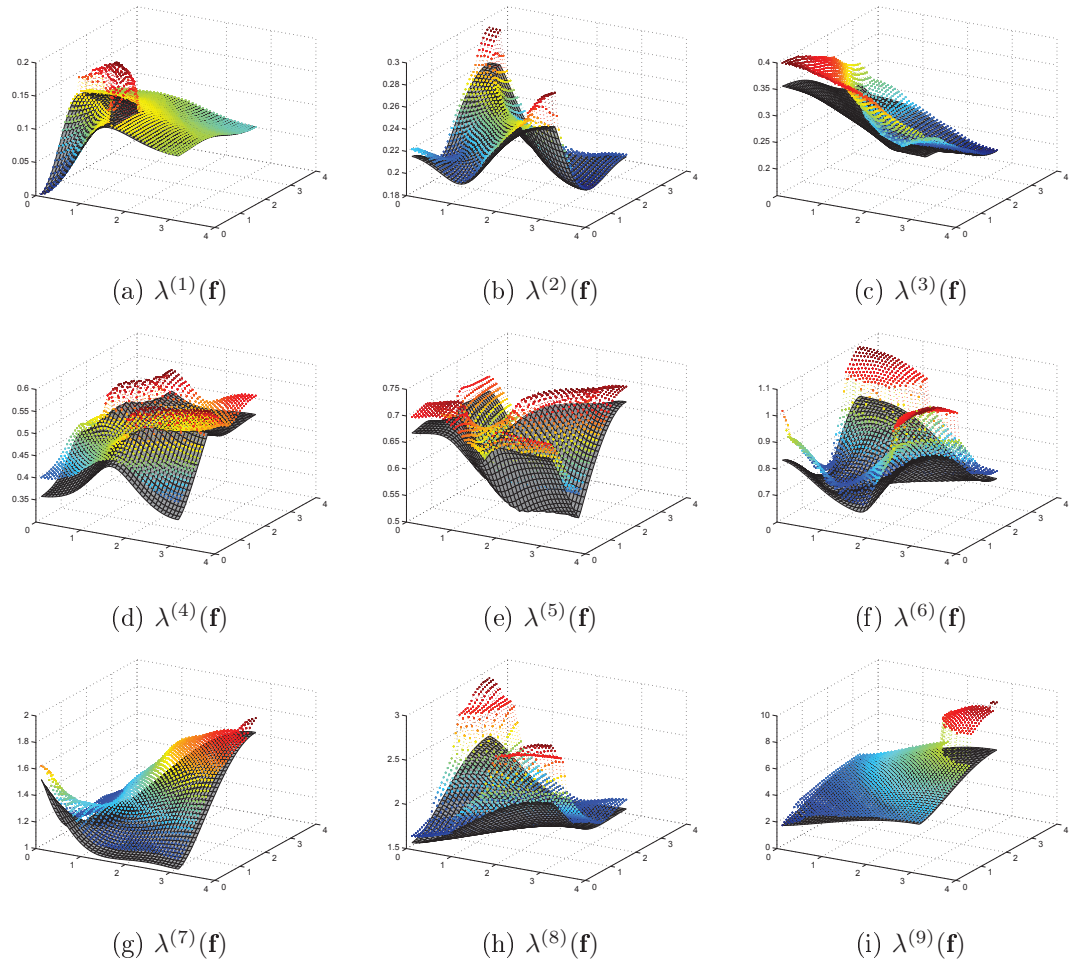
$$\left(\lambda^{(1)}(\mathbf{f})\right)(\theta_1,\theta_2) < \left(\lambda^{(l)}(\mathbf{f})\right)(\theta_1,\theta_2), \quad l = 2,\ldots,9, \quad (\theta_1,\theta_2) \in \mathcal{I}_2^+, \tag{II.21}$$

it is sufficient to study $\lambda^{(1)}(\mathbf{f})$ in $(0,0)$. Because of (II.21), the behaviour of $\lambda^{(1)}(\mathbf{f})$ in $(0,0)$ is equivalent to the one of

$$\det \mathbf{f}(\theta_1,\theta_2) = \prod_{i=1}^{9} \left(\lambda^{(i)}(\mathbf{f})\right)(\theta_1,\theta_2)$$

at the same point, which as a product of nonnegative functions is still a nonnegative function. We numerically checked that

$$\begin{aligned} \det \mathbf{f}(\theta_1,\theta_2)|_{(0,0)} &= 0, \\ \left.\frac{\partial \det \mathbf{f}(\theta_1,\theta_2)}{\partial \theta_1}\right|_{(0,0)} &= \left.\frac{\partial \det \mathbf{f}(\theta_1,\theta_2)}{\partial \theta_2}\right|_{(0,0)} = 0, \\ \left.\frac{\partial^2 \det \mathbf{f}(\theta_1,\theta_2)}{\partial \theta_2 \partial \theta_1}\right|_{(0,0)} &= \left.\frac{\partial^2 \det \mathbf{f}(\theta_1,\theta_2)}{\partial \theta_1 \partial \theta_2}\right|_{(0,0)} = 0, \end{aligned}$$



(a) $\lambda^{(1)}(\mathbf{f})$ (b) $\lambda^{(2)}(\mathbf{f})$ (c) $\lambda^{(3)}(\mathbf{f})$

(d) $\lambda^{(4)}(\mathbf{f})$ (e) $\lambda^{(5)}(\mathbf{f})$ (f) $\lambda^{(6)}(\mathbf{f})$

(g) $\lambda^{(7)}(\mathbf{f})$ (h) $\lambda^{(8)}(\mathbf{f})$ (i) $\lambda^{(9)}(\mathbf{f})$

Figure II.7: Comparison between the eigenvalues of K_N and $\lambda^{(l)}(\mathbf{f})$, $l = 1, \ldots, 9$ displayed as a mesh on G_n, when $n = 40$

$$\left.\frac{\partial^2 \det \mathbf{f}(\theta_1,\theta_2)}{\partial {\theta_1}^2}\right|_{(0,0)} = \left.\frac{\partial^2 \det \mathbf{f}(\theta_1,\theta_2)}{\partial {\theta_2}^2}\right|_{(0,0)} = \frac{53}{3912}.$$

Therefore,

$$(\nabla \det\ \mathbf{f}(\theta_1,\theta_2))|_{(0,0)} = \begin{bmatrix} \left.\frac{\partial \det \mathbf{f}(\theta_1,\theta_2)}{\partial \theta_1}\right|_{(0,0)} \\ \left.\frac{\partial \det \mathbf{f}(\theta_1,\theta_2)}{\partial \theta_2}\right|_{(0,0)} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix},$$

and

$$(H_{\det \mathbf{f}})|_{(0,0)} = \begin{bmatrix} \left.\frac{\partial^2 \det \mathbf{f}(\theta_1,\theta_2)}{\partial {\theta_1}^2}\right|_{(0,0)} & \left.\frac{\partial^2 \det \mathbf{f}(\theta_1,\theta_2)}{\partial \theta_1 \partial \theta_2}\right|_{(0,0)} \\ \left.\frac{\partial^2 \det \mathbf{f}(\theta_1,\theta_2)}{\partial \theta_2 \partial \theta_1}\right|_{(0,0)} & \left.\frac{\partial^2 \det \mathbf{f}(\theta_1,\theta_2)}{\partial {\theta_2}^2}\right|_{(0,0)} \end{bmatrix} = \begin{bmatrix} \frac{53}{3912} & 0 \\ 0 & \frac{53}{3912} \end{bmatrix},$$

that is the Hessian matrix $(H_{\det \mathbf{f}})|_{(0,0)}$ is positive definite. As a consequence,

$$\det \mathbf{f}(\theta_1,\theta_2) = \det \mathbf{f}(\theta_1,\theta_2)|_{(0,0)} + (\nabla \det \mathbf{f}(\theta_1,\theta_2))^T|_{(0,0)} \begin{bmatrix}\theta_1\\ \theta_2\end{bmatrix} +$$
$$+\frac{1}{2}\begin{bmatrix}\theta_1\\ \theta_2\end{bmatrix}^T (H_{\det \mathbf{f}})|_{(0,0)} \begin{bmatrix}\theta_1\\ \theta_2\end{bmatrix} + o\left(\|\theta\|_2^2\right), = \frac{53}{3912}(\theta_1^2+\theta_2^2) + o\left(\|\theta\|_2^2\right),$$

where $\|\theta\|_2^2 = \theta_1^2 + \theta_2^2$.

Hence, in a neighborhood of $(0,0)$ $\det \mathbf{f}(\theta_1,\theta_2)$ behaves as a quadratic form and

$$\lim_{\|\theta\|_2 \to 0} \frac{\det \mathbf{f}(\theta_1,\theta_2)}{\|\theta\|_2^2} = \frac{53}{3912},$$

which means that $\det \mathbf{f}(\theta_1,\theta_2)$ and then $\lambda^{(1)}(\mathbf{f})$ have a zero of order 2 in $(0,0)$, as confirmed by Figure II.8 and Figure II.3(a), respectively. Finally, in the light of the third Item of Theorem

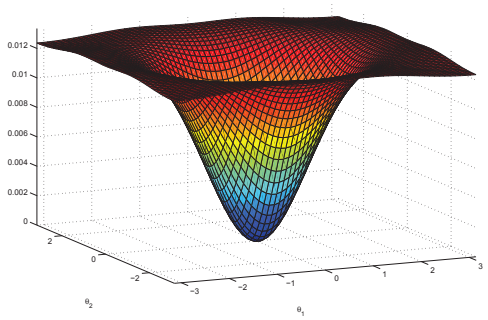


Figure II.8: $\det \mathbf{f}(\theta_1,\theta_2)$, $(\theta_1,\theta_2) \in \mathcal{I}_2$

I.4.4, we conclude that the minimal eigenvalue of $T_{\mathbf{n}}(\mathbf{f})$ goes to zero as $(N(\mathbf{n}))^{-1}$.

II.2.4 Spectral analysis of K_N via low rank perturbations

In this subsection we study the extremal behaviour of the matrix K_N, by making a careful analysis of the low rank matrix $E_{\mathbf{n}}$, defined in Section II.2.1. In particular, we show that $E_{\mathbf{n}}$ affects the number of outliers of K_N but does not influence the behaviour of the minimum eigenvalue of K_N with respect to that of $T_{\mathbf{n}}(\mathbf{f})$.
As shown in Section II.2.1, the matrix K_N is the sum of two Hermitian matrices, $T_{\mathbf{n}}(\mathbf{f})$ and $E_{\mathbf{n}}$. The structural and the spectral feature of $T_{\mathbf{n}}(\mathbf{f})$ have already been discussed in Section II.2.1, while $E_{\mathbf{n}}$ is a block diagonal matrix with $9n \times 9n$ block diagonal blocks. In particular there are just 3 types of nonzero blocks in the matrix $E_{\mathbf{n}}$.

1. $E_{\mathbf{n}}^{(l)}$, that is in the top left corner,
2. $E_{\mathbf{n}}^{(r)}$, that is in the bottom right corner,
3. $E_{\mathbf{n}}^{(c)}$, that is repeated $n-2$ times in the centre of the matrix.

We will prove that $E_{\mathbf{n}}^{(l)}$, $E_{\mathbf{n}}^{(r)}$ are positive definite, while $E_{\mathbf{n}}^{(c)}$ is nonnegative definite. This allows us to conclude that $E_{\mathbf{n}}$ is a nonnegative definite matrix.
Let us start by observing that $E_{\mathbf{n}}^{(l)}$, $E_{\mathbf{n}}^{(r)}$ and $E_{\mathbf{n}}^{(c)}$ are block diagonal themselves with 9×9 diagonal blocks. In detail, $E_{\mathbf{n}}^{(l)}$ and $E_{\mathbf{n}}^{(r)}$ are composed by n blocks of fixed dimension 9×9, $e_i^{(l)}$ and $e_i^{(r)}$, $i = 1, \ldots, n$, respectively, ordered in non decreasing way from the top left to the bottom right. Moreover we have

$$e_i^{(l)} = e_{i+1}^{(l)}, \quad i = 2, \ldots, n-2 \tag{II.22}$$

$$e_i^{(r)} = e_{i+1}^{(r)}, \quad i = 2, \ldots, n-2 \tag{II.23}$$

and

$$e_1^{(r)} = \mathcal{J} e_n^{(l)} \mathcal{J},$$

$$e_n^{(r)} = \mathcal{J} e_1^{(l)} \mathcal{J}$$

$$e_i^{(r)} = \mathcal{J} e_i^{(l)} \mathcal{J}, \quad i = 2, \ldots, n-1$$

where $\mathcal{J}$ is the 9×9 Hankel flip-matrix

$$\mathcal{J} = \begin{bmatrix} & & & & 1 \\ & & & 1 & \\ & & \iddots & & \\ & 1 & & & \end{bmatrix}.$$

Note that $\mathcal{J} = \mathcal{J}^{-1}$, then

$$e_1^{(r)} \sim e_n^{(l)}, \tag{II.24}$$

$$e_n^{(r)} \sim e_1^{(l)}, \tag{II.25}$$

$$e_i^{(r)} \sim e_i^{(l)}, \qquad i = 2, \ldots, n-1. \tag{II.26}$$

A direct computation shows that $e_1^{(l)}$, $e_2^{(l)}$, $e_n^{(l)}$ are positive definite, therefore according to relations (II.22)-(II.23) and (II.24)-(II.26) we can conclude that $E_{\mathbf{n}}^{(l)}$, $E_{\mathbf{n}}^{(r)}$ are positive definite.

The matrix $E_{\mathbf{n}}^{(c)}$ has only 2 nonzero 9×9 blocks, $e_1^{(c)}$, $e_n^{(c)}$ in the top left and bottom right corner respectively, such that

$$e_n^{(c)} = \mathcal{J} e_1^{(c)} \mathcal{J}, \tag{II.27}$$

while

$$e_i^{(c)} = O_9, \quad i = 2, \ldots, n-1.$$

Because of equation (II.27) it holds that

$$e_1^{(c)} \sim e_n^{(c)},$$

then, checking directly that $e_1^{(c)}$ is positive definite, we have proved that $E_{\mathbf{n}}^{(c)}$ is nonnegative definite.

Summarizing, since $E_{\mathbf{n}}^{(l)}$, $E_{\mathbf{n}}^{(r)}$ are positive definite, while $E_{\mathbf{n}}^{(c)}$ is nonnegative definite, we can conclude that $E_{\mathbf{n}}$ is a nonnegative definite matrix.
Let

$$\lambda_1(T_{\mathbf{n}}(\mathbf{f})) \leq \lambda_2(T_{\mathbf{n}}(\mathbf{f})) \leq \cdots \leq \lambda_N(T_{\mathbf{n}}(\mathbf{f}))$$

be the eigenvalues of $T_{\mathbf{n}}(\mathbf{f})$. Since $E_{\mathbf{n}}$ is nonnegative definite, the Interlacing Theorem [13] , applied to the matrices K_N, $T_{\mathbf{n}}$ and $E_{\mathbf{n}}$, leads to the relation

$$\lambda_j(T_{\mathbf{n}}(\mathbf{f})) \leq \lambda_j(K_N) \leq \lambda_{\gamma+j}(T_{\mathbf{n}}(\mathbf{f})) \tag{II.28}$$

for $1 \leq j \leq N - \gamma$, where γ is the rank of $E_{\mathbf{n}}(\mathbf{f})$.
This relation is useful for the study of the conditioning of the matrix K_N.
As shown in the last subsection

$$\lambda_1(T_{\mathbf{n}}(\mathbf{f})) \overset{\mathbf{n}\to\infty}{\sim} (N(\mathbf{n}))^{-1},$$

and in addition, from Section II.2.1, $\{K_N\}_N \sim_\lambda (\mathbf{f}, \mathcal{I}_2)$ and $\lambda_1(\mathbf{f}(0,0)) = 0$, with $\mathbf{f}$ nonnegative definite. Hence the minimum eigenvalue of K_N, $\lambda_1(K_N)$, has to go to zero.
The relation (II.28) provides a lower bound for the convergence speed of $\lambda_1(K_N)$ to zero, in fact, choosing in (II.28) $j = 1$,

$$\lambda_1(T_{\mathbf{n}}(\mathbf{f})) \leq \lambda_1(K_N), \tag{II.29}$$

and this implies that $\lambda_1(K_N)$ does not go to zero faster than $\lambda_1(T_{\mathbf{n}}(\mathbf{f}))$.
This means that the system (II.10) has the coefficient matrix K_N with a better conditioning, with respect to that of the matrix $T_{\mathbf{n}}(\mathbf{f})$, which is quadratic with the inverse of the mesh size.
In Subsection II.2.2, Table II.2, we have seen that the ratio between the number of outliers of K_N with respect to the sampling of $\lambda^{(9)}(\mathbf{f})$ and $\sqrt{9n^2}$ is constantly equal to $\frac{4}{3}$, so the number of outliers of K_N is $\frac{4}{3}\sqrt{9n^2} = 4n$.

Because of the fact that the matrix $E_{\mathbf{n}}$ is a block diagonal matrix with precisely $2n+2(n-2) = 4n-4$ of its 9×9 blocks positive definite, we have that $E_{\mathbf{n}}$ has exactly

$$9(2n) + 9(2(n-2)) = 36n - 36$$

linearly independent rows and then γ grows exactly as $36n - 36$ (see Table II.3).
This value is greater than the number of outliers, but asymptotically has the same order and the latter is in line with the theoretical forecasts induced by the Interlacing Theorem.

II.2.5 Further variations

The numerical tests in Subsection II.2.2 are done using Dirichlet pressure boundary conditions everywhere and a standard nodal approach of conforming continuous finite elements, in order to develop the basis functions (the Lagrange interpolation polynomials passing through the given set of nodes), which are needed to compute the values in K_N.
Two simple but important changes can be considered, but their detailed analysis will be the subject of future research:

n	$Rank(E_{\mathbf{n}}(\mathbf{f}))$
10	324
15	504
20	684
25	864
30	1044
35	1224
40	1404

Table II.3: $Rank(E_{\mathbf{n}}(\mathbf{f}))$ with increasing n

- using periodic boundary conditions;
- considering another standard basis of Lagrange interpolation polynomials, passing through the Gauss-Legendre quadrature points.

The first is motivated by the fact that several important numerical tests use this kind of boundary condition, the second one by the fact that this important kind of polynomial basis constitute an orthogonal basis. In this way the mass matrices used in the numerical method become diagonal and hence require less memory and computational effort (see, e.g., [65]). Here we give some details on the first Item.

Indeed, if we use periodic boundary conditions, then we obtain a sequence of linear systems analogous to (II.10) of the form

$$C_N x = b, \quad C_N \in \mathbb{R}^{N\times N}, \quad x, b \in \mathbb{R}^N. \tag{II.30}$$

The symmetric matrix $C_N \equiv C_{\mathbf{n}}(\mathbf{f})$ is the circulant matrix generated by the symbol $\mathbf{f} : \mathcal{I}_2 \to \mathbb{C}^{s\times s}$, $s = (p+1)^2$, described in Section II.2.1

$$\mathbf{f}(\theta_1, \theta_2) = \hat{\mathbf{f}}_{(0,0)} + \hat{\mathbf{f}}_{(-1,0)} e^{-\mathbf{i}\theta_1} + \hat{\mathbf{f}}_{(0,-1)} e^{-\mathbf{i}\theta_2} + \hat{\mathbf{f}}_{(1,0)} e^{\mathbf{i}\theta_1} + \hat{\mathbf{f}}_{(0,1)} e^{\mathbf{i}\theta_2},$$

Because $\mathbf{f}$ is a trigonometric polynomial, taking into account Theorem I.6.2 and Remark 3, for $\mathbf{n}$ sufficiently large we have

$$C_{\mathbf{n}}(\mathbf{f}) = (F_{\mathbf{n}} \otimes I_s) D_{\mathbf{n}}(\mathbf{f}) (F_{\mathbf{n}} \otimes I_s)^*, \tag{II.31}$$

with $D_{\mathbf{n}}(\mathbf{f})$ as in (I.22).
In (II.31), as stated in Theorem I.6.2, the matrix $F_{\mathbf{n}} \otimes I_s$ is unitary and $D_{\mathbf{n}}(\mathbf{f})$ is a block diagonal matrix with Hermitian blocks, $\mathbf{f}\left(\theta_{\mathbf{r}}^{(\mathbf{n})}\right)$, so we have

$$\Lambda(C_{\mathbf{n}}(\mathbf{f})) = \left\{ \lambda^{(l)} \left(\mathbf{f}\left(\theta_{\mathbf{r}}^{(\mathbf{n})}\right)\right) \, : \, \mathbf{r} = \mathbf{0}, \ldots, \mathbf{n} - \mathbf{e}\,; l = 1 \ldots, s \right\}, \tag{II.32}$$

where, for a fixed $\theta_{\mathbf{r}}^{(\mathbf{n})}$, $\lambda^{(l)} \left(\mathbf{f}\left(\theta_{\mathbf{r}}^{(\mathbf{n})}\right)\right)$ $l = 1 \ldots, s$ are the eigenvalues of $\mathbf{f}\left(\theta_{\mathbf{r}}^{(\mathbf{n})}\right)$.
Fixed $\mathbf{n} = (n_1, n_2)$, with $n_1 = n_2 = n$, and $p = 2$ the eigenvalues of C_N, with $N = 9n^2$, are a

sampling of the eigenvalue functions $\lambda^{(1)}(\mathbf{f}),\dots,\lambda^{(9)}(\mathbf{f})$ on a equispaced grid on $[0,2\pi]^2$,

$$J_n = \left\{ \left(\frac{2\pi j}{n}, \frac{2\pi k}{n} \right), \quad j,k = 0,\dots,n-1 \right\}.$$

Regarding the case of a possible change of the basis functions used for representing our numerical solution, we just mention that the new coefficient matrix is of the form $\tilde{K}_N = T_{\mathbf{n}}(\tilde{\mathbf{f}}) + \tilde{E}_{\mathbf{n}}$, with the same dimensions and structure seen in (II.11) but with different coefficients. The symbol $\tilde{\mathbf{f}}$ is again a trigonometric polynomial of the form described before and we obtain, with the same argument, $\{\tilde{K}_N\}_N \sim_\lambda (\tilde{\mathbf{f}}, \mathcal{I}_2)$. However, the analytical behavior of $\tilde{\mathbf{f}}$ has to be studied in detail and this will be considered in a future work.

II.3 Numerical experiments

In this section we numerically verify the spectral properties derived in Section II.2 on several applications of the staggered DG method [65] for the incompressible Navier-Stokes equations (II.1)-(II.2). In particular we evaluate the computational effort needed for solving the main linear system for the calculation of the discrete pressure using successive refinements of a regular grid with $n := n_1 = n_2 = \ldots = n_k$ on a square computational domain Ω. From the analysis given in Section II.2 we expect a condition number $\kappa = \kappa(K_N) \approx cN^{\frac{2}{k}}$ (the analysis has been done for $k = 2$ but it is easily extendible to any $k > 2$) where k represents the space dimension, $N = n^k(p+1)^k$ is the matrix size, p the polynomial degree of the DG discretization, and c is a positive real constant. Due to the use of the CG method and the spectral distribution/conditioning results, the expected number of iterations for reaching a precision ϵ can be expressed as

$$Iter(n) \approx \frac{1}{2}\sqrt{c}\log\left(\frac{2||r_0||}{\epsilon}\right)(p+1)n \qquad k = 2,3. \tag{II.33}$$

where $r_0 = \mathbf{p}^{\tau+\delta\tau} - \mathbf{p}_0^{\tau+\delta\tau}$ is the initial residual between the numerical solution $\mathbf{p}$ at the new time step $\tau+\delta\tau$ and the initial guess for the CG method that is indicated with $\mathbf{p}_0^{\tau+\delta\tau}$. In particular we will use a trivial initial guess $\mathbf{p}_0^{\tau+\delta\tau} = b^\tau$ or a better one that is based on the solution at the previous time τ, i.e. $\mathbf{p}_0^{\tau+\delta\tau} = \mathbf{p}^\tau$. In the following we will indicate with the term "*IG*" this second choice for the initial guess. Furthermore, ϵ is set to 10^{-8} for all the simulations.

II.3.1 Taylor Green vortex

First of all we take a classical test problem, the two and three dimensional Taylor Green vortex. The initial condition is given by

$$u(\mathbf{x},0) = \sin(x)\cos(y), \quad v(\mathbf{x},0) = -\cos(x)\sin(y), \quad p(\mathbf{x},0) = \frac{1}{4}\left[\cos(2x)+\cos(2y)\right], \tag{II.34}$$

for $k = 2$ and

$$u(\mathbf{x},0) = \sin(x)\cos(y)\cos(z), \qquad v(\mathbf{x},0) = -\cos(x)\sin(y)\cos(z),$$
$$w(\mathbf{x},0) = 0, \qquad p(\mathbf{x},0) = \frac{1}{16}\left[\cos(2x)+\cos(2y)\right]\left[\cos(2z)+2\right], \tag{II.35}$$

for $k = 3$. The behavior of the solution for $k = 3$ was numerically studied by Brachet et al. in [22] and consists in a fast generation of small scale structures, whose kinetic energy dissipation was monitored for several Reynolds numbers, see, e.g., [22, 65, 137]. For $k = 2$ and short times there is an analytical representation of the energy dissipation due to friction phenomena and hence this test can be used to check the accuracy of the numerical algorithm, see [65]. We consider $\Omega = [0, 2\pi]^k$; $\delta\tau = 5 \cdot 10^{-3}$; $\tau_{end} = 2$; Reynolds number $Re = 800$ and periodic boundary conditions everywhere. The resulting final pressure at $\tau = \tau_{end}$ is shown in Figure $II.9$ for $k = 2$ and 3. The obtained average number of iterations needed to compute the solution is reported in Table $II.4$ and Figure $II.10$ for the two particular choices of $\mathbf{p}_0^{\tau+\delta\tau} = b^\tau$ and a better initial guess $\mathbf{p}_0^{\tau+\delta\tau} = \mathbf{p}^\tau$. The expected linear behavior for both two and three dimensional case is achieved according to equation (II.33). Note that the choice of the initial guess $\mathbf{p}_0^{\tau+\delta\tau} = \mathbf{p}^\tau$ becomes particularly good when the solution is steady or quasi-steady, since $\mathbf{p}^{\tau+\delta\tau} - \mathbf{p}^\tau \approx (K_N)^{-1}(b^\tau - b^{\tau-\delta\tau})$ and $b^\tau - b^{\tau-\delta\tau}$ contains essentially the variation of the convective-viscous contribution. Hence, for quasi stationary problems or small perturbations around a steady state, $\mathbf{p}^\tau$ is a good candidate for the initial guess of the CG algorithm. In practice, what we observe is indeed that the needed number of iterations tends to decrease due to a better choice of the initial guess, as suggested in equation (II.33). Note, however, that the asymptotic behavior remains the same, i.e. linear in n, see Figure $II.10$ for a graphical representation.

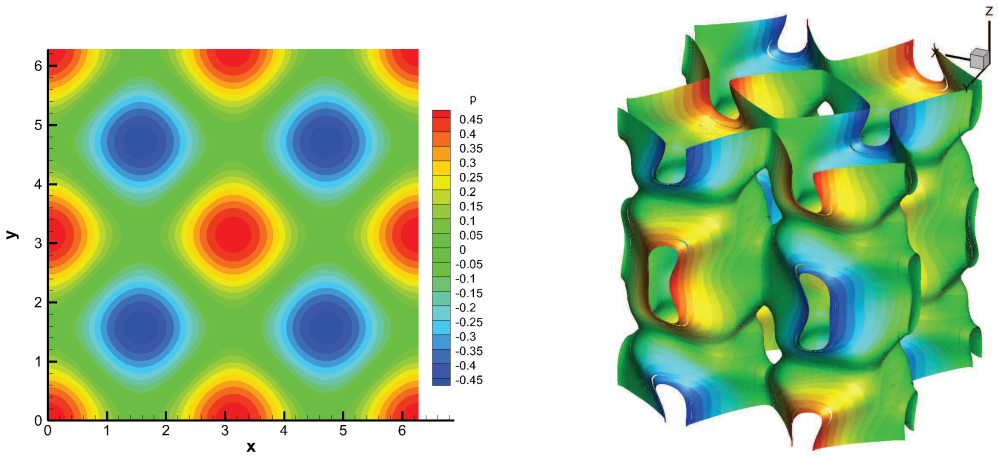


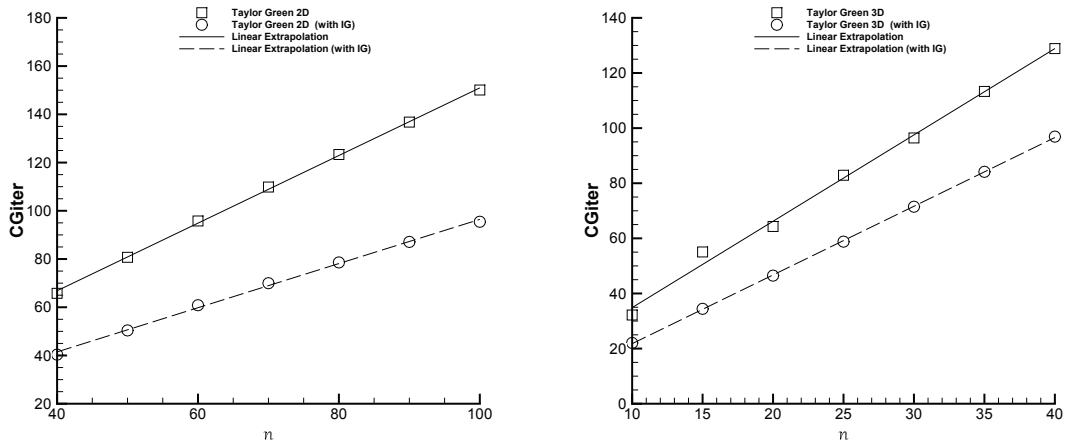
Figure II.9: Pressure profiles at $\tau = \tau_{end}$. Pressure contours for $k = 2$ (left) and isosurfaces for $k = 3$ (right).

II.3.2 Modified double shear layer

The previous test manifests at $\tau_{end} = 2$ a relatively complex behavior for $k = 3$ but a simple one involving sinusoidal functions for $k = 2$. In this section we want to test the behavior of the number of iterations in a variant of the classical $2D$ double shear layer originally studied in [12]. For this test case we consider the same initial condition as the one used in [136]. In the original study there is a regular jet region with $\mathbf{v} = (1, 0)$ in a fluid with velocity $\mathbf{v} = (-1, 0)$. The flow is characterized by two shear layers with high velocity gradient in the y-direction. This steady state is physically unstable due to the Kelvin-Helmholtz instability and tends to generate also

$k=2$				$k=3$			
n	N	Iter	Iter with IG	n	N	Iter	Iter with IG
40	14400	65.8	40.3	10	27000	32.2	22.1
50	22500	80.7	50.3	15	91125	55.1	34.6
60	32400	95.8	60.8	20	216000	64.3	46.5
70	44100	109.8	69.9	25	421875	82.9	58.8
80	57600	123.3	78.5	30	729000	96.4	71.5
90	72900	136.7	87.0	35	1157625	113.3	84.1
100	90000	150.0	95.4	40	1728000	128.9	96.6

Table II.4: Resulting average number of CG iterations for $\tau \in [0,2]$ with the choice of $\mathbf{p}_0^{\tau+\delta\tau} = b^\tau$ (Iter) and the use of the initial guess $\mathbf{p}_0^{\tau+\delta\tau} = \mathbf{p}^\tau$ (Iter with IG) for $k = 2, 3$.



(a) $k = 2$ (b) $k = 3$

Figure II.10: Resulting average number of CG iterations as a function of n with and without the IG initial guess compared with the linear extrapolation of the data, for $k = 2, 3$.

in this case vortical structures close to the shear layers. In order to drive this instability, a small perturbation is introduced in the vertical velocity directly at $\tau = 0$. In [12] the evolution of this instability was performed for periodic boundary conditions everywhere.

For this test we take $p = 2$, $\tau_{end} = 1$; $Re = 800$ but pressure boundary condition everywhere in order to introduce the important perturbation matrix $E_{\mathbf{n}}$ discussed in Section $II.2.1$. In this case we expect a slightly different behavior with respect to what is observed in [12], owing to the use of a different type of boundary conditions. In any case the resulting pressure field will not maintain a simple sinusoidal structure for $k = 2$. The resulting numerical solution at $\tau = \tau_{end}$ for the finest grid is reported in Figure $II.11$ while the obtained average number of iterations is shown in Table $II.5$ and the corresponding plot in Figure $II.12$. As expected, also in this case the behavior for the number of iterations is linear with respect to $N^{1/k}$.

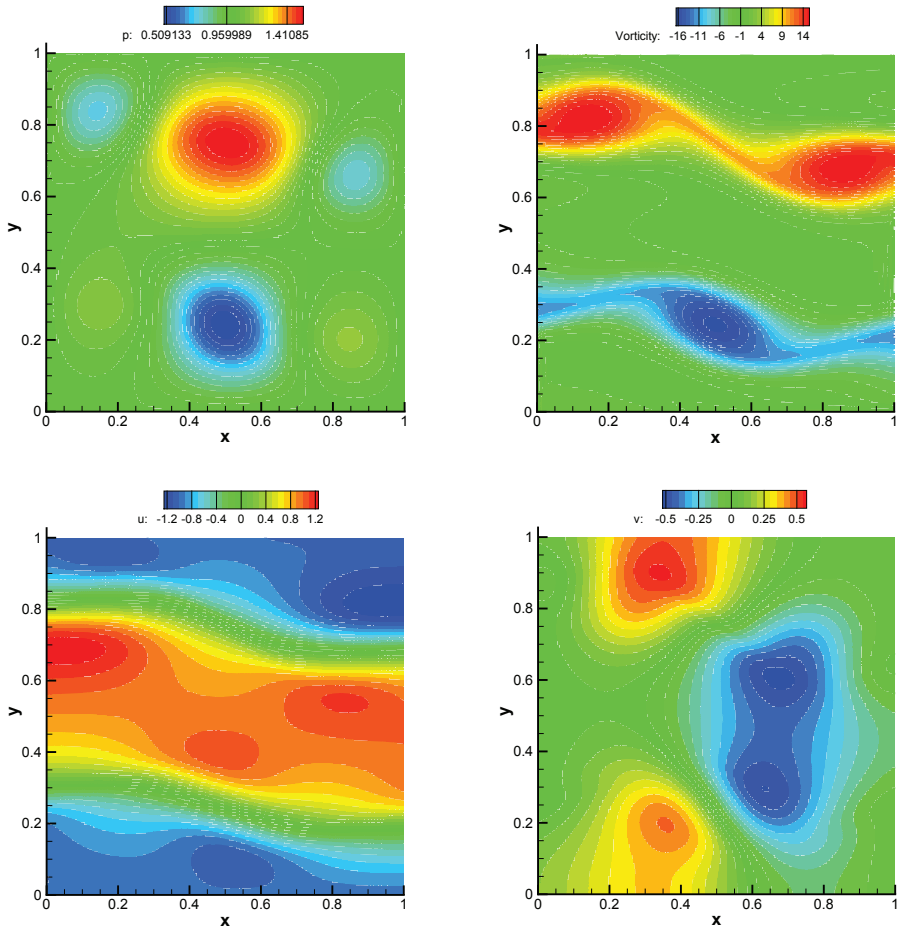


Figure II.11: Numerical solution for the modified double shear layer at $\tau = 1$. Top: the pressure and the vorticity. Bottom: from left to right, u and v velocity component, respectively.

II.3.3 Preconditioning

A simple preconditioner is based on the use of the two-level circulant matrix $C_{\mathbf{n}}(\mathbf{f})$ that is directly associated to the fully periodic boundary case. In this case we can choose as preconditioner the matrix $C_{\mathbf{n}}(\mathbf{f})$ with the Strang correction $P_{\mathbf{n}}(\mathbf{f}) = C_{\mathbf{n}}(\mathbf{f}) + \mathbf{e}\mathbf{e}^{\top}\frac{1}{N^2}$ where $\mathbf{e}^T = (1, \ldots, 1)$ is the

n	N	Iter	Iter with IG
10	900	98.7	74.1
20	3600	195.9	138.9
30	8100	297.1	201.8
40	14400	400.1	264.3
50	22500	504.0	325.3
60	32400	607.1	386.3
70	44100	711.3	447.0

Table II.5: Resulting average number of CG iterations for $\tau \in [0, 1]$ with and without the IG initial guess.

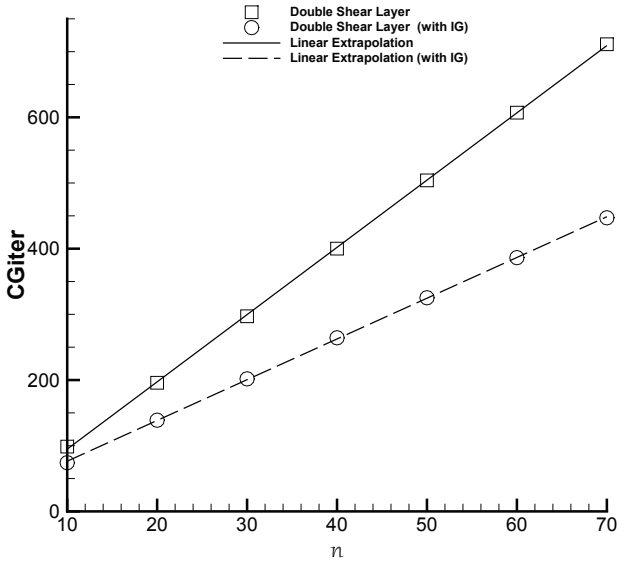


Figure II.12: Average number of CG iterations as function of n obtained in the modified double shear layer test case with and without the IG initial guess compared with the linear extrapolation of the data.

$N-$dimensional unitary vector. The inverse of this matrix is still a circulant matrix and so its computation can be done at the cost of $O(N \log N)$. In this section we want to investigate the impact of this simple preconditioner on the number of iterations in the complete case where the coefficient matrix is K_N (see Subsections II.2.2.1, II.2.2.2). For this test we take the same framework as in the previous numerical experiment, using $\mathbf{p}^\tau$ as initial guess. The resulting number of iterations is reported in Table II.6

The use of this preconditioner drastically reduces the number of iterations as well as the behavior that seems to be sub-linear and almost flat with respect to the case without preconditioner, see Figure $II.13$. A comment on the latter fact has to be made. First of all we observe that the matrix sequences $\{K_N\}_{\mathbf{n}}$ and $\{P_{\mathbf{n}}(\mathbf{f})\}_{\mathbf{n}}$ share the same spectral symbol $\mathbf{f}$, by Items **GLT1**, **GLT2**, and **GLT4**, since the rank of the differences $K_N - T_{\mathbf{n}}(\mathbf{f}) = E_{\mathbf{n}}$ and $P_{\mathbf{n}}(\mathbf{f}) - T_{\mathbf{n}}(\mathbf{f})$ grows as $N^{1/2}$. As a consequence, since $\mathbf{f}$ is singular only at a unique point, that is $(0, 0)$, again by Item **GLT2** we deduce that the preconditioned matrix-sequence $\{P_{\mathbf{n}}^{-1}(\mathbf{f})K_N\}_{\mathbf{n}}$ is a GLT

n	N	Iter (CG method)	Iter (PCG method)
10	900	74.1	24.6
20	3600	138.9	30.1
30	8100	201.8	33.3
40	14400	264.3	35.8
50	22500	325.3	38.3

Table II.6: Resulting average number of iterations for $\tau \in [0,1]$ with CG and PCG whose preconditioner is the 2-level circulant $P_{\mathbf{n}}(\mathbf{f})$.

sequence with symbol 1. Since for every N, $\mathbf{n}$, the matrix $P_{\mathbf{n}}^{-1}(\mathbf{f})K_N$ is similar to a symmetric we deduce that 1 is the spectral symbol of the preconditioned matrix-sequence. In addition the analysis of the rank corrections tells that the number of outlying eigenvalues grows at most as $O(N^{1/2})$. However, from the classical theory of the (preconditioned) CG convergence we know that small outliers negatively affect the convergence more than large outliers. Now, since the coefficient matrix K_N can be written as $T_{\mathbf{n}}(\mathbf{f}) + E_{\mathbf{n}}$ and since we proved that $E_{\mathbf{n}}$ is nonnegative definite, we expect that the outliers are large (as practically observed in the numerical experiments) and this is the reason why the number of iterations seems to grow slower than the number of estimated outlying eigenvalues.

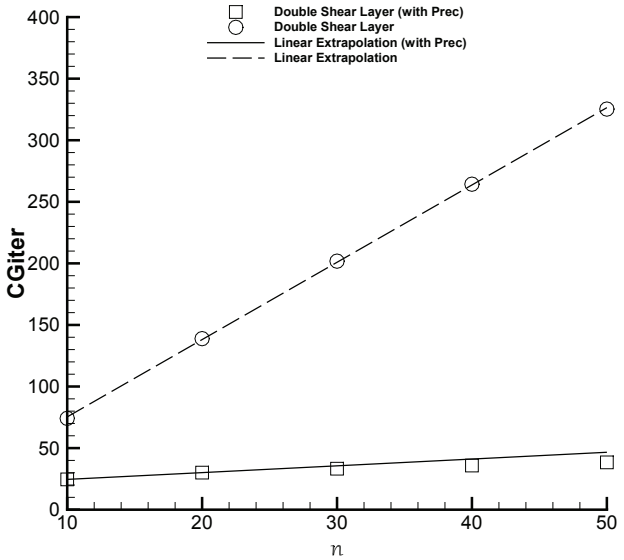


Figure II.13: Average number of iterations obtained in the modified double shear layer test case with CG and PCG whose preconditioner is the 2-level circulant $P_{\mathbf{n}}(\mathbf{f})$.

Let us now take a look at the gain in terms of CPU time obtained by the use of this simple preconditioner. Since $P_{\mathbf{n}}(\mathbf{f})$ is a circulant matrix, we can diagonalize it as FDF^* where $F = F_n \otimes F_n \otimes F_9$ is the three-level Fourier matrix and D is a block diagonal matrix. We can then use the Fast Fourier Transform (FFT) to construct the matrix $D = F^* P_{\mathbf{n}}(\mathbf{f}) F$ and then D^{-1} by

inverting each single block. Once D^{-1} is known, we can easily compute $P_{\mathbf{n}}^{-1}(\mathbf{f})x = F^* D^{-1} F x$ using the three-level FFT algorithm to compute first $x_1 = Fx$ at the cost of $O(N \log N)$. Then we have to compute $x_2 = D^{-1} x_1$ at a linear cost and finally we obtain $P_{\mathbf{n}}^{-1}(\mathbf{f})x = F^* x_2$ again at the cost of $O(N \log N)$. A particular test when we can really take advantage of this procedure is the fully periodic case so that the considered test becomes the classical double shear layer test case. The resulting total CPU time as well as the total CPU time needed to compute only the linear system is reported in Table II.7 for the fully periodic case (i.e. classical double shear layer). In Table II.8 we report the obtained results for the case with pressure boundary conditions everywhere (i.e. modified double shear layer).

		No Preconditioner				With Preconditioner			
n	N_{step}	T_{TOT}	$\frac{T_{TOT}}{N_{step}}$	T_{LS}	$\frac{T_{LS}}{N_{step}}$	T_{TOT}	$\frac{T_{TOT}}{N_{step}}$	T_{LS}	$\frac{T_{LS}}{N_{step}}$
32	709	195.51	0.28	126.92	0.18	77.56	0.11	6.42	0.01
64	1411	2298.6	1.63	1793.5	1.27	609.06	0.43	48.76	0.03
128	2829	31284.	11.06	27218.	9.62	4434.81	1.57	367.12	0.13

Table II.7: Number of time steps N_{step}, total and relative (small numbers) CPU time for the solution of the main linear system for the pressure (T_{LS}) and the entire CPU time (T_{TOT}) for fully periodic boundary conditions. Note that in this test $p = 2$, $k = 2$ and $N = (p+1)^k n^k$.

		No Preconditioner				With Preconditioner			
n	N_{step}	T_{TOT}	$\frac{T_{TOT}}{N_{step}}$	T_{LS}	$\frac{T_{LS}}{N_{step}}$	T_{TOT}	$\frac{T_{TOT}}{N_{step}}$	T_{LS}	$\frac{T_{LS}}{N_{step}}$
32	696	371.36	0.53	296.50	0.43	219.20	0.31	142.20	0.20
64	1410	4868.4	3.45	4280.4	3.04	2034.6	1.44	1419.9	1.01
128	2853	72713.	25.49	67693.	23.73	20509.	7.19	15320.	5.37

Table II.8: Number of time steps N_{step}, total and relative (small numbers) CPU time for the solution of the main linear system for the pressure (T_{LS}) and the entire CPU time (T_{TOT}) for pressure boundary conditions everywhere. Note that in this test $p = 2$, $k = 2$ and $N = (p+1)^k n^k$.

As expected, since the symbol fully represents the periodic case, the gain on T_{LS} obtained by introducing the preconditioner is impressive. In fact, the computational cost is *essentially* the cost of a fully explicit formula for large N. In fact, a solution method in this case is just the application of a standard inversion formula for block circulant matrices.

In the worst case where we introduce pressure boundary conditions everywhere, we observe a gain factor $T_{LS}^{nopre}/T_{LS}^{pre}$ of 2.0, 3.0, 4.4 for $n = 32, 64, 128$, respectively. Hence, the advantage of using the basic Strang-type preconditioner suggested by our spectral analysis is verified both for periodic and non periodic case.

II.3.4 A multigrid approach

The PCG procedure defined in the previous subsection is practically effective, but still there is room for improvements: in fact, the cost of each PCG iteration is $O(N \log N)$ because of the use

of a block Fast Fourier Transform (FFT) and the number of iterations grows at most linearly with the partial sizes, for moderate matrix sizes, due to the rank of the difference between the actual matrix and the Strang-type preconditioner. In conclusion the computational cost results in $O(N^{3/2}\log N)$, which is not optimal. Here for optimality we mean a total cost for solving the linear system with a preassigned accuracy proportional to the cost of the matrix-vector product, where the matrix is the coefficient matrix and the vector is generic. Since the coefficient matrix is sparse the optimality amounts in a cost of $O(N)$.

By exploiting the spectral analysis provided so far, a way for recovering optimality relies in following a multigrid approach, which we briefly sketch below.

Consider the linear system

$$A_N x_N = b_N \tag{II.36}$$

where x_N, $b_N \in \mathbb{C}^N$, $A_N = \mathcal{W}_N - \mathcal{B}_N \in \mathbb{C}^N \times \mathbb{C}^N$, $\mathcal{W}_N$ non singular matrix. Let

$$x^{(j+1)} = V_N x^{(j)} + b_1 := V_N(x^{(j)}, b_1) \tag{II.37}$$

be an iterative method for the solution of system (II.36), where $b_1 := \mathcal{W}_N^{-1} b \in \mathbb{C}^N$ and $V_N := I_N - \mathcal{W}_N^{-1} A_N \in \mathbb{C}^N \times \mathbb{C}^N$. Let $p_N^M \in \mathbb{C}^N \times \mathbb{C}^M$ be a full-rank matrix, with $M < N$. A Two-Grid Method (TGM) is defined by the following algorithm [145]

1. $d_N = A_N x^{(j)} - b$
2. $d_M = (p_N^M)^* d_N$
3. $A_M = (p_N^M)^* A_N (p_N^M)$
4. Solve $A_M y = d_M$
5. $\hat{x}^{(j)} = x^{(j)} - p_N^M y$
6. $x^{(j+1)} = V_N^\mu(\hat{x}^{(j)}, b_1)$

Step 6 consists in applying the "smoothing iteration" (II.37) μ times while steps 1-5 define the "coarse grid correction", that depends only on the projection operator p_N^M. The global iteration matrix of the TGM is given by

$$TGM(V_N, p_N^M) = V_N^\mu \left[I - p_N^M \left(\left(p_N^M\right)^* A_N p_N^M \right)^{-1} \left(p_N^M\right)^* A_N \right].$$

We remind that, if step 4 is replaced by a recursive call to the same algorithm (until the size M is bounded from above by a fixed constant), then the scheme given before defines a V-cycle procedure.

Our idea is to follow the same proposal as in [128] for Toeplitz structures generated by a scalar-valued symbol, where the scalar generating function considered in [128] is replaced by the minimal eigenvalue function of our matrix-valued symbol. According to this choice, since the minimal eigenvalue function is of Laplacian type that is a nonnegative function with a unique zero at $(0,0)$ of order two, then the projector has the form

$$p_{\mathbf{n}}^{\mathbf{n}/2} = T_{\mathbf{n}}(p)(Z_{\mathbf{n}}^{\mathbf{n}/2} \otimes I_9),$$

where the generating function associated to the projector is

$$p(\theta_1, \theta_2) = \left[(2 + 2\cos(\theta_1))\,(2 + 2\cos(\theta_2)) \right] I_9 \in \mathbb{R}^{9\times 9}, \tag{II.38}$$

$$T_{\mathbf{n}}(p) = \begin{bmatrix} 2 & 1 & & & \\ 1 & 2 & 1 & & \\ & 1 & 2 & 1 & \\ & & & \ddots & 1 \\ & & & 1 & 2 \end{bmatrix} \otimes \begin{bmatrix} 2 & 1 & & & \\ 1 & 2 & 1 & & \\ & 1 & 2 & 1 & \\ & & & \ddots & 1 \\ & & & 1 & 2 \end{bmatrix} \otimes I_9, \tag{II.39}$$

$$Z_{\mathbf{n}}^{\mathbf{n}/2} = Z_{n_1}^{n_1/2} \otimes Z_{n_2}^{n_2/2},$$

and $Z_m^{m/2}$ is the $m \times \frac{m-1}{2}$ matrix given by

$$(Z_m^{m/2})_{i,j} = \begin{cases} 1 & \text{for } i = 2j \\ 0 & \text{otherwise} \end{cases} \quad i = 1, \ldots, m, \quad j = 1, \ldots, \frac{m-1}{2}, \tag{II.40}$$

with m of the form $m = 2^t - 1$, with t positive integer.

In the following ϵ is set to 10^{-8} for all the simulations, b and x_0 are the known term and the initial guess, respectively. We use as Pre/Post- smoother 1 iteration of Gauss-Seidel. In Table II.9 we compare the iterations of the block TGM with those of our PCG with Strang-type preconditioner, when increasing the size N. We can observe that the number of iterations of block TGM for achieving a precision ϵ remains constant, $cost = 18$, when increasing the size N. Here the right-hand side is the sampling of a smooth functions but no qualitative variations are observed with different choices.

n	$N = 9n^2$	PCG	TGM
15	2025	22	18
21	3969	26	18
25	5625	27	18
31	8649	28	18
35	11025	30	18
41	15129	32	18
45	18225	32	18

Table II.9: Number of iterations for $T_{\mathbf{n}}(\mathbf{f})$ provided by PCG and TGM with 1 Pre/Post-smoothing Gauss-Seidel iteration.

Now we check the TGM optimality in the complete case of

$$A_N = T_{\mathbf{n}}(\mathbf{f}) + E_{\mathbf{n}}.$$

Because of the fact that $E_{\mathbf{n}}$ is nonnegative definite and $T_{\mathbf{n}}(\mathbf{f})$ is positive definite we have

$$A_N \geq T_{\mathbf{n}}(\mathbf{f}); \tag{II.41}$$

Hence, according to the result in Remark 5 [127], we have that the same TGM, designed for $T_{\mathbf{n}}(\mathbf{f})$, has to be optimal also for A_N.

In Table II.10 we can observe that the number of iterations of block TGM, needed for achieve the tolerance ϵ, remains constant ($cost \approx 30$), when increasing the size n. Therefore the application

n	$N = 9n^2$	PCG	TGM
15	2025	40	29
21	3969	45	29
25	5625	47	30
31	8649	50	30
35	11025	51	30
41	15129	53	30
45	18225	54	30

Table II.10: Number of iterations for $A_N = T_{\mathbf{n}}(\mathbf{f}) + E_{\mathbf{n}}$ provided by PCG and TGM with 1 Pre/Post-smoothing Gauss-Seidel iteration.

of the V-cycle for solving a linear system with our Strang-type preconditioner will decrease the computational cost from $O(N^{3/2}\log N)$ to $O(N^{3/2})$, which is a slight improvement, while the use of the V-cycle algorithm directly on the original linear system in connection with the basic Gauss-Seidel smoother induces a $O(N)$ solver, that is a solver with an optimal cost. For the formal proof of convergence of the Two-Grid method it is enough to mimic the same steps as in [128], taking into account the order-relation results in [127] and the spectral study in which we proved that: **a)** the minimal eigenvalue function is of Laplacian type that is a nonnegative function with a unique zero at $(0,0)$ of order two, and **b)** $A_N \geq T_{\mathbf{n}}(\mathbf{f})$. The V-cycle analysis should follow the more advanced tools in [3], which again rely strongly on the analytical information regarding the spectral symbol studied in the previous sections. A formal study of these issues and more efficient combinations involving multigrid schemes and preconditioned Krylov techniques will be the subject of future researches.

Chapter III

Asymptotic Expansion: an algorithm for preconditioned matrices

III.1 Generalization of the preconditioned Asymptotic Expansion

The present chapter is devoted to present the asymptotic spectral expansion for the eigenvalues of preconditioned Toeplitz matrices $\mathcal{P}_n(f,g) = T_n^{-1}(g)T_n(f)$. We consider the case where f is a trigonometric polynomial, g is a nonnegative and not identically zero trigonometric polynomial.

We provide numerical evidence that few of the assumptions of [16, 17, 19] can be relaxed, accompanied by an appropriate error analysis and numerical experiments.

Main contributions

The main results of the Chapter can be summarized as follows.

1. We provide numerical evidence of a precise asymptotic expansion for the eigenvalues of $\mathcal{P}_n(f,g)$. Precisely, we show through numerical experiments that, under the assumption that $r = f/g$ is monotone, for every integer $\alpha \geq 0$, every n and every $j = 1, \ldots, n$, the following asymptotic expansion holds:

$$\lambda_j(\mathcal{P}_n(f,g)) = r(\theta_{j,n}) + \sum_{k=1}^{\alpha} c_k(\theta_{j,n})h^k + E_{j,n,\alpha}, \tag{III.1}$$

   where:

   - the eigenvalues of $\mathcal{P}_n(f,g)$ are arranged in nondecreasing or nonincreasing order, depending on whether r is increasing or decreasing;
   - $\{c_k\}_{k=1,2,\ldots}$ is a sequence of functions from $[0,\pi]$ to $\mathbb{R}$ which depends only on r;
   - $h = \frac{1}{n+1}$ and $\theta_{j,n} = \frac{j\pi}{n+1} = j\pi h$;
   - $E_{j,n,\alpha} = O(h^{\alpha+1})$ is the remainder (the error), which satisfies the inequality $|E_{j,n,\alpha}| \leq C_\alpha h^{\alpha+1}$ for some constant C_α depending only on α and r.

We refer the reader to the **Chapter VI** Section VI.2 for a proof of the expansion (III.1) for $\alpha = 0$.

We note that (III.1) is formally the same as the expansions for the eigenvalues of Toeplitz matrices, which have been conjectured and validated through numerical experiments in [62].

2. Based on the expansion (III.1) and drawing inspiration from [61], we propose a parallel interpolation–extrapolation algorithm for computing the eigenvalues of $\mathcal{P}_n(f,g)$. The computation is performed for very large n, when the eigenvalues of $\mathcal{P}_{n_i}(f,g)$ have been computed, for moderate values of n_i, $i = 1,\dots,\alpha$, with α a fixed small number. The performance of the algorithm is illustrated through numerical experiments.

The context we consider is that of a scalar univariate generating function ϕ. In addition all the functions involved are real-valued, hence, by the properties seen in Section I.4.1, all the Toeplitz matrices $T_n(\phi)$ are Hermitian. From the results seen in **Chapter I** much is known regarding their spectral properties: from the localization of the eigenvalues to the asymptotic spectral distribution in the Weyl sense. Indeed we recall that, under these hypothesis ϕ is the spectral symbol of $\{T_n(\phi)\}_n$, see [20, 77] and the references therein.

In addition, if ϕ is real-valued and not identically constant, then any eigenvalue of $T_n(\phi)$ belongs to the open set (m_ϕ, M_ϕ), with m_ϕ, M_ϕ being the essential infimum, the essential supremum of ϕ, respectively. Notice that the case of a constant ϕ is trivial: in that case if $\phi = \gamma$ almost everywhere then $T_n(\phi) = \gamma I_n$.

Hence if $M_\phi > 0$ and ϕ is nonnegative almost everywhere, then $T_n(\phi)$ is Hermitian positive definite.

In this chapter we focus our attention on the following setting.

- We consider two real-valued cosine trigonometric polynomials (RCTPs) f, g, that is

$$f(\theta) = \hat{f}_0 + 2\sum_{k=1}^{d_1} \hat{f}_k \cos(k\theta), \qquad \hat{f}_0, \hat{f}_1, \dots, \hat{f}_{d_1} \in \mathbb{R}, \qquad d_1 \in \mathbb{N},$$

$$g(\theta) = \hat{g}_0 + 2\sum_{k=1}^{d_2} \hat{g}_k \cos(k\theta), \qquad \hat{g}_0, \hat{g}_1, \dots, \hat{g}_{d_2} \in \mathbb{R}, \qquad d_2 \in \mathbb{N},$$

  so that $T_n(f)$, $T_n(g)$ are both real symmetric.

- We assume that $M_g = \max g > 0$ and $m_g = \min g \geq 0$, so that $T_n(g)$ is positive definite.

- We consider $\mathcal{P}_n(f,g) = T_n^{-1}(g)T_n(f)$ the "preconditioned" matrix and we define the new symbol $r = f/g$.

The nth Toeplitz matrix generated by $\phi \in \{f, g\}$ is the real symmetric banded matrix of

bandwidth $2d+1$, $d \in \{d_1, d_2\}$ ($d = d_1$ if $\phi = f$ and $d = d_2$ if $\phi = g$), given by

$$T_n(\phi) = \begin{bmatrix}
\hat{\phi}_0 & \hat{\phi}_1 & \cdots & \hat{\phi}_d & & & & & & & \\
\hat{\phi}_1 & \ddots & \ddots & & \ddots & & & & & & \\
\vdots & \ddots & \ddots & \ddots & & \ddots & & & & & \\
\hat{\phi}_d & & \ddots & \ddots & \ddots & & \ddots & & & & \\
& \ddots & & \ddots & \ddots & \ddots & & \ddots & & & \\
& & \boxed{\hat{\phi}_d} & \cdots & \boxed{\hat{\phi}_1} & \boxed{\hat{\phi}_0} & \boxed{\hat{\phi}_1} & \cdots & \boxed{\hat{\phi}_d} & & \\
& & & \ddots & & \ddots & \ddots & \ddots & & \ddots & \\
& & & & \ddots & & \ddots & \ddots & \ddots & & \hat{\phi}_d \\
& & & & & \ddots & & \ddots & \ddots & \ddots & \vdots \\
& & & & & & \ddots & & \ddots & \ddots & \hat{\phi}_1 \\
& & & & & & & \hat{\phi}_d & \cdots & \hat{\phi}_1 & \hat{\phi}_0
\end{bmatrix}.$$

Matrices of the form $\mathcal{P}_n(f,g)$ are important for the fast solution of large Toeplitz linear systems (in connection with the preconditioned conjugate gradient method [32, 34, 47, 116] or of more general preconditioned Krylov methods [88, 89]). Furthermore, up to low rank corrections, they appear in the context of the spectral approximation of differential operators in which a low rank correction of $T_n(g)$ is the mass matrix and a low rank correction of $T_n(f)$ is the stiffness matrix.

Their spectral features have been studied in detail. More precisely, under the assumption that $r = \gamma$ identically $\mathcal{P}_n(f,g) = \gamma I_n$. When $m_r < M_r$ then any eigenvalue of $\mathcal{P}_n(f,g)$ belongs to the open set (m_r, M_r), see [47], and the whole sequence $\{\mathcal{P}_n(f,g)\}_n$ is spectrally distributed in the Weyl sense as $r = f/g$ (see [118]).

In our context, we say that a function is monotone if it is either increasing or decreasing over the interval $[0, \pi]$.

Under the assumption that $r = f/g$ is monotone, we show experimentally in this chapter that for every integer $\alpha \geq 0$, every n and every $j = 1, \ldots, n$, the following asymptotic expansion holds:

$$\lambda_j(\mathcal{P}_n(f,g)) = r(\theta_{j,n}) + \sum_{k=1}^{\alpha} c_k(\theta_{j,n}) h^k + E_{j,n,\alpha}, \tag{III.2}$$

where:

- the eigenvalues of $\mathcal{P}_n(f,g)$ are arranged in nondecreasing or nonincreasing order, depending on whether r is increasing or decreasing;
- $\{c_k\}_{k=1,2,\ldots}$ is a sequence of functions from $[0,\pi]$ to $\mathbb{R}$ which depends only on r;
- $h = \frac{1}{n+1}$ and $\theta_{j,n} = \frac{j\pi}{n+1} = j\pi h$;
- $E_{j,n,\alpha} = O(h^{\alpha+1})$ is the remainder (the error), which satisfies the inequality $|E_{j,n,\alpha}| \leq C_\alpha h^{\alpha+1}$ for some constant C_α depending only on α and r.

In the pure Toeplitz case, that is for $g = 1$ identically, so that $\mathcal{P}_n(f,g) = T_n(f)$ and $r = f$, the result is proven in [16, 17, 19], if the RCTP f is monotone and satisfies certain additional assumptions, which include the requirements that $f'(\theta) \neq 0$ for $\theta \in (0,\pi)$ and $f''(\theta) \neq 0$ for $\theta \in \{0,\pi\}$. The symbols

$$f_q(\theta) = (2 - 2\cos\theta)^q, \qquad q = 1, 2, \ldots, \tag{III.3}$$

arise in the discretization of differential equations and are therefore of particular interest. Unfortunately, for these symbols the requirement that $f''(0) \neq 0$ is not satisfied if $q \geq 2$. In [62] several numerical evidences are reported, showing that the higher order approximation (III.2) holds even in this "degenerate case".

Here, as a first purpose, we show numerically the same for the preconditioned matrices $\mathcal{P}_n(f,g)$ and, from a theoretical point of view, the numerical testing is complemented in Section VI.2 of the **Chapter VI** by the proof of the above conjecture in the basic case of $\alpha = 0$.

Furthermore, in [62], the authors employed the asymptotic expansion (III.2) for computing an accurate approximation of $\lambda_j(T_n(f))$ for very large n, provided that the values

$$\lambda_{j_1}(T_{n_1}(f)), \ldots, \lambda_{j_\alpha}(T_{n_\alpha}(f))$$

are available for moderate sizes $n_1, \ldots, n_s$ with $\theta_{j_1,n_1} = \cdots = \theta_{j_\alpha,n_\alpha} = \theta_{j,n}$, $\alpha \geq 2$. The second and main purpose of this chapter is to carry out this idea and to support it by numerical experiments, accompanied by an appropriate error analysis in the more general case of the preconditioned matrices $\mathcal{P}_n(f,g)$. In particular, we devise an algorithm to compute $\lambda_j(\mathcal{P}_n(f,g))$ with a high level of accuracy and a relatively low computational cost. The algorithm is completely analogous to the extrapolation procedure, which is employed in the context of Romberg integration (to obtain high precision approximations of an integral from a few coarse trapezoidal approximations [132, Section 3.4], see also [23] for more advanced algorithms). In this regard, the asymptotic expansion (III.2) plays here the same role as the Euler–Maclaurin summation formula [132, Section 3.3].

The third and last purpose of this chapter is to formulate, on the basis of numerical experiments, a conjecture on the higher-order asymptotic of the eigenvalues if the monotonicity assumption on $r = f/g$ is not in force. We also illustrate how this conjecture can be used along with our extrapolation algorithm in order to compute some of the eigenvalues of $\mathcal{P}_n(f,g)$ in the case where r is non-monotone.

III.2 Implicit Errors expansion

The proposed approach is based on the classical concept of the symbol, but with an innovative view on the errors of the approximation of eigenvalues by the uniform sampling of the symbol. In particular our advantage is that of manipulating the error expression implicitly given in (III.2). In fact, if we assume that the relations in (III.2) hold, then we can write

$$E_{j,n,0} = \sum_{k=1}^{\alpha} c_k(\theta_{j,n})\, h^k + E_{j,n,\alpha}\,, \tag{III.4}$$

where $E_{j,n,0} = \lambda_j(\mathcal{P}_n(f,g)) - r(\theta_{j,n})$.

We now suppose to know the eigenvalues for different (small) n_i namely

$$\{(n_1, \lambda_{j_1}(\mathcal{P}_{n_1}(f,g))), (n_2, \lambda_{j_2}(\mathcal{P}_{n_2}(f,g))), \cdots, (n_\alpha, \lambda_{j_\alpha}(\mathcal{P}_{n_\alpha}(f,g)))\},$$

where $n_1, n_2, \cdots, n_\alpha$ and $j_1, j_2, \cdots, j_\alpha$ are chosen in such a way that $j_1/(n_1+1) = j_2/(n_2+1) = \cdots = j_\alpha/(n_\alpha+1)$.

By defining $h_1 = 1/(n_1+1), h_2 = 1/(n_2+1), \ldots, h_\alpha = 1/(n_\alpha+1)$, for a given set of eigenvalues, equation (III.4) can be written as

$$\begin{aligned}
E_{j_1,n_1,0} &= \sum_{k=1}^{\alpha} c_k(\theta_{j_1,n_1})\, h_1^k + E_{j_1,n_1,\alpha},\\
E_{j_2,n_2,0} &= \sum_{k=1}^{\alpha} c_k(\theta_{j_2,n_2})\, h_2^k + E_{j_2,n_2,\alpha},\\
E_{j_3,n_3,0} &= \sum_{k=1}^{\alpha} c_k(\theta_{j_3,n_3})\, h_3^k + E_{j_3,n_3,\alpha},\\
&\vdots\\
E_{j_\alpha,n_\alpha,0} &= \sum_{k=1}^{\alpha} c_k(\theta_{j_\alpha,n_\alpha})\, h_\alpha^k + E_{j_\alpha,n_\alpha,\alpha}.
\end{aligned} \tag{III.5}$$

Let c, $\tilde{c}$ be the vectors

$$c = [c_1, c_2, \ldots, c_\alpha]^T; \qquad \tilde{c} = [\tilde{c}_1, \tilde{c}_2, \ldots, \tilde{c}_\alpha]^T,$$

and let A be the coefficient matrix of size $\alpha \times \alpha$ with components $A_{i,j} = h_i^j$. Hence the set of equations (III.5) can be written in matrix form as

$$Ac = b_0 - b_\alpha\,, \tag{III.6}$$

where $b_0 = [E_{j_1,n_1,0}, E_{j_2,n_2,0}, \ldots, E_{j_\alpha,n_\alpha,0}]^T$ and $b_\alpha = [E_{j_1,n_1,\alpha}, E_{j_2,n_2,\alpha}, \ldots, E_{j_\alpha,n_\alpha,\alpha}]^T$. Furthermore, by neglecting the higher order errors, we may define an approximation $\tilde{c}$ of c according to the expression below

$$A\tilde{c} = b_0\,. \tag{III.7}$$

In the next we analyse in more detail the properties of the matrix A. It must be highlighted that the matrix involved is typically ill-conditioned. However, the approximation of c is easily obtained by solving the linear system of equations above, since the matrix size is in practice very small.

Indeed assume we are interested in calculating the eigenvalues, which are of the order $\mathcal{O}(1)$, of a large matrix, for example, of the order $\mathcal{O}(10^6)$, and with c_k function of the order $\mathcal{O}(1)$. Then, when using the approximated $\tilde{c}_k$ and $h = \mathcal{O}(10^{-6})$, we have the term $\tilde{c}_3 h^3 = \mathcal{O}(10^{-18})$ in the algorithm, which is beyond machine precision of the order $\mathcal{O}(10^{-16})$, for 64 bit double precision computations. Therefore it is sufficient using a small α in the asymptotic expansion (and consequently a small size of A) for reaching an accurate approximation of the c_k functions, using double precision arithmetic computations.

III.2.1 Error bounds for the coefficients c_k in the Asymptotic Expansion

In the current subsection we derive upper-bounds for $|\tilde{c} - c|$: in reality, equations (III.6) and (III.7) leads to

$$A(\tilde{c} - c) = b_\alpha. \tag{III.8}$$

If we define $\Delta c = \tilde{c} - c$ and $\eta_i = \frac{E_{j_i,n_{j_i},\alpha}}{h_i^{\alpha+1}}$ for $i = 1, \ldots, \alpha$, then the system (III.8) can be written as

$$A\Delta c = \begin{bmatrix} \eta_1 h_1^{\alpha+1} \\ \eta_2 h_2^{\alpha+1} \\ \vdots \\ \eta_\alpha h_\alpha^{\alpha+1} \end{bmatrix}, \tag{III.9}$$

with $|\eta_i| \leq C_\alpha$ for $i = 1, \ldots, \alpha$, where C_α is a constant. The coefficient matrix can be expressed as

$$A = \begin{bmatrix} h_1 & h_1^2 & \ldots & h_1^\alpha \\ h_2 & h_2^2 & \ldots & h_2^\alpha \\ \vdots & \vdots & & \vdots \\ h_\alpha & h_\alpha^2 & \ldots & h_\alpha^\alpha \end{bmatrix} = \begin{bmatrix} h_1 & & & \\ & h_2 & & \\ & & \ddots & \\ & & & h_\alpha \end{bmatrix} V(h_1, \ldots, h_\alpha),$$

where $V(h_1, \ldots, h_\alpha)$ is the Vandermonde matrix of order α corresponding to $h_1, \ldots, h_\alpha$.
By assuming $W = V^{-1}(h_1, \ldots, h_\alpha)$, we deduce

$$(W)_{i,j} = \begin{cases} (-1)^{\alpha-i} \left(\dfrac{\displaystyle\sum_{\substack{1 \leq k_1 < \ldots < k_{\alpha-i} \leq \alpha \\ k_1, \ldots, k_{\alpha-i} \neq j}} h_{k_1} \cdots h_{k_{\alpha-i}}}{\displaystyle\prod_{\substack{1 \leq k \leq \alpha \\ k \neq j}} (h_j - h_k)} \right) & 1 \leq i < \alpha, \\ \dfrac{1}{\displaystyle\prod_{\substack{1 \leq k \leq \alpha \\ k \neq j}} (h_j - h_k)} & i = \alpha. \end{cases} \tag{III.10}$$

Therefore for the inversion of the matrix A we have

$$(A^{-1})_{i,j} = \begin{cases} (-1)^{\alpha-i} \left(\dfrac{\displaystyle\sum_{\substack{1 \leq k_1 < \ldots < k_{\alpha-i} \leq \alpha \\ k_1, \ldots, k_{\alpha-i} \neq j}} h_{k_1} \cdots h_{k_{\alpha-i}}}{h_j \displaystyle\prod_{\substack{1 \leq k \leq \alpha \\ k \neq j}} (h_j - h_k)} \right) & 1 \leq i < \alpha, \\ \dfrac{1}{h_j \displaystyle\prod_{\substack{1 \leq k \leq \alpha \\ k \neq j}} (h_j - h_k)} & i = \alpha, \end{cases} \tag{III.11}$$

and we can obtain an explicit expression for $(\Delta c)_i$, $i = 1, \ldots, \alpha$, that is

$$(\Delta c)_i = \sum_{j=1}^{\alpha} (A^{-1})_{i,j} \eta_j h_j^{\alpha+1}. \tag{III.12}$$

Case 1. If $i = \alpha$, then

$$(\Delta c)_\alpha = \sum_{j=1}^{\alpha} \frac{\eta_j h_j^{\alpha+1}}{h_j \prod\limits_{\substack{1 \le k \le \alpha \\ k \neq j}} (h_j - h_k)}.$$

Whence, from the fact that $|\eta_i| \le C_\alpha$ for $i = 1, \dots, \alpha$,

$$|(\Delta c)_\alpha| \le \sum_{j=1}^{\alpha} \frac{|\eta_j| h_j^{\alpha+1}}{h_j \prod\limits_{\substack{1 \le k \le \alpha \\ k \neq j}} |h_j - h_k|} \le \sum_{j=1}^{\alpha} \frac{C_\alpha h_j^{\alpha}}{\prod\limits_{\substack{1 \le k \le \alpha \\ k \neq j}} |h_j - h_k|}.$$

With the choice $h_j = \frac{1}{\gamma^{j-1}} h_1$ for $j = 1, \dots, \alpha$, γ positive integer, we have

$$|(\Delta c)_\alpha| \le C_\alpha \sum_{j=1}^{\alpha} \frac{(\frac{h_1}{\gamma^{j-1}})^\alpha}{\prod\limits_{\substack{1 \le k \le \alpha \\ k \neq j}} h_1 \left| \frac{1}{\gamma^{j-1}} - \frac{1}{\gamma^{k-1}} \right|} = C_\alpha h_1^\alpha \sum_{j=1}^{\alpha} \frac{(\frac{1}{\gamma^{j-1}})^\alpha}{h_1^{\alpha-1} \prod\limits_{\substack{1 \le k \le \alpha \\ k \neq j}} \left| \frac{1}{\gamma^{j-1}} - \frac{1}{\gamma^{k-1}} \right|} =$$

$$= h_1 C_\alpha \sum_{j=1}^{\alpha} \frac{(\frac{1}{\gamma^{j-1}})^\alpha}{\prod\limits_{\substack{1 \le k \le \alpha \\ k \neq j}} \left| \frac{1}{\gamma^{j-1}} - \frac{1}{\gamma^{k-1}} \right|} = O(h_1).$$

Case 2. If $i = 1, \dots, \alpha - 1$, then

$$(\Delta c)_i = \sum_{j=1}^{\alpha} (-1)^{\alpha-i} \eta_j h_j^{\alpha+1} \frac{\sum\limits_{\substack{1 \le k_1 < \dots < k_{\alpha-i} \le \alpha \\ k_1, \dots, k_{\alpha-i} \neq j}} h_{k_1} \cdots h_{k_{\alpha-i}}}{h_j \prod\limits_{\substack{1 \le k \le \alpha \\ k \neq j}} (h_j - h_k)},$$

that is different from the case $i = \alpha$ just for the numerator

$$\sum_{\substack{1 \le k_1 < \dots < k_{\alpha-i} \le \alpha \\ k_1, \dots, k_{\alpha-i} \neq j}} h_{k_1} \cdots h_{k_{\alpha-i}}.$$

As a consequence

$$|(\Delta c)_i| \le C_\alpha \sum_{j=1}^{\alpha} h_j^\alpha \frac{\sum\limits_{\substack{1 \le k_1 < \dots < k_{\alpha-i} \le \alpha \\ k_1, \dots, k_{\alpha-i} \neq j}} h_{k_1} \cdots h_{k_{\alpha-i}}}{\prod\limits_{\substack{1 \le k \le \alpha \\ k \neq j}} |h_j - h_k|}.$$

With the choice $h_j = \frac{1}{\gamma^{j-1}} h_1$ for $j = 1, \ldots, \alpha$, we infer

$$|(\Delta c)_i| \leq C_\alpha \sum_{j=1}^{\alpha} \left(\frac{h_1}{\gamma^{j-1}}\right)^\alpha \frac{\sum_{\substack{1 \leq k_1 < \ldots < k_{\alpha-i} \leq \alpha \\ k_1, \ldots, k_{\alpha-i} \neq j}} h_1^{\alpha-i} \left(\frac{1}{\gamma^{k_1-1}} \frac{1}{\gamma^{k_2-1}} \cdots \frac{1}{\gamma^{k_{\alpha-i}-1}}\right)}{\prod_{\substack{1 \leq k \leq \alpha \\ k \neq j}} h_1 \left|\frac{1}{\gamma^{j-1}} - \frac{1}{\gamma^{k-1}}\right|}$$

$$= C_\alpha \sum_{j=1}^{\alpha} \left(\frac{1}{\gamma^{j-1}}\right)^\alpha \left(\frac{h_1^\alpha h_1^{\alpha-i}}{h_1^{\alpha-1}}\right) \frac{\sum_{\substack{1 \leq k_1 < \ldots < k_{\alpha-i} \leq \alpha \\ k_1, \ldots, k_{\alpha-i} \neq j}} \left(\frac{1}{\gamma^{k_1-1}} \frac{1}{\gamma^{k_2-1}} \cdots \frac{1}{\gamma^{k_{\alpha-i}-1}}\right)}{\prod_{\substack{1 \leq k \leq \alpha \\ k \neq j}} \left|\frac{1}{\gamma^{j-1}} - \frac{1}{\gamma^{k-1}}\right|}$$

$$= h_1^{\alpha-i+1} C_\alpha \sum_{j=1}^{\alpha} \left(\frac{1}{\gamma^{j-1}}\right)^\alpha \frac{\sum_{\substack{1 \leq k_1 < \ldots < k_{\alpha-i} \leq \alpha \\ k_1, \ldots, k_{\alpha-i} \neq j}} \left(\frac{1}{\gamma^{k_1-1}} \frac{1}{\gamma^{k_2-1}} \cdots \frac{1}{\gamma^{k_{\alpha-i}-1}}\right)}{\prod_{\substack{1 \leq k \leq \alpha \\ k \neq j}} \left|\frac{1}{\gamma^{j-1}} - \frac{1}{\gamma^{k-1}}\right|} = O(h_1^{\alpha-i+1}).$$

As a conclusion, with the choice $h_j = \frac{1}{\gamma^{j-1}} h_1$ for $j = 1, \ldots, \alpha$ and under the assumption that the asymptotic expansion reported in (III.2) is true, we deduce

$$|(\Delta c)_i| = O(h_1^{\alpha-i+1}), \tag{III.13}$$

for $i = 1, \ldots, \alpha$.

III.3 Error bounds for numerically approximated eigenvalues

The goal of this section is to provide error bounds based on the linear system in (III.7) for the computation of the eigenvalues of $\mathcal{P}_n(f, g)$: of course these error bounds are based on the conjecture that the relations reported in (III.2) are true. However, as we can see in Section III.4, the numerical tests fully support the existence of the considered asymptotic expansion.

Indeed, as already observed, by solving (III.7), we can compute the approximations $\tilde{c}_k$ of c_k. Once we have the values of $\tilde{c}_k$, we can calculate the eigenvalues $\tilde{\lambda}_{j_\beta}$ of a large dimension matrix of size n_β, here $n_\beta + 1 = \gamma^{\beta-1}(n_1 + 1)$. The asymptotic expansion (III.4) can be written as

$$E_{j_\beta, n_\beta, 0} = \bar{h}_\beta^T c + E_{j_\beta, n_\beta, \alpha}\,. \tag{III.14}$$

By subtraction $\bar{h}_\beta^T \tilde{c}$ from both sides of the equation above, we find

$$\begin{aligned}
E_{j_\beta, n_\beta, 0} - \bar{h}_\beta^T \tilde{c} &= \bar{h}_\beta^T (c - \tilde{c}) + E_{j_\beta, n_\beta, \alpha}, \\
\lambda_j(\mathcal{P}_{n_\beta}(f, g)) - r(\theta_{j, n_\beta}) - \bar{h}_\beta^T \tilde{c} &= \bar{h}_\beta^T \Delta c + E_{j_\beta, n_\beta, \alpha}, \\
\left|\lambda_j(\mathcal{P}_{n_\beta}(f, g)) - r(\theta_{j, n_\beta}) - \bar{h}_\beta^T \tilde{c}\right| &\leq \sum_{i=1}^{\alpha} h_\beta^i |(\Delta c)_i| + |E_{j_\beta, n_\beta, \alpha}|, \\
\left|\lambda_j(\mathcal{P}_{n_\beta}(f, g)) - r(\theta_{j, n_\beta}) - \bar{h}_\beta^T \tilde{c}\right| &\leq \sum_{i=1}^{\alpha} h_\beta^i |(\Delta c)_i| + C_\alpha h_\beta^{\alpha+1},
\end{aligned} \tag{III.15}$$

where $\bar{h}_\beta = [h_\beta, h_\beta^2, \cdots, h_\beta^\alpha]^T$, $|E_{j_\beta,n_\beta,\alpha}| \leq C_\alpha h_\beta^{\alpha+1}$ for some constant C_α and $|(\Delta c)_i|$ is given in (III.13).

III.4 Numerical tests

In this section we want to present a few numerical experiments to support the asymptotic expansion (III.2) in the case where one or more properties of the following list are satisfied:

1. $f''(0) \neq 0$ (see Example 1, Example 3, and Example 5),
2. $f''(0) = 0$ (see Example 2 and Example 4),
3. $\min g > 0$ (see Example 1, Example 2, and Example 5),
4. $\min g = 0$ (see Example 3 and Example 4),
5. $r = f/g$ is non-monotone (see Example 5).

The approximation of eigenvalues of large matrices in each case is also computed. The expansion (III.2) for $\alpha = 4$ is

$$\begin{aligned}\lambda_j(\mathcal{P}_n(f,g)) &= r(\theta_{j,n}) + c_1(\theta_{j,n})\,h + c_2(\theta_{j,n})\,h^2 + c_3(\theta_{j,n})\,h^3 + c_4(\theta_{j,n})\,h^4 + E_{j,n,4},\\ E_{j,n,0} &= \lambda_j(\mathcal{P}_n(f,g)) - r(\theta_{j,n}) = c_1(\theta_{j,n})\,h + c_2(\theta_{j,n})\,h^2 + c_3(\theta_{j,n})\,h^3 + c_4(\theta_{j,n})\,h^4 + E_{j,n,4}\,.\end{aligned} \tag{III.16}$$

In all numerical examples we choose four matrix-size values, that is n_i for $i \in \{1,2,3,4\}$, in a way that they satisfy $n_i = \gamma^{i-1}(n_1+1) - 1$, with γ being a positive integer. The expansion (III.16) for the set of the four dimensions n_i can be written as

$$\begin{aligned}E_{j_1,n_1,0} &= c_1(\theta_{j_1,n_1})\,h_1 + c_2(\theta_{j_1,n_1})\,h_1^2 + c_3(\theta_{j_1,n_1})\,h_1^3 + c_4(\theta_{j_1,n_1})\,h_1^4 + E_{j_1,n_1,4},\\ E_{j_2,n_2,0} &= c_1(\theta_{j_2,n_2})\,h_2 + c_2(\theta_{j_2,n_2})\,h_2^2 + c_3(\theta_{j_2,n_2})\,h_2^3 + c_4(\theta_{j_2,n_2})\,h_2^4 + E_{j_2,n_2,4},\\ E_{j_3,n_3,0} &= c_1(\theta_{j_3,n_3})\,h_3 + c_2(\theta_{j_3,n_3})\,h_3^2 + c_3(\theta_{j_3,n_3})\,h_3^3 + c_4(\theta_{j_3,n_3})\,h_3^4 + E_{j_3,n_3,4},\\ E_{j_4,n_4,0} &= c_1(\theta_{j_4,n_4})\,h_4 + c_2(\theta_{j_4,n_4})\,h_4^2 + c_3(\theta_{j_4,n_4})\,h_4^3 + c_4(\theta_{j_4,n_4})\,h_4^4 + E_{j_4,n_4,4},\end{aligned} \tag{III.17}$$

where $h_i = \frac{1}{n_i+1}$ and $j_i = \gamma^{i-1}\,j_1$ for $i \in \{1,2,3,4\}$. Notice that $\theta_{j_i,n_i} = \theta_{j_1,n_1} = \bar{\theta}$ for a fixed $j_1 \in \{1,2,\cdots,n_1\}$. We are interested in the numerical approximation of $c_i(\bar{\theta})$ for $i \in \{1,2,3,4\}$ and then in the precise numerical approximation of the eigenvalue of $\mathcal{P}_n(f,g)$ for large n. The set of equations (III.17) can be written as

$$\begin{aligned}E_{j_1,n_1,0} &= \tilde{c}_1(\bar{\theta})\,h_1 + \tilde{c}_2(\bar{\theta})\,h_1^2 + \tilde{c}_3(\bar{\theta})\,h_1^3 + \tilde{c}_4(\bar{\theta})\,h_1^4,\\ E_{j_2,n_2,0} &= \tilde{c}_1(\bar{\theta})\,h_2 + \tilde{c}_2(\bar{\theta})\,h_2^2 + \tilde{c}_3(\bar{\theta})\,h_2^3 + \tilde{c}_4(\bar{\theta})\,h_2^4,\\ E_{j_3,n_3,0} &= \tilde{c}_1(\bar{\theta})\,h_3 + \tilde{c}_2(\bar{\theta})\,h_3^2 + \tilde{c}_3(\bar{\theta})\,h_3^3 + \tilde{c}_4(\bar{\theta})\,h_3^4,\\ E_{j_4,n_4,0} &= \tilde{c}_1(\bar{\theta})\,h_4 + \tilde{c}_2(\bar{\theta})\,h_4^2 + \tilde{c}_3(\bar{\theta})\,h_4^3 + \tilde{c}_4(\bar{\theta})\,h_4^4.\end{aligned} \tag{III.18}$$

We solve the system of linear equations above for $j_1 \in \{1,2,\cdots,n_1\}$ to compute $\tilde{c}_i(\bar{\theta})$. The computed $\tilde{c}_i$ are used to approximate the eigenvalues of large size n_β by exploiting the following relation

$$\tilde{\lambda}_{j_\beta}(\mathcal{P}_{n_\beta}(f,g)) = r(\theta_{j_\beta,n_\beta}) + \bar{h}_\beta^T \tilde{c}\,. \tag{III.19}$$

Example 1. *Let g, f, and r be the functions defined as*

$$f(\theta) = 4 - 2\cos(\theta) - 2\cos(2\theta) = (2 - 2\cos(\theta))(3 + 2\cos(\theta))\,,$$
$$g(\theta) = 3 + 2\cos(\theta)\,,$$
$$r(\theta) = \frac{f(\theta)}{g(\theta)} = 2 - 2\cos(\theta)\,,$$

where $\theta \in [0, \pi]$. The graphs of generating functions are shown in the left panel of Figure III.1, and the approximations $\tilde{c}_k$, for $k = 1, 2, 3, 4$ are shown in the right panel. Note that $g(\theta) > 0$, $\forall\, \theta \in [0, \pi]$, $f''(0) \neq 0$, and furthermore $r(\theta)$ is monotone. We set $n_1 \in \{40, 60, 80, 100\}$ and $\gamma = 2$.

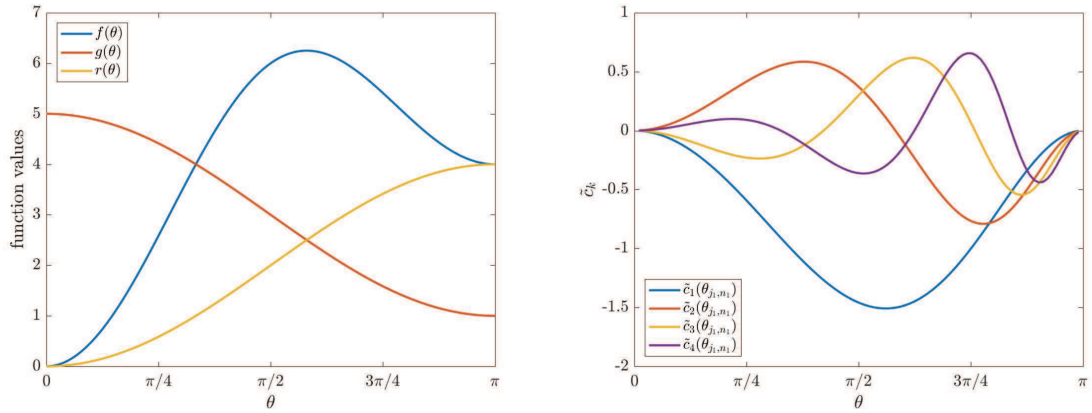


Figure III.1: Example 1: Generating functions (f, g, and r) and $\tilde{c}_k$ for $k = 1, 2, 3, 4$.

Example 2. *Let g, f, and r be the functions defined as*

$$f(\theta) = 20 - 30\cos(\theta) + 12\cos(2\theta) - 2\cos(3\theta) = (2 - 2\cos(\theta))^3\,,$$
$$g(\theta) = 3 + 2\cos(\theta)\,,$$
$$r(\theta) = \frac{f(\theta)}{g(\theta)} = \frac{(2 - 2\cos(\theta))^3}{3 + 2\cos(\theta)}\,,$$

where $\theta \in [0, \pi]$. The graphs of generating functions are shown in the left panel of Figure III.2, and the approximations $\tilde{c}_k$, for $k = 1, 2, 3, 4$ are shown in the right panel. Remark that $g(\theta) > 0$, $\forall\, \theta \in [0, \pi]$, $f''(0) = 0$, and furthermore $r(\theta)$ is monotone. We set $n = n_1 \in \{40, 60, 80, 100\}$ and $\gamma = 2$.

There is an important issue to discuss here. Both the functions f and r attain the minimum at $\theta = 0$ with a very high order. Indeed we have $f(\theta), r(\theta) \approx \theta^6$, with $\phi_1 \approx \phi_2$ being the symmetric, transitive relation telling that there exist positive constants $c, C > 0$ such that $c\phi_1 \leq \phi_2 \leq C\phi_1$ on the whole definition domain $[0, \pi]$. Therefore for fixed j (independent of n) the jth smallest eigenvalue of $\mathcal{P}_n(f, g)$ is asymptotic to $k_j h^6$, k_j a positive constant depending on j but not on n: the reader is referred to [114] for the preconditioned case with the limitation $j = 1$ and to [8] and references therein for very elegant and precise estimates regarding the pure Toeplitz case.

Now if we fix j and we put together $\lambda_j\left(\mathcal{P}_n(f, g)\right) \approx h^6$ with relations (III.4)–(III.5) then the only possibility for avoiding a contradiction is that the functions $c_1(\theta), c_2(\theta), c_3(\theta), c_4(\theta), c_5(\theta)$ all vanish at $\theta = 0$.

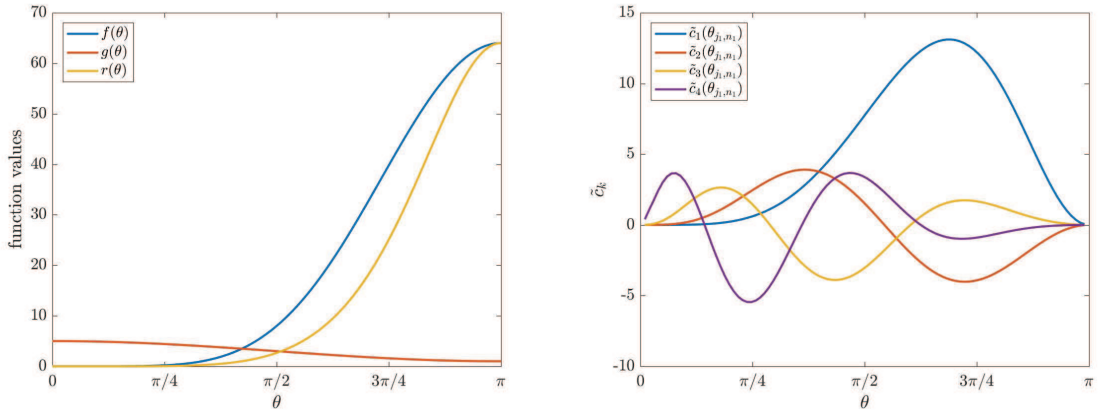


Figure III.2: Example 2: Generating functions ($f, g,$ and r) and $\tilde{c}_k$ for $k = 1, 2, 3, 4$.

The approximations $\tilde{c}_k$, for $k = 1, 2, 3, 4$ shown in the right panel of Figure III.2 are coherent with the above mathematical conclusion and in fact all these approximations vanish simultaneously at $\theta = 0$ (the fifth is not displayed, but we computed it and it also equals to zero at $\theta = 0$, while, as expected from an extension of the results by [8] to the preconditioned Toeplitz case, the sixth is nonzero at $\theta = 0$).

Since the argument and the conclusions are the very same, we anticipate that the discussion can be repeated verbatim for Example 4, where the functions f and r attain the minimum at $\theta = 0$ with order 10. As a consequence, we expect that the functions $c_1(\theta), \ldots, c_9(\theta)$ all simultaneously vanish at $\theta = 0$, while $c_{10}(0) \neq 0$: this is confirmed for the first four of them as reported in the right panel of Figure III.4.

Example 3. *Let g, f, and r be the functions defined as*

$$
\begin{aligned}
f(\theta) &= 1 + \cos(\theta) + \frac{1}{4}\cos(2\theta) + \frac{1}{5}\cos(3\theta) + \frac{1}{10}\cos(4\theta) + \frac{1}{10}\cos(5\theta)\,,\\
g(\theta) &= 2 - 2\cos(\theta)\,,\\
r(\theta) &= \frac{f(\theta)}{g(\theta)} = \frac{1 + \cos(\theta) + \frac{1}{4}\cos(2\theta) + \frac{1}{5}\cos(3\theta) + \frac{1}{10}\cos(4\theta) + \frac{1}{10}\cos(5\theta)}{2 - 2\cos(\theta)}\,,
\end{aligned}
$$

where $\theta \in [0, \pi]$. The graphs of generating functions are shown in the left panel of Figure III.3, and the approximations $\tilde{c}_k$, for $k = 1, 2, 3, 4$ are shown in the right panel. Note that $\min g(\theta) = 0$, $\forall\, \theta \in [0, \pi]$, $f''(0) \neq 0$, and furthermore $r(\theta)$ is monotone. We set $n = n_1 \in \{40, 60, 80, 100\}$ and $\gamma = 2$.

Example 4. *Let g, f, and r be the functions defined as*

$$
\begin{aligned}
f(\theta) &= 252 - 420\cos(\theta) + 240\cos(2\theta) - 90\cos(3\theta) + 20\cos(4\theta) - 2\cos(5\theta) = (2 - 2\cos(\theta))^5\,,\\
g(\theta) &= 2 + 2\cos(\theta)\,,\\
r(\theta) &= \frac{f(\theta)}{g(\theta)} = \frac{(2 - 2\cos(\theta))^5}{2 + 2\cos(\theta)}\,,
\end{aligned}
$$

where $\theta \in [0, \pi]$. The graphs of generating functions are shown in the left panel of Figure III.4, and the approximations $\tilde{c}_k$, for $k = 1, 2, 3, 4$ are shown in the right panel. Remark that $\min g(\theta) =$

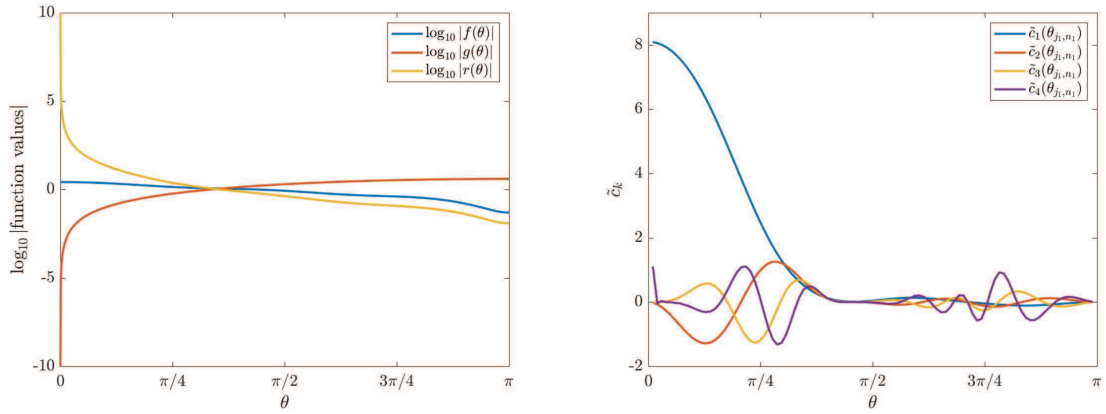


Figure III.3: Example 3: Generating functions ($f, g,$ and r) and $\tilde{c}_k$ for $k = 1, 2, 3, 4$.

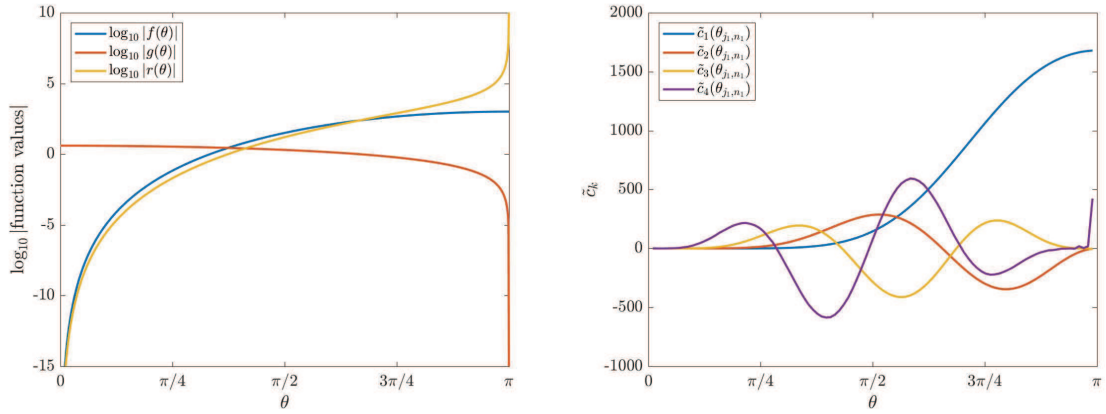


Figure III.4: Example 4: Generating functions ($f, g,$ and r) and $\tilde{c}_k$ for $k = 1, 2, 3, 4$.

0, $\forall\, \theta \in [0, \pi]$, $f''(0) = 0$, *and furthermore* $r(\theta)$ *is monotone. We set* $n = n_1 \in \{40, 60, 80, 100\}$ *and* $\gamma = 2$.

Example 5. *Let* g, f, *and* r *be the functions defined as*

$$f(\theta) = \frac{136}{17} + \frac{56}{17}\cos(\theta) - \frac{2}{17}\cos(2\theta) + \frac{5}{17}\cos(3\theta) = (3 - \cos(\theta) + \frac{5}{17}\cos(2\theta))(3 + 2\cos(\theta))\,,$$
$$g(\theta) = 3 + 2\cos(\theta)\,,$$
$$r(\theta) = \frac{f(\theta)}{g(\theta)} = 3 - \cos(\theta) + \frac{5}{17}\cos(2\theta)\,,$$

where $\theta \in [0, \pi]$. *The graphs of generating functions are shown in the left panel of Figure III.5, and the approximations* $\tilde{c}_k$, *for* $k = 1, 2, 3, 4$ *are shown in the right panel. Notice that* $\min g(\theta) > 0$, $\forall\, \theta \in [0, \pi]$, $f''(0) \neq 0$, *and furthermore* r *is non-monotone. We set* $n = n_1 \in \{40, 60, 80, 100\}$ *and* $\gamma = 2$.

The numerical tests related to Examples 1 and 2, as in Figures III.6 and III.7, show that the error expansion (III.2) behaves as expected. In Figure III.11 we also see that the approximated $\tilde{c}_k$ can be used for a large n to approximate the error term to (or almost to) machine precision.

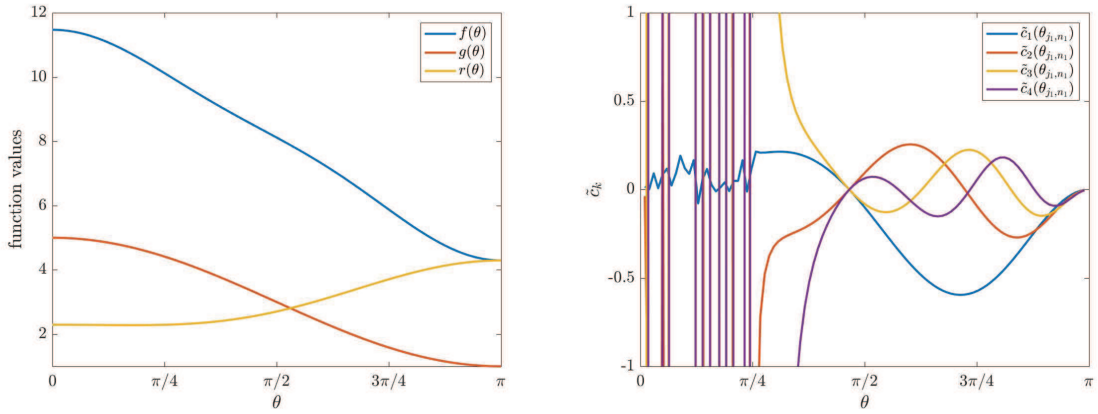


Figure III.5: Example 5: Generating functions (f, g, and r) and $\tilde{c}_k$ for $k = 1, 2, 3, 4$.

In the numerical tests associated with Examples 3 and 4, as in Figures III.8 and III.9, we observe again that the error expansion is in accordance with (III.2). We also note a slight deviation for the largest eigenvalue and this has to be expected since we have $r(\theta_{1,n}) \to \infty$ as $n \to \infty$ for Example 3 (on the other hand for Example 4 we notice $r(\theta_{n,n}) \to \infty$ as $n \to \infty$). However, the approximation of the eigenvalues of $\mathcal{P}_n(f, g)$ is excellent and almost to machine precision as reported in Figure III.12.

In the numerical test related to Example 5 we have a non-monotone region for

$$\theta \in [0, 2\tan^{-1}(\sqrt{3/17})]$$

where the proposed expansion does not work. Indeed additional errors are introduced when compared to $E_{j,n,0}$, since the sampling of $r(\theta_{j_1,n_1})$ leads to a poorer approximation after ordering than the procedure given by sampling $r(\theta_{j,n_7})$ first and then picking samples after ordering. However, the expansion is confirmed for the rest of the domain, as seen in Figure III.10. Furthermore, in Figure III.13 the expansion works well again for the monotone part, by allowing an approximation almost to machine precision of the eigenvalues of $\mathcal{P}_n(f, g)$.

However, even if the eigenvalues lying in the non-monotone region give raise to an irregular error pattern, it seems that there exists a kind of "deformed" periodicity in the error, like it is formally proven, without deformations, for the eigenvalues of $T_n(f)$, $f(\theta) = 2 - 2\cos(\omega\theta)$, $\omega \geq 2$ integer, and $g(\theta) = 1$ (see [63]). The latter observation indicates that a more complete study of this "deformed" periodicity has to be considered in the future.

We finally observe that the remarkable numerical results for the eigenvalues of $\mathcal{P}_n(f, g)$, as reported in Figures III.11, III.12, III.13, positively answer the question:

Q 1. *"Are the eigenvalues of preconditioned banded symmetric Toeplitz matrices known in almost closed form?".*

In fact, we obtain almost machine precision for the computation of the spectrum of $\mathcal{P}_n(f, g)$, for large n and only working with few really small matrices.

At this point our goal will be to ascertain the existence of an asymptotic eigenvalue expansion for PDE discretization matrices and exploit this expansion (if any) for computing the eigenvalues themselves through fast interpolation–extrapolation procedures.

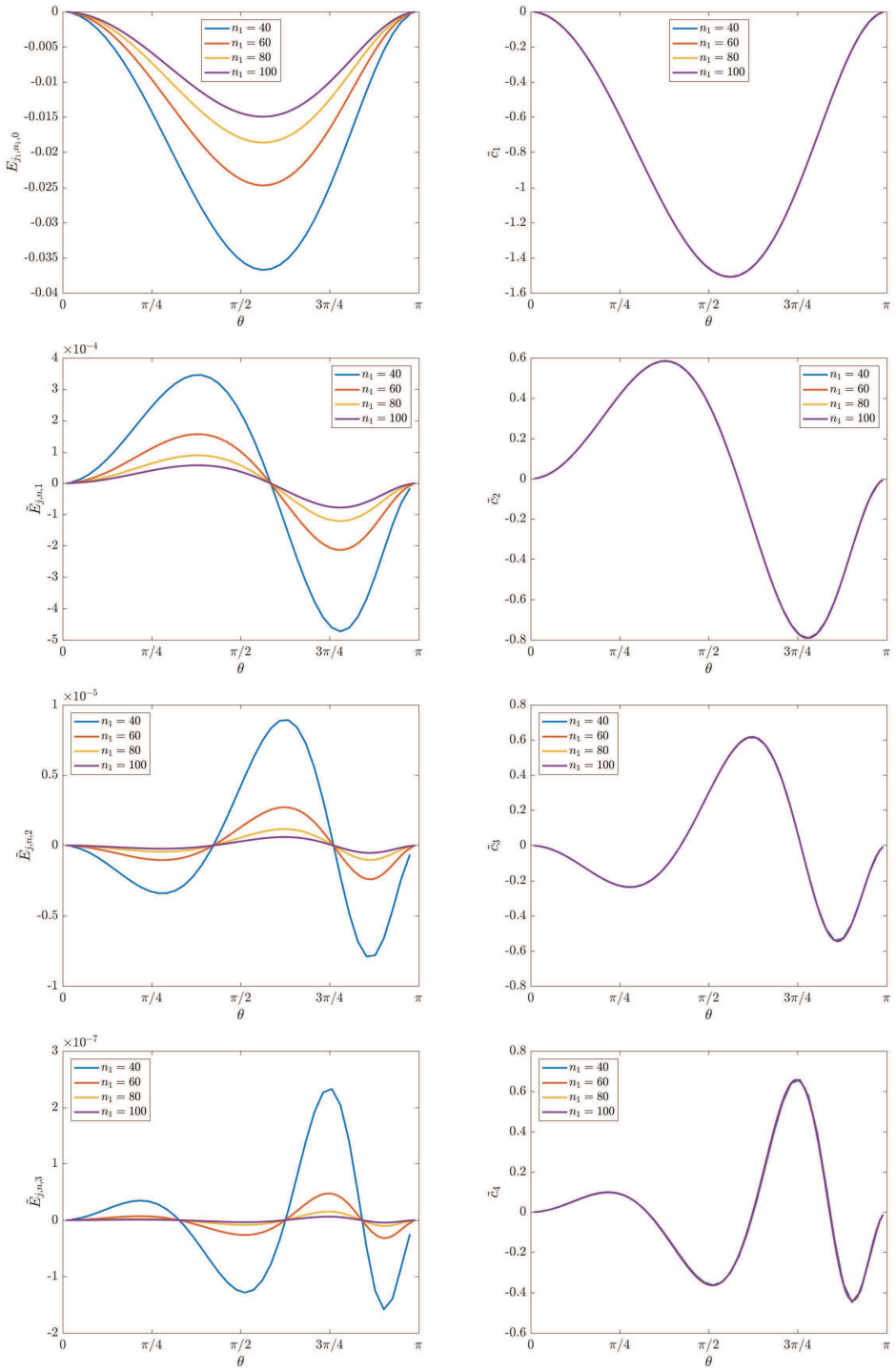


Figure III.6: Example 1: $E_{j,n,0}$, $\tilde{E}_{j,n,k}$ $(k = 1, 2, 3)$, and $\tilde{c}_k$ $(k = 1, 2, 3, 4)$, for $n = n_1 = \{40, 60, 80, 100\}$.

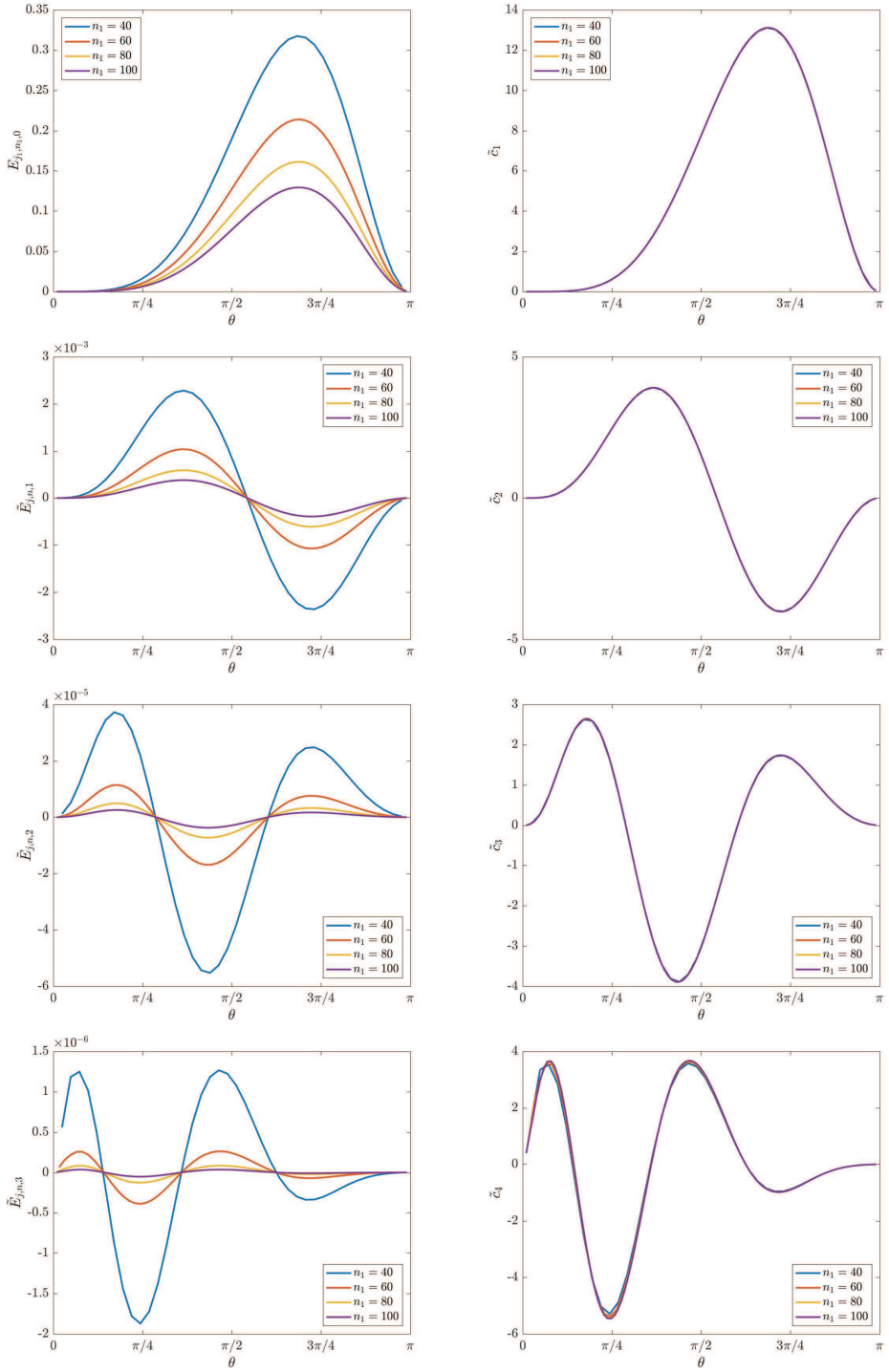


Figure III.7: Example 2: $E_{j,n,0}$, $\tilde{E}_{j,n,k}$ $(k = 1, 2, 3)$, and $\tilde{c}_k$ $(k = 1, 2, 3, 4)$, for $n = n_1 = \{40, 60, 80, 100\}$.

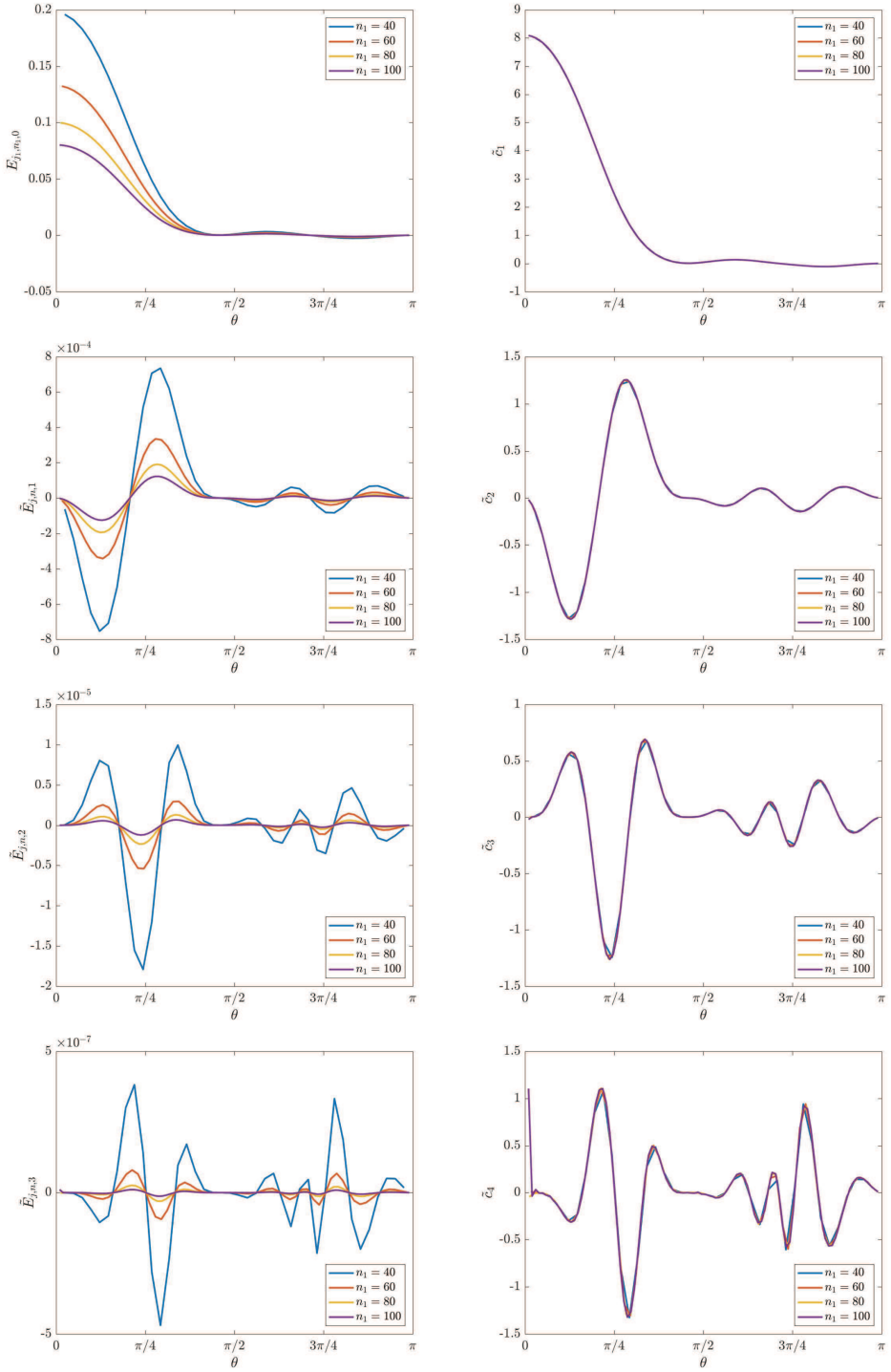


Figure III.8: Example 3: $E_{j,n,0}$, $\tilde{E}_{j,n,k}$ $(k = 1, 2, 3)$, and $\tilde{c}_k$ $(k = 1, 2, 3, 4)$, for $n = n_1 = \{40, 60, 80, 100\}$.

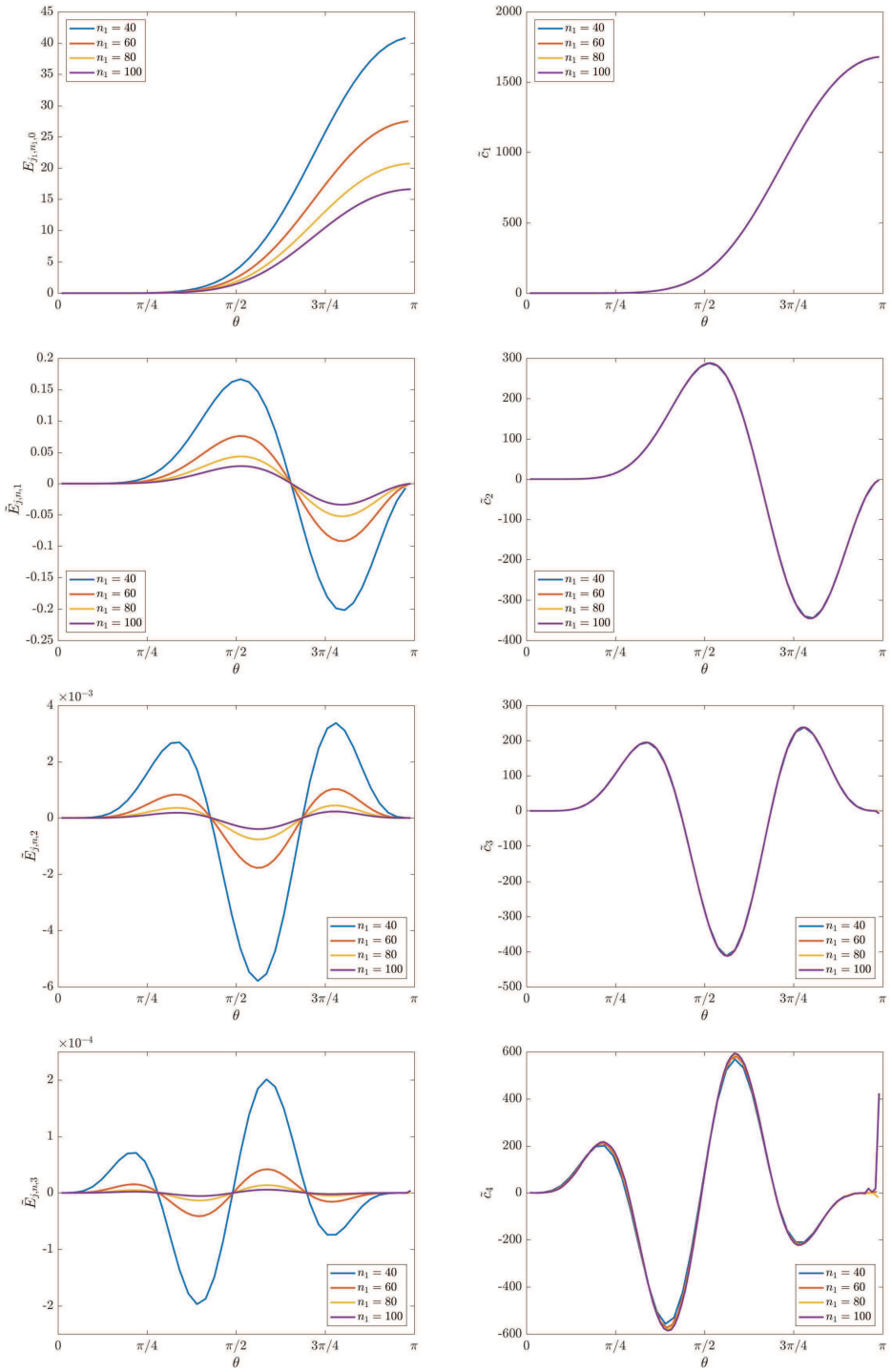


Figure III.9: Example 4: $E_{j,n,0}$, $\tilde{E}_{j,n,k}$ $(k = 1, 2, 3)$, and $\tilde{c}_k$ $(k = 1, 2, 3, 4)$, for $n = n_1 = \{40, 60, 80, 100\}$.

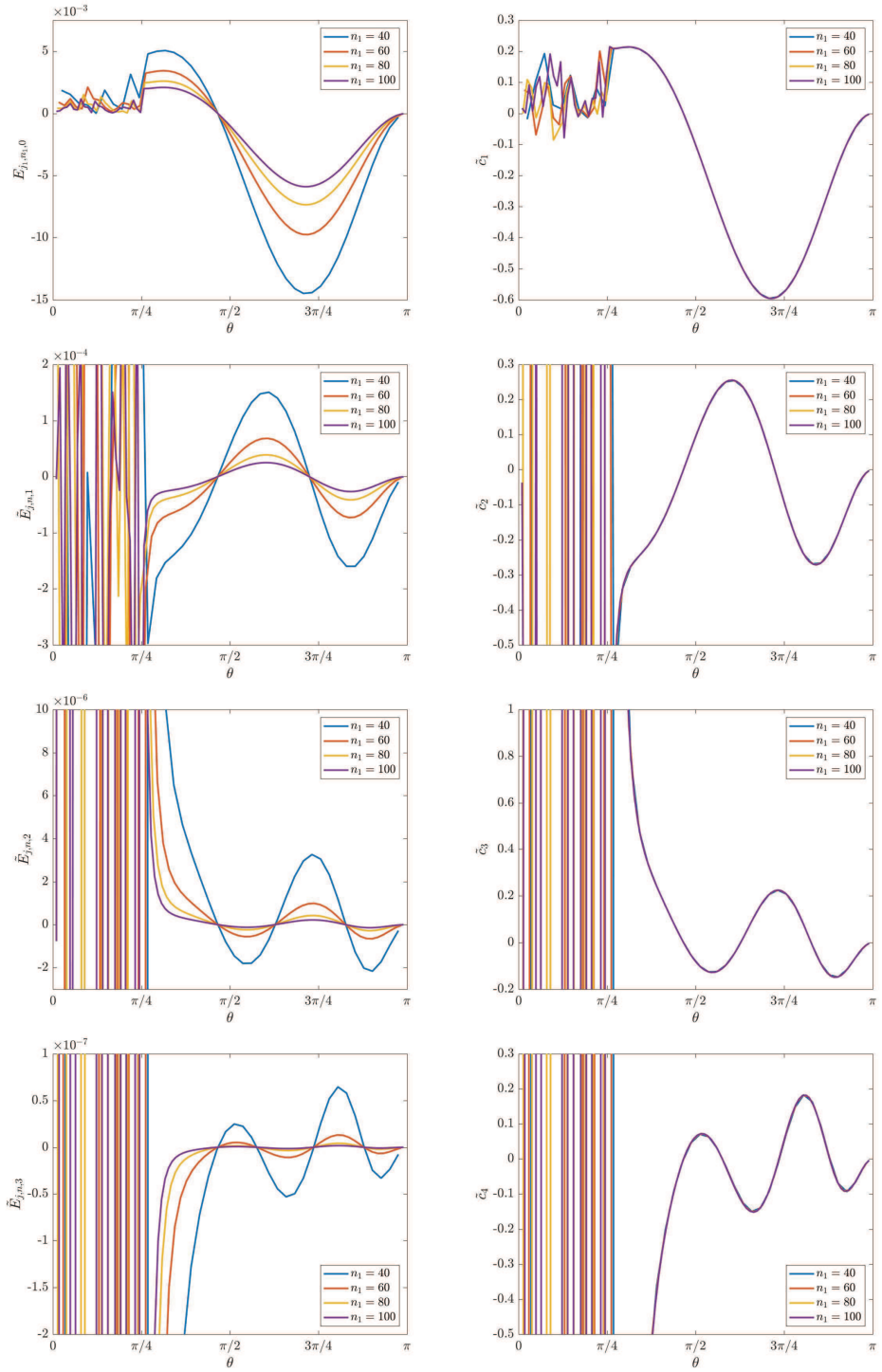


Figure III.10: Example 5: $E_{j,n,0}$, $\tilde{E}_{j,n,k}$ $(k = 1, 2, 3)$, and $\tilde{c}_k$ $(k = 1, 2, 3, 4)$, for $n = n_1 = \{40, 60, 80, 100\}$.

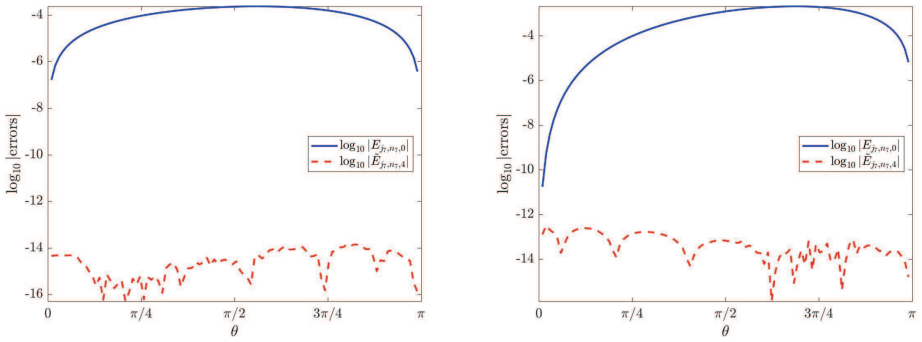


Figure III.11: Example 1 and 2: The errors $\log_{10}|E_{j_7,n_7,0}|$ and $\log_{10}|\tilde{E}_{j_7,n_7,4}|$ for the 100 indices j_7 of $n_7 = 6463$ in (III.19), corresponding to $n_1 = 100$, and using $\tilde{c}_k, k = 1, 2, 3, 4$, computed with $\gamma = 2$.

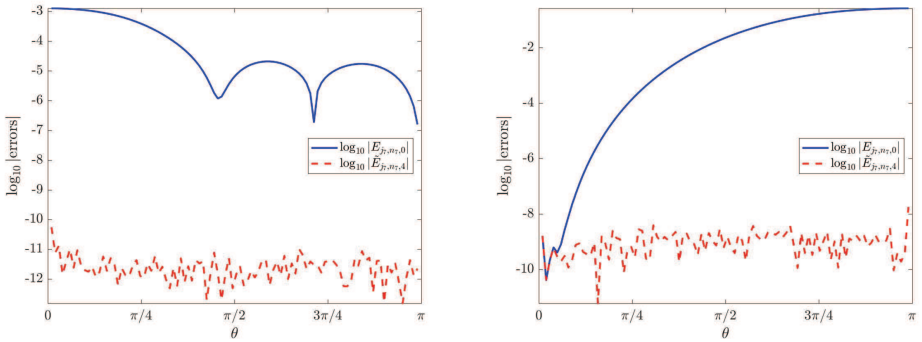


Figure III.12: Example 3 and 4: The errors $\log_{10}|E_{j_7,n_7,0}|$ and $\log_{10}|\tilde{E}_{j_7,n_7,4}|$ for the 100 indices j_7 of $n_7 = 6463$ in (III.19), corresponding to $n_1 = 100$, and using $\tilde{c}_k, k = 1, 2, 3, 4$, computed with $\gamma = 2$.

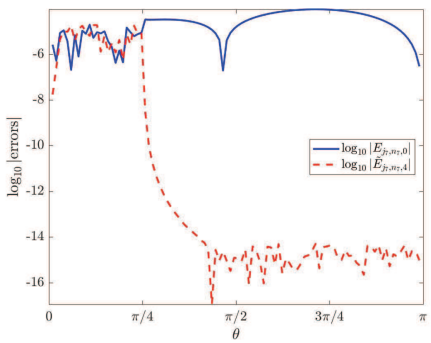


Figure III.13: Example 5: The errors $\log_{10}|E_{j_7,n_7,0}|$ and $\log_{10}|\tilde{E}_{j_7,n_7,4}|$ for the 100 indices j_7 of $n_7 = 6463$ in (III.19), corresponding to $n_1 = 100$, and using $\tilde{c}_k, k = 1, 2, 3, 4$, computed with $\gamma = 2$. Note the non-monotone part, $\theta \in [0, 2\tan^{-1}(\sqrt{3/17})]$, where the error is not improved.

In the next chapter we provide a positive answer in the case where the PDE is a the Laplacian eigenproblem and the discretization method is the B-spline IgA. We observe that the question **Q 1.** can have interesting consequences since it opens the doors to a series of possible future researches.

Chapter IV

Asymptotic Expansion: applied to the IgA discretization

In the present chapter, motivated by the aforesaid interest, we perform a detailed spectral analysis of the matrices stemming from the B-spline Isogeometric Analysis (IgA) discretization of the Laplacian eigenproblem $-\Delta u = \lambda u$.

IgA is a modern paradigm for analyzing problems governed by Partial Differential Equations (PDEs); see [41]. Because of its capability to enhance the connection between numerical simulation and Computer-Aided Design (CAD) systems, IgA is gaining more and more attention over time. In particular, the spectral investigation of matrices arising from the IgA discretization of PDEs has become a topic of interest in the scientific community, mainly because of the superiority of IgA over the classical Finite Element Analysis (FEA) in approximating the spectrum of the underlying differential operator; see, e.g., [42, 80, 90, 92, 103]. It is also worth recalling that recent spectral distribution results for IgA discretization matrices [51, 71, 72, 73, 74, 76, 77] turned out to be the keystone for designing fast IgA solvers [49, 50, 52].

Our main results, which will be detailed in Subsection IV.1, complement those of [51, 71, 72, 73, 74, 76, 77] and deliver a fast (parallel) interpolation–extrapolation algorithm for computing the eigenvalues of the considered IgA matrices.

IV.1 Problem setting

Consider the one-dimensional Laplacian eigenproblem with homogeneous Dirichlet boundary conditions

$$\begin{cases} -u''(x) = \lambda u(x), & x \in (0,1), \\ u(0) = u(1) = 0. \end{cases} \tag{IV.1}$$

The corresponding weak formulation reads as follows: find eigenvalues $\lambda \in \mathbb{R}^+$ and eigenfunctions $u \in H_0^1(0,1)$ such that, for all $v \in H_0^1(0,1)$,

$$\mathrm{a}(u,v) = \lambda(u,v),$$

where

$$\mathrm{a}(u,v) = \int_0^1 u'(x)v'(x)\mathrm{d}x, \qquad (u,v) = \int_0^1 u(x)v(x)\mathrm{d}x.$$

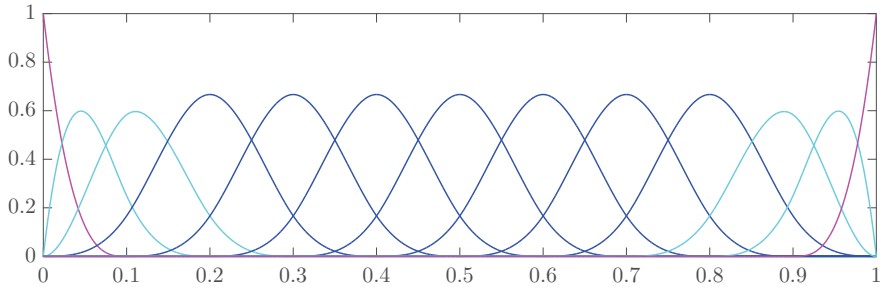


Figure IV.1: Cubic B-splines $\{N_{1,[3]}, \ldots, N_{n+3,[3]}\}$ for the knot sequence $\{0,0,0,0,\frac{1}{n},\frac{2}{n},\ldots,\frac{n-1}{n},1,1,1,1\}$ ($n = 10$).

In Galerkin's method, we choose a finite-dimensional vector space $\mathscr{W} \subset H_0^1(0,1)$, we set $N = \dim \mathscr{W}$, and we look for approximations of the exact eigenpairs

$$\lambda_j = j^2\pi^2, \qquad u_j(x) = \sin(j\pi x), \qquad j \geq 1, \tag{IV.2}$$

by solving the following Galerkin problem: find $\lambda_{j,\mathscr{W}} \in \mathbb{R}^+$ and $u_{j,\mathscr{W}} \in \mathscr{W}$, for $j = 1, \ldots, N$, such that, for all $v \in \mathscr{W}$,

$$\mathrm{a}(u_{j,\mathscr{W}}, v) = \lambda_{j,\mathscr{W}}(u_{j,\mathscr{W}}, v). \tag{IV.3}$$

Assuming the numerical eigenvalues $\lambda_{j,\mathscr{W}}$ are arranged in non decreasing order, the pair $(\lambda_{j,\mathscr{W}}, u_{j,\mathscr{W}})$ is taken as an approximation of the pair

$$(\lambda_j, u_j)$$

for all $j = 1, \ldots, N$. The numbers $\lambda_{j,\mathscr{W}}/\lambda_j - 1$, $j = 1, \ldots, N$, are referred to as the (relative) eigenvalue errors. If $\{\varphi_1, \ldots, \varphi_N\}$ is a basis of $\mathscr{W}$, in view of the canonical identification between each $v \in \mathscr{W}$ and its coefficient vector with respect to $\{\varphi_1, \ldots, \varphi_N\}$, solving the Galerkin problem (IV.3) is equivalent to solving the generalized eigenvalue problem

$$K\mathbf{u}_{j,\mathscr{W}} = \lambda_{j,\mathscr{W}} M\mathbf{u}_{j,\mathscr{W}}, \tag{IV.4}$$

where $\mathbf{u}_{j,\mathscr{W}}$ is the coefficient vector of $u_{j,\mathscr{W}}$ with respect to $\{\varphi_1, \ldots, \varphi_N\}$ and

$$K = [\mathrm{a}(\varphi_j, \varphi_i)]_{i,j=1}^N = \left[\int_0^1 \varphi_j'(x)\varphi_i'(x)\mathrm{d}x\right]_{i,j=1}^N, \tag{IV.5}$$

$$M = [(\varphi_j, \varphi_i)]_{i,j=1}^N = \left[\int_0^1 \varphi_j(x)\varphi_i(x)\mathrm{d}x\right]_{i,j=1}^N. \tag{IV.6}$$

The matrices K and M are referred to as the stiffness matrix and the mass matrix, respectively. Both K and M are always symmetric positive definite, regardless of the chosen basis functions $\varphi_1, \ldots, \varphi_N$. Moreover, it is clear from (IV.4) that the numerical eigenvalues $\lambda_{j,\mathscr{W}}$, $j = 1, \ldots, N$, are just the eigenvalues of the matrix

$$L = M^{-1}K. \tag{IV.7}$$

Now, for $p, n \geq 1$ let

$$N_{i,[p]}, \qquad i = 1, \ldots, n+p, \tag{IV.8}$$

be the B-splines of degree $p \geq 1$ and smoothness $C^{p-1}(\mathbb{R})$ defined over the knot sequence

$$\underbrace{0,\ldots,0}_{p+1},\frac{1}{n},\frac{2}{n},\ldots,\frac{n-1}{n},\underbrace{1,\ldots,1}_{p+1}.$$

The B-splines (IV.8) form a basis for the spline space

$$\mathscr{V}_{n,[p]} = \big\{v \in C^{p-1}[0,1] : \ v_{|[\frac{i}{n},\frac{i+1}{n})} \in \mathbb{P}_p \text{ for } i = 0,\ldots,n-1\big\},$$

where $\mathbb{P}_p$ is the space of polynomials of degree at most p. Except for the first and the last one, all the other B-splines vanish on the boundary of $[0,1]$. In particular, the B-splines

$$N_{i+1,[p]}, \qquad i = 1,\ldots,n+p-2, \tag{IV.9}$$

form a basis for the space

$$\mathscr{W}_{n,[p]} = \big\{v \in \mathscr{V}_{n,[p]} : \ v(0) = v(1) = 0\big\}.$$

We refer the reader to Figure IV.1 for the graphs of the B-splines (IV.8) corresponding to the degree $p = 3$. For more on B-splines, including the precise definition of the functions (IV.8), see [45, 112].

In the IgA approximation of (IV.1) based on uniform B-splines of degree $p \geq 1$, we look for approximations of the exact eigenpairs (IV.2) by using the Galerkin method described above, in which the basis functions $\varphi_1,\ldots,\varphi_N$ are chosen as the B-splines $N_{2,[p]},\ldots,N_{n+p-1,[p]}$ and, consequently, the vector space $\mathscr{W}$ is equal to $\mathscr{W}_{n,[p]}$. The resulting stiffness and mass matrices (IV.5)–(IV.6) are given by

$$K_n^{[p]} = \left[\int_0^1 N'_{j+1,[p]}(x)N'_{i+1,[p]}(x)\mathrm{d}x\right]_{i,j=1}^{n+p-2}, \tag{IV.10}$$

$$M_n^{[p]} = \left[\int_0^1 N_{j+1,[p]}(x)N_{i+1,[p]}(x)\mathrm{d}x\right]_{i,j=1}^{n+p-2}, \tag{IV.11}$$

and the numerical eigenvalues $\lambda_{j,n}^{[p]}$, $j = 1,\ldots,n+p-2$, are the eigenvalues of the matrix

$$L_n^{[p]} = (M_n^{[p]})^{-1}K_n^{[p]}. \tag{IV.12}$$

For more details on IgA, we refer the reader to [41].

Let ϕ_q be the B-spline of degree $q \geq 0$ corresponding to the knot sequence $\{0,1,\ldots,q+1\}$. The function ϕ_q is usually referred to as the cardinal B-spline of degree q and it is recursively defined as follows [45]:

$$\begin{aligned}
&\phi_0(t) = \chi_{[0,1)}(t), \qquad t \in \mathbb{R},\\
&\phi_q(t) = \frac{t}{q}\phi_{q-1}(t) + \frac{q+1-t}{q}\phi_{q-1}(t-1), \qquad t \in \mathbb{R}, \qquad q \geq 1,
\end{aligned}$$

where $\chi_{[0,1)}$ is the characteristic (indicator) function of the interval $[0,1)$. Let

$$f_p : [0,\pi] \to \mathbb{R}, \quad f_p(\theta) = -\phi''_{2p+1}(p+1) - 2\sum_{k=1}^{p} \phi''_{2p+1}(p+1-k)\cos(k\theta), \quad p \geq 1, \tag{IV.13}$$

$$g_p : [0,\pi] \to \mathbb{R}, \quad g_p(\theta) = \phi_{2p+1}(p+1) + 2\sum_{k=1}^{p} \phi_{2p+1}(p+1-k)\cos(k\theta), \quad p \geq 0, \tag{IV.14}$$

$$e_p : [0,\pi] \to \mathbb{R}, \quad e_p(\theta) = \frac{f_p(\theta)}{g_p(\theta)}, \quad p \geq 1. \tag{IV.15}$$

It was proved in [72, Section 3] that [1]

$$f_p(\theta) = (2 - 2\cos(\theta))g_{p-1}(\theta), \qquad \theta \in [0,\pi], \qquad p \geq 1, \tag{IV.16}$$

$$\left(\frac{4}{\pi^2}\right)^{p+1} \leq g_p(\theta) \leq g_p(0) = 1, \qquad \theta \in [0,\pi], \qquad p \geq 0, \tag{IV.17}$$

so in particular the function $e_p(\theta)$ is well-defined. From the analysis in [77, Section 10.7], we know that the three sequences of matrices $\{n^{-1}K_n^{[p]}\}_n$, $\{nM_n^{[p]}\}_n$, $\{n^{-2}L_n^{[p]}\}_n$ have an asymptotic spectral distribution (in the Weyl sense) described by the functions $f_p(\theta)$, $g_p(\theta)$, $e_p(\theta)$, respectively; that is, for any sufficiently large n, up to a small number of outliers, the eigenvalues of $n^{-1}K_n^{[p]}$ (resp., $nM_n^{[p]}$, $n^{-2}L_n^{[p]}$) are approximately given by the samples of $f_p(\theta)$ (resp., $g_p(\theta)$, $e_p(\theta)$) over some uniform grid in $[0,\pi]$. This is illustrated in Figure IV.2 for the matrix $n^{-2}L_n^{[p]}$ and for $p = 1,\dots,6$. Following the terminology in [77, Section 3.1], we refer to $f_p(\theta)$, $g_p(\theta)$, $e_p(\theta)$ as the spectral symbols of $\{n^{-1}K_n^{[p]}\}_n$, $\{nM_n^{[p]}\}_n$, $\{n^{-2}L_n^{[p]}\}_n$, respectively. For more details on the spectral distribution of a sequence of matrices, see [77].

Main contributions

The main contributions of this Chapter can be summarized as follows. Throughout this chapter, we will use the notations $n_p^{\text{out}} = p - 2 + \text{mod}(p,2)$ and $N^{n,p} = n + p - 2$.

1. We prove several important analytic properties of the spectral symbol $e_p(\theta)$. In particular, we show that $e_p(\theta)$ is monotone increasing on $[0,\pi]$ for all $p \geq 1$ and that $e_p(\theta) \to \theta^2$ uniformly on $[0,\pi]$ as $p \to \infty$. Incidentally, we also show that the ratio $w_p(\theta) = g_p(\theta)/g_{p-1}(\theta)$ satisfies $1/3 \leq w_p(\theta) \leq 1$ for all $p \geq 1$ and $\theta \in [0,\pi]$. The latter result was already conjectured in [50, 52] on the basis of numerical experiments, and it was therein exploited to design/analyze fast solvers for IgA discretization matrices.

2. For $p = 1$ and $p = 2$, we compute eigenvalues and eigenvectors of $L_n^{[p]}$. In both cases, the eigenvalues are given by $e_p(\theta_{j,n})$ for $j = 1,\dots,n+p-2$, where $\theta_{j,n} = j\pi/n$. The exact computation of eigenvalues and eigenvectors is made possible by the fact that the matrices $K_n^{[p]}$, $M_n^{[p]}$, $L_n^{[p]}$ belong to the same matrix algebra, which is the tau algebra $\tau_{n-1}(0,0)$ for $p = 1$ and the algebra $\tau_n(-1,-1)$ for $p = 2$ (we are using the notations of [21]). It is worth noting that both these algebras are related to fast unitary sine transforms [21], which implies that many numerical linear algebra computations involving the matrices $K_n^{[p]}$, $M_n^{[p]}$, $L_n^{[p]}$ (matrix-vector products, solutions of linear systems, inversions, etc.) are stable and fast [101, 102].

[1] Note that in [72] the function $g_p(\theta)$ is denoted by $h_p(\theta)$.

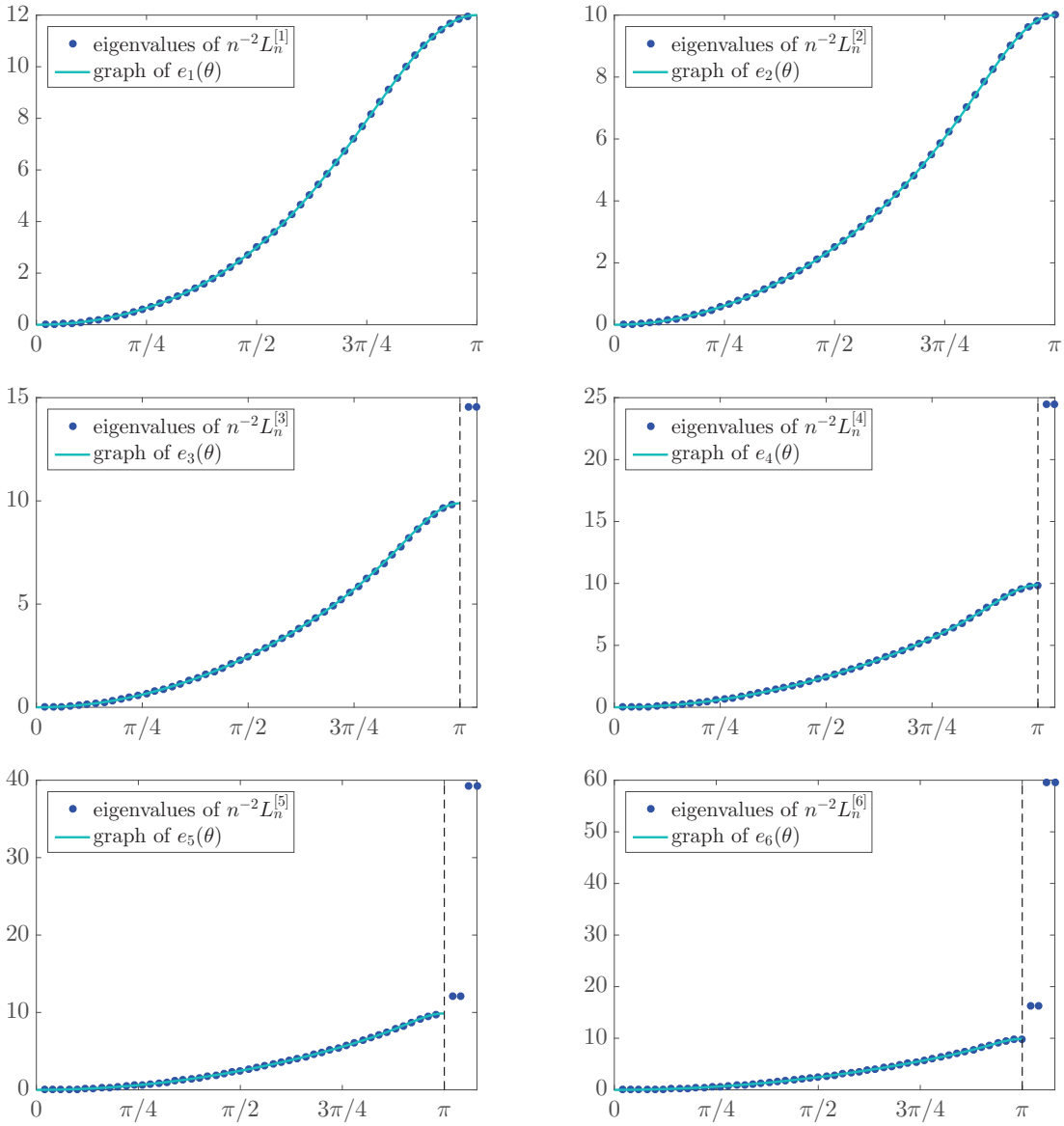


Figure IV.2: Comparison between the eigenvalues of $n^{-2}L_n^{[p]}$ and the graph of $e_p(\theta)$ for $n = 50$ and $p = 1, \ldots, 6$. The eigenvalues of $n^{-2}L_n^{[p]}$ are sorted in non decreasing order and are represented by the thick dots placed at the points $(\theta_{j,n}, \lambda_j(n^{-2}L_n^{[p]}))$, $j = 1, \ldots, n - \mathrm{mod}(p, 2)$, where $\theta_{j,n} = j\pi/n$. The eigenvalues $\lambda_j(n^{-2}L_n^{[p]})$ for $j > n - \mathrm{mod}(p, 2)$ are the so-called outliers and are positioned outside the domain $[0, \pi]$.

3. For $p \geq 3$, we provide numerical evidence of a precise asymptotic expansion for the eigenvalues of $n^{-2}L_n^{[p]}$. Such an expansion, which obviously begins with the spectral symbol $e_p(\theta)$, is in force for the whole of the spectrum except for the largest n_p^{out} eigenvalues (the so-called outliers; see Figure IV.2). To be more precise, we show through numerical experiments that for every $p \geq 3$, every integer $\alpha \geq 0$, every n, and every $j = 1, \ldots, N^{n,p} - n_p^{\text{out}} = n - \text{mod}(p, 2)$, we have

$$\lambda_j(n^{-2}L_n^{[p]}) = e_p(\theta_{j,n}) + \sum_{k=1}^{\alpha} c_k^{[p]}(\theta_{j,n})h^k + E_{j,n,\alpha}^{[p]}, \tag{IV.18}$$

where:

- the eigenvalues of $n^{-2}L_n^{[p]}$ are arranged in non decreasing order, $\lambda_1(n^{-2}L_n^{[p]}) \leq \ldots \leq \lambda_{n+p-2}(n^{-2}L_n^{[p]})$;
- $\{c_k^{[p]}\}_{k=1,2,\ldots}$ is a sequence of functions from $[0, \pi]$ to $\mathbb{R}$ which depends only on p;
- $h = \frac{1}{n}$ and $\theta_{j,n} = \frac{j\pi}{n} = j\pi h$ for $j = 1, \ldots, n$;
- $E_{j,n,\alpha}^{[p]} = O(h^{\alpha+1})$ is the remainder (the error), which satisfies the inequality $|E_{j,n,\alpha}^{[p]}| \leq C_\alpha^{[p]} h^{\alpha+1}$ for some constant $C_\alpha^{[p]}$ depending only on α and p.

We refer the reader to the **Chapter VI** Section VI.4 for a proof of the expansion (IV.18) for $\alpha = 0$ and $j = 1, \ldots, N^{n,p} - (4p - 2)$, where $4p - 2$ represents an estimate, solely based on interlacing/rank-correction arguments, of the actual number of outliers n_p^{out}. We note that (IV.18) is formally the same as the expansions for the eigenvalues of Toeplitz and preconditioned Toeplitz matrices, which have been conjectured and validated through numerical experiments in [1, 62]. Furthermore, basic eigenvalue expansions (and related extrapolation techniques) have been used in [37, 151] in the context of finite element approximations of differential problems. In the light of these considerations, the expansion (IV.18) is not completely unexpected, because $n^{-2}L_n^{[p]}$ is "almost" a preconditioned Toeplitz matrix as $n^{-2}L_n^{[p]} = (nM_n^{[p]})^{-1}(n^{-1}K_n^{[p]})$ and $nM_n^{[p]}$, $n^{-1}K_n^{[p]}$ are Toeplitz matrices, up to low rank corrections. To be precise,

$$n^{-1}K_n^{[p]} = T_{n+p-2}(f_p) + R_n^{[p]}, \tag{IV.19}$$
$$nM_n^{[p]} = T_{n+p-2}(g_p) + S_n^{[p]}, \tag{IV.20}$$

where f_p, g_p are defined in (IV.13)–(IV.14) and

$$(R_n^{[p]})_{ij} = 0, \quad 2p \leq i \leq n - p - 1 \quad \Longrightarrow \quad \text{rank}(R_n^{[p]}) \leq 4p - 2, \tag{IV.21}$$
$$(S_n^{[p]})_{ij} = 0, \quad 2p \leq i \leq n - p - 1 \quad \Longrightarrow \quad \text{rank}(S_n^{[p]}) \leq 4p - 2; \tag{IV.22}$$

see [72, Subsection 4.1].

4. We show through numerical experiments that, for $p \geq 3$ and $k \geq 1$, there exists a point $\theta(p, k) \in (0, \pi)$ such that $c_k^{[p]}(\theta)$ vanishes over $[0, \theta(p, k)]$. Moreover, as it is suggested by the numerics of this chapter, it is very likely that $y_p = \inf_{k \geq 1} \theta(p, k) > 0$ for all $p \geq 3$. This is consistent with another crucial numerical observation, namely the fact that, for all

$p \geq 3$, the equation $\lambda_j(n^{-2}L_n^{[p]}) = e_p(\theta_{j,n})$ holds numerically whenever $\theta_{j,n} < \theta(p)$, with $\theta(p)$ being a point in $(0, y_p]$. In addition, $\theta(p)$ apparently grows with p, i.e., the portion of the spectrum of $\lambda_j(n^{-2}L_n^{[p]})$ which is exactly described by $e_p(\theta)$, at least from a numerical viewpoint, increases with p.

5. For $p \geq 3$, based on the expansion (IV.18) and drawing inspiration from [58], we propose a parallel interpolation–extrapolation algorithm for computing the eigenvalues of $L_n^{[p]}$, excluding the n_p^{out} outliers. The performance of the algorithm is illustrated through numerical experiments. Note that we actually need to compute only the eigenvalues of $L_n^{[p]}$ corresponding in the expansion (IV.18) to points $\theta_{j,n} \geq \theta(p)$, because whenever $\theta_{j,n} < \theta(p)$ we numerically have $\lambda_j(L_n^{[p]}) = n^2 e_p(\theta_{j,n})$ by the previous Item 4.

6. We present a detailed extension of the whole analysis to the general k-dimensional setting, in which problem (IV.1) is replaced by (IV.32). By using tensor-product arguments, we show that the eigenvalue–eigenvector structure of the matrix arising from the IgA approximation of (IV.32) is completely determined by the eigenvalue–eigenvector structure of the matrix $L_n^{[p]}$. In short, the analysis of $L_n^{[p]}$ is enough to cover also the multidimensional case.

The Chapter is organized as follows. In Section IV.2 we report the properties of $e_p(\theta)$ (and $w_p(\theta)$); for ease of reading, the corresponding technical proofs are deferred to **Chapter VI** Section VI.3. In Section IV.3 we compute eigenvalues and eigenvectors of the matrix $L_n^{[p]}$ for $p = 1$ and $p = 2$. In Section IV.4, assuming the asymptotic eigenvalue expansion (IV.18), we present our parallel interpolation–extrapolation algorithm for computing the eigenvalues of $L_n^{[p]}$ for $p \geq 3$, excluding the n_p^{out} outliers. In Section IV.5 we provide numerical experiments in support of both the asymptotic eigenvalue expansion (IV.18) and the properties described in Item 4 of Subsection IV.1. Moreover, we numerically illustrate the performance of the algorithm presented in Section IV.4. In Section IV.6 we extend the whole analysis carried out in Sections IV.3–IV.5 to the multidimensional setting by showing through appropriate tensor-product arguments that the multidimensional case reduces to the unidimensional case.

IV.2 Properties of the spectral symbol $e_p(\theta)$

The spectral symbol $e_p(\theta)$ enjoys the properties reported in Theorems IV.2.1 and IV.2.2, whose proofs are collected in the **Chapter VI** Section VI.3. We note that the convergence expressed in Theorem IV.2.1 was numerically observed in [80] and represents a starting point for the research program outlined in [75, Remark 15].

Theorem IV.2.1. *The function $e_p(\theta)$ converges uniformly to θ^2 on $[0, \pi]$ as $p \to \infty$.*

Theorem IV.2.2. *The function $e_p(\theta)$ is monotone increasing on $[0, \pi]$ for all $p \geq 1$.*

As a byproduct of the proofs of Theorems IV.2.1 and IV.2.2, we also prove the following result for the function

$$w_p : [0, \pi] \to \mathbb{R}, \qquad w_p(\theta) = \frac{g_p(\theta)}{g_{p-1}(\theta)}, \qquad p \geq 1.$$

Theorem IV.2.3. *For $p \geq 1$ and $\theta \in [0, \pi]$ we have*

$$\frac{1}{3} \leq w_p(\theta) \leq 1. \tag{IV.23}$$

Note that the bounds in (IV.23) are sharp. Indeed, $w_p(0) = 1$ for all $p \geq 1$ and $w_1(\pi) = 1/3$. Theorem IV.2.3 provides theoretical support to the numerically observed p-robustness of the solvers devised in [50, 52] for IgA linear systems; see in particular [50, Section 5.5].

IV.3 Eigenvalues and eigenvectors of $L_n^{[p]}$ for $p = 1$ and $p = 2$

In this section we compute the exact spectral decomposition of the matrix $L_n^{[p]}$ for $p = 1$ and $p = 2$. As a preliminary step, we recall some properties of the matrix algebras $\tau_n(\epsilon, \phi)$ introduced in [21] for $\epsilon, \phi \in \{0, 1, -1\}$. It will turn out that $K_n^{[1]}, M_n^{[1]}, L_n^{[1]}$ belong to $\tau_{n-1}(0, 0)$ and $K_n^{[2]}, M_n^{[2]}, L_n^{[2]}$ belong to $\tau_n(-1, -1)$, and this will be the key for computing eigenvalues and eigenvectors of both $L_n^{[1]}$ and $L_n^{[2]}$.

IV.3.1 The matrix algebras $\tau_m(\epsilon, \phi)$ for $\epsilon, \phi \in \{0, 1, -1\}$

Following [21], for any $m \geq 2$ and any $\epsilon, \phi \in \{0, 1, -1\}$ we define the tridiagonal matrix

$$H_m(\epsilon, \phi) = \begin{bmatrix} \epsilon & 1 & 0 & \cdots & 0 \\ 1 & 0 & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & 0 & 1 \\ 0 & \cdots & 0 & 1 & \phi \end{bmatrix} = T_m(2\cos(\theta)) + \epsilon \mathbf{e}_1 \mathbf{e}_1^T + \phi \mathbf{e}_m \mathbf{e}_m^T,$$

Since $H_m(\epsilon, \phi)$ is real and symmetric, it can be decomposed as

$$H_m(\epsilon, \phi) = Q_m(\epsilon, \phi) D_m(\epsilon, \phi) Q_m(\epsilon, \phi)^T,$$

where $Q_m(\epsilon, \phi)$ is a real unitary matrix and $D_m(\epsilon, \phi)$ is a real diagonal matrix. The matrix algebra generated by $H_m(\epsilon, \phi)$ is denoted by $\tau_m(\epsilon, \phi)$ and is given by

$$\tau_m(\epsilon, \phi) = \{Q_m(\epsilon, \phi) D_m Q_m(\epsilon, \phi)^T : \ D_m \text{ is a diagonal matrix of size } m\}.$$

It turns out that the matrix $Q_m(\epsilon, \phi)$ is a fast trigonometric transform such that the matrix-vector product $Q_m(\epsilon, \phi)\mathbf{v}$ can be computed in $O(m \log m)$ operations. Moreover, the diagonal entries of the matrix $D_m(\epsilon, \phi)$ (i.e., the eigenvalues of $H_m(\epsilon, \phi)$) are equal to the samples of the function $2\cos(\theta)$ over a uniform grid in $[0, \pi]$.

The cases of interest in this chapter are $\epsilon = \phi = 0$ and $\epsilon = \phi = -1$. For $\epsilon = \phi = 0$, the matrix algebra $\tau_m(0, 0)$ is the so-called tau algebra, which was originally introduced in [14]. In this case, the sampling grid is

$$\frac{j\pi}{m+1}, \qquad j = 1, \ldots, m,$$

and we have

$$D_m(0,0) = \operatorname*{diag}_{j=1,\dots,m}\left[2\cos\Big(\frac{j\pi}{m+1}\Big)\right],$$
$$Q_m(0,0) = \sqrt{\frac{2}{m+1}}\left[\sin\Big(\frac{ij\pi}{m+1}\Big)\right]_{i,j=1}^{m}.$$

For $\epsilon = \phi = -1$, the sampling grid is

$$\frac{j\pi}{m}, \qquad j = 1, \dots, m,$$

and we have

$$D_m(-1,-1) = \operatorname*{diag}_{j=1,\dots,m}\left[2\cos\Big(\frac{j\pi}{m}\Big)\right],$$
$$Q_m(-1,-1) = \sqrt{\frac{2}{m}}\left[k_j \sin\Big(\frac{(2i-1)j\pi}{2m}\Big)\right]_{i,j=1}^{m}, \qquad k_j = \begin{cases} 1/\sqrt{2}, & \text{if } j = m, \\ 1, & \text{otherwise.} \end{cases}$$

For more details on the matrix algebras $\tau_m(\epsilon, \phi)$ we refer the reader to [21].

IV.3.2 Eigenvalues and eigenvectors of $L_n^{[p]}$ for $p = 1, 2$

In the case $p = 1$, the stiffness and mass matrices $K_n^{[1]}$ and $M_n^{[1]}$ have size $n-1$ and a direct computation shows that

$$n^{-1}K_n^{[1]} = \begin{bmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 2 \end{bmatrix} = T_{n-1}(f_1) = 2I_{n-1} - H_{n-1}(0,0),$$

$$nM_n^{[1]} = \frac{1}{6}\begin{bmatrix} 4 & 1 & & & \\ 1 & 4 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & 4 & 1 \\ & & & 1 & 4 \end{bmatrix} = T_{n-1}(g_1) = \frac{2}{3}I_{n-1} + \frac{1}{6}H_{n-1}(0,0),$$

where f_1, g_1 are given by (IV.13)–(IV.14) for $p = 1$, i.e.,

$$f_1(\theta) = 2 - 2\cos(\theta),$$
$$g_1(\theta) = \frac{2}{3} + \frac{1}{3}\cos(\theta).$$

It follows that both $K_n^{[1]}$ and $M_n^{[1]}$ belong to the tau algebra $\tau_{n-1}(0,0)$. Moreover, based on the results of Subsection IV.3.1, we have

$$n^{-1}K_n^{[1]} = 2I_{n-1} - H_{n-1}(0,0) = Q_{n-1}(0,0)\left(\operatorname*{diag}_{j=1,\dots,n-1}\left[f_1\Big(\frac{j\pi}{n}\Big)\right]\right)Q_{n-1}(0,0)^T,$$
$$nM_n^{[1]} = \frac{2}{3}I_{n-1} + \frac{1}{6}H_{n-1}(0,0) = Q_{n-1}(0,0)\left(\operatorname*{diag}_{j=1,\dots,n-1}\left[g_1\Big(\frac{j\pi}{n}\Big)\right]\right)Q_{n-1}(0,0)^T.$$

Given the algebra structure of $\tau_{n-1}(0,0)$, we obtain

$$n^{-2}L_n^{[1]} = (nM_n^{[1]})^{-1}(n^{-1}K_n^{[1]}) = Q_{n-1}(0,0)\left(\operatorname*{diag}_{j=1,\dots,n-1}\left[e_1\Big(\frac{j\pi}{n}\Big)\right]\right)Q_{n-1}(0,0)^T,$$

where

$$e_1(\theta) = \frac{f_1(\theta)}{g_1(\theta)} = \frac{6(1-\cos(\theta))}{2+\cos(\theta)},$$

as defined by (IV.15) for $p=1$. In particular, $L_n^{[1]}$ belongs to the tau algebra $\tau_{n-1}(0,0)$ just like $K_n^{[1]}$ and $M_n^{[1]}$, and the eigenvalues and eigenvectors of $L_n^{[1]}$ are given by

$$\begin{aligned} &n^2 e_1\Big(\frac{j\pi}{n}\Big), \qquad j=1,\dots,n-1,\\ &\sqrt{\frac{2}{n}}\left[\sin\Big(\frac{ij\pi}{n}\Big)\right]_{i=1}^{n-1}, \qquad j=1,\dots,n-1. \end{aligned}$$

In the case $p=2$, the stiffness and mass matrices $K_n^{[2]}$ and $M_n^{[2]}$ have size n and a direct computation shows that

$$n^{-1}K_n^{[2]} = \frac{1}{6}\begin{bmatrix} 8 & -1 & -1 & & & & \\ -1 & 6 & -2 & -1 & & & \\ -1 & -2 & 6 & -2 & -1 & & \\ & \ddots & \ddots & \ddots & \ddots & \ddots & \\ & & -1 & -2 & 6 & -2 & -1 \\ & & & -1 & -2 & 6 & -1 \\ & & & & -1 & -1 & 8 \end{bmatrix} = T_n(f_2) + R_n^{[2]},$$

$$nM_n^{[2]} = \frac{1}{120}\begin{bmatrix} 40 & 25 & 1 & & & & \\ 25 & 66 & 26 & 1 & & & \\ 1 & 26 & 66 & 26 & 1 & & \\ & \ddots & \ddots & \ddots & \ddots & \ddots & \\ & & 1 & 26 & 66 & 26 & 1 \\ & & & 1 & 26 & 66 & 25 \\ & & & & 1 & 25 & 40 \end{bmatrix} = T_n(g_2) + S_n^{[2]},$$

where f_2, g_2 are given by (IV.13)–(IV.14) for $p=2$, i.e.,

$$\begin{aligned} f_2(\theta) &= 1 - \frac{2}{3}\cos(\theta) - \frac{1}{3}\cos(2\theta),\\ g_2(\theta) &= \frac{11}{20} + \frac{13}{30}\cos(\theta) + \frac{1}{60}\cos(2\theta), \end{aligned}$$

and $R_n^{[2]}, S_n^{[2]}$ are matrices of rank 4 given by

$$R_n^{[2]} = \frac{1}{6}\begin{bmatrix} 2 & 1 & & & \\ 1 & & & & \\ & & & & \\ & & & & 1 \\ & & & 1 & 2 \end{bmatrix},$$

$$S_n^{[2]} = \frac{1}{120}\begin{bmatrix} -26 & -1 & & & \\ -1 & & & & \\ & & & & \\ & & & & -1 \\ & & & -1 & -26 \end{bmatrix}.$$

We note that both $n^{-1}K_n^{[2]}$ and $nM_n^{[2]}$ are of the form

$$A_n(a,b,c) = T_n(a + 2b\cos(\theta) + 2c\cos(2\theta)) + R_n(b,c), \qquad R_n(b,c) = -\begin{bmatrix} b & c & & & \\ c & & & & \\ & & & & \\ & & & & c \\ & & & c & b \end{bmatrix}. \tag{IV.24}$$

Indeed,

$$n^{-1}K_n^{[2]} = A_n\Big(1, -\frac{1}{3}, -\frac{1}{6}\Big),$$
$$nM_n^{[2]} = A_n\Big(\frac{11}{20}, \frac{13}{60}, \frac{1}{120}\Big).$$

Now, any matrix of the form (IV.24) is a polynomial in $H_n(-1,-1)$, and precisely

$$A_n(a,b,c) = (a-2c)I_n + bH_n(-1,-1) + cH_n(-1,-1)^2.$$

It follows that $A_n(a,b,c)$ belongs to the matrix algebra $\tau_n(-1,-1)$. Moreover, based on the results of Subsection IV.3.1, we have

$$A_n(a,b,c) = Q_n(-1,-1)\left(\operatorname*{diag}_{j=1,\dots,n}\Big[a + 2b\cos\Big(\frac{j\pi}{n}\Big) + 2c\cos\Big(\frac{2j\pi}{n}\Big)\Big]\right)Q_n(-1,-1)^T.$$

In particular, $K_n^{[2]}$ and $M_n^{[2]}$ belong to $\tau_n(-1,-1)$ and

$$n^{-1}K_n^{[2]} = Q_n(-1,-1)\left(\operatorname*{diag}_{j=1,\dots,n}\Big[f_2\Big(\frac{j\pi}{n}\Big)\Big]\right)Q_n(-1,-1)^T,$$
$$nM_n^{[2]} = Q_n(-1,-1)\left(\operatorname*{diag}_{j=1,\dots,n}\Big[g_2\Big(\frac{j\pi}{n}\Big)\Big]\right)Q_n(-1,-1)^T.$$

Given the algebra structure of $\tau_n(-1,-1)$, we obtain

$$n^{-2}L_n^{[2]} = (nM_n^{[2]})^{-1}(n^{-1}K_n^{[2]}) = Q_n(-1,-1)\left(\operatorname*{diag}_{j=1,\dots,n}\Big[e_2\Big(\frac{j\pi}{n}\Big)\Big]\right)Q_n(-1,-1)^T,$$

where

$$e_2(\theta) = \frac{f_2(\theta)}{g_2(\theta)} = \frac{20(3 - 2\cos(\theta) - \cos(2\theta))}{33 + 26\cos(\theta) + \cos(2\theta)},$$

as defined by (IV.15) for $p = 2$. In particular, $L_n^{[2]}$ belongs to the algebra $\tau_n(-1,-1)$ just like $K_n^{[2]}$ and $M_n^{[2]}$, and the eigenvalues and eigenvectors of $L_n^{[2]}$ are given by

$$n^2 e_2\Big(\frac{j\pi}{n}\Big), \qquad j = 1, \dots, n,$$
$$\sqrt{\frac{2}{n}}\left[k_j \sin\Big(\frac{(2i-1)j\pi}{2n}\Big)\right]_{i=1}^{n}, \qquad k_j = \begin{cases} 1/\sqrt{2}, & \text{if } j = n, \\ 1, & \text{otherwise,} \end{cases} \qquad j = 1, \dots, n.$$

Remark 6. *In a recent work [134], Tani proposed a preconditioner based on the fast sine transform $Q_n(-1,-1)$ for solving linear systems arising from the IgA discretization of unidimensional differential problems. For the case $p = 2$, the performance of the preconditioner was extremely good: just one Krylov iteration! The theoretical explanation of such an excellent behavior lies precisely in the exact spectral decompositions obtained in this subsection, where it is shown that $Q_n(-1,-1)$ diagonalizes simultaneously the three matrices $K_n^{[2]}, M_n^{[2]}, L_n^{[2]}$. Note that decompositions of this kind can also be used for accelerating the convergence of recently proposed iterative solvers for IgA linear systems, such as multigrid-based and preconditioned Krylov-based methods; see [50, 52, 111] and the references therein.*

Remark 7. *The results of Subsection IV.3 show that $K_n^{[p]}, M_n^{[p]}, L_n^{[p]}$ belong to the same matrix algebra for $p = 1, 2$. Does this property remains true for $p \geq 3$? The answer is "no". Indeed, if $K_n^{[p]}, M_n^{[p]}, L_n^{[p]}$ belong to the same matrix algebra, then $K_n^{[p]}$ and $M_n^{[p]}$ commute. We numerically verified that $K_n^{[p]}$ and $M_n^{[p]}$ do not commute for $p \geq 3$.*

IV.4 Algorithm for computing the eigenvalues of $L_n^{[p]}$ for $p \geq 3$

Assuming the expansion (IV.18) and drawing inspiration from [58], in this section we propose a parallel interpolation–extrapolation algorithm for computing the eigenvalues of $L_n^{[p]}$, excluding the n_p^{out} outliers. In what follows, for each positive integer $n \in \mathbb{N} = \{1, 2, 3, \dots\}$ and each $p \geq 3$ we define $n^{[p]} = n - \text{mod}(p, 2)$. Moreover, with each positive integer n we associate the stepsize $h = \frac{1}{n}$ and the grid points $\theta_{j,n} = j\pi h$, $j = 1, \dots, n$. For notational convenience, unless otherwise stated, we will always denote a positive integer and the associated stepsize in the same way. For example, if the positive integer is n, the associated stepsize is h; if the positive integer is n_1, the associated stepsize is h_1; if the positive integer is $\bar{n}$, the associated stepsize is $\bar{h}$; etc. Throughout this section, we make the following assumptions.

- $p \geq 3$ and $n, n_1, \alpha \in \mathbb{N}$ are fixed parameters.
- $n_k = 2^{k-1} n_1$ for $k = 1, \dots, \alpha$.
- $j_k = 2^{k-1} j_1$ for $j_1 = 1, \dots, n_1$ and $k = 1, \dots, \alpha$; j_k is the index in $\{1, \dots, n_k\}$ such that $\theta_{j_k,n_k} = \theta_{j_1,n_1}$.

A graphical representation of the grids $\{\theta_{1,n_k}, \dots, \theta_{n_k,n_k}\}$, $k = 1, \dots, \alpha$, is reported in Figure IV.3 for $n_1 = 5$ and $\alpha = 4$.

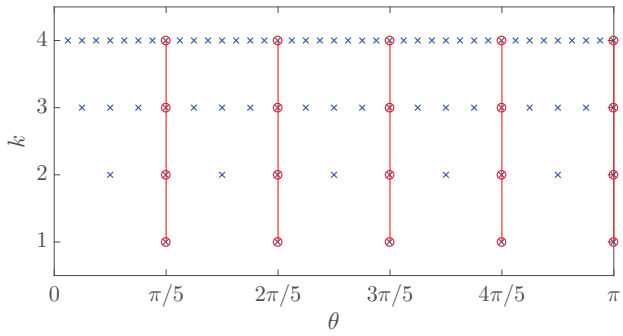


Figure IV.3: Representation of the grids $\theta_{[n_k]}$, $k = 1, \ldots, \alpha$, for $n_1 = 5$ and $\alpha = 4$. Red circles represent the grid points θ_{j_k,n_k} and blue x's represent the rest of the grid points of $\theta_{[n_k]}$.

For each fixed $j_1 = 1, \ldots, n_1^{[p]}$ we apply α times the expansion (IV.18) with $n = n_1, n_2, \ldots, n_\alpha$ and $j = j_1, j_2, \ldots, j_\alpha$. Since $\theta_{j_1,n_1} = \theta_{j_2,n_2} = \ldots = \theta_{j_\alpha,n_\alpha}$ (by definition of $j_2, \ldots, j_\alpha$), we obtain

$$\begin{cases} E^{[p]}_{j_1,n_1,0} = c_1^{[p]}(\theta_{j_1,n_1})h_1 + c_2^{[p]}(\theta_{j_1,n_1})h_1^2 + \ldots + c_\alpha^{[p]}(\theta_{j_1,n_1})h_1^\alpha + E^{[p]}_{j_1,n_1,\alpha} \\ E^{[p]}_{j_2,n_2,0} = c_1^{[p]}(\theta_{j_1,n_1})h_2 + c_2^{[p]}(\theta_{j_1,n_1})h_2^2 + \ldots + c_\alpha^{[p]}(\theta_{j_1,n_1})h_2^\alpha + E^{[p]}_{j_2,n_2,\alpha} \\ \quad \vdots \\ E^{[p]}_{j_\alpha,n_\alpha,0} = c_1^{[p]}(\theta_{j_1,n_1})h_\alpha + c_2^{[p]}(\theta_{j_1,n_1})h_\alpha^2 + \ldots + c_\alpha^{[p]}(\theta_{j_1,n_1})h_\alpha^\alpha + E^{[p]}_{j_\alpha,n_\alpha,\alpha} \end{cases}, \tag{IV.25}$$

where

$$E^{[p]}_{j_k,n_k,0} = \lambda_{j_k}(n_k^{-2}L^{[p]}_{n_k}) - e_p(\theta_{j_1,n_1}), \qquad k = 1, \ldots, \alpha,$$

and

$$|E^{[p]}_{j_k,n_k,\alpha}| \leq C_\alpha^{[p]} h_k^{\alpha+1}, \qquad k = 1, \ldots, \alpha. \tag{IV.26}$$

Let $\tilde{c}_1^{[p]}(\theta_{j_1,n_1}), \ldots, \tilde{c}_\alpha^{[p]}(\theta_{j_1,n_1})$ be the approximations of $c_1^{[p]}(\theta_{j_1,n_1}), \ldots, c_\alpha^{[p]}(\theta_{j_1,n_1})$ obtained by removing all the errors $E^{[p]}_{j_1,n_1,\alpha}, \ldots, E^{[p]}_{j_\alpha,n_\alpha,\alpha}$ in (IV.25) and by solving the resulting linear system:

$$\begin{cases} E^{[p]}_{j_1,n_1,0} = \tilde{c}_1^{[p]}(\theta_{j_1,n_1})h_1 + \tilde{c}_2^{[p]}(\theta_{j_1,n_1})h_1^2 + \ldots + \tilde{c}_\alpha^{[p]}(\theta_{j_1,n_1})h_1^\alpha \\ E^{[p]}_{j_2,n_2,0} = \tilde{c}_1^{[p]}(\theta_{j_1,n_1})h_2 + \tilde{c}_2^{[p]}(\theta_{j_1,n_1})h_2^2 + \ldots + \tilde{c}_\alpha^{[p]}(\theta_{j_1,n_1})h_2^\alpha \\ \quad \vdots \\ E^{[p]}_{j_\alpha,n_\alpha,0} = \tilde{c}_1^{[p]}(\theta_{j_1,n_1})h_\alpha + \tilde{c}_2^{[p]}(\theta_{j_1,n_1})h_\alpha^2 + \ldots + \tilde{c}_\alpha^{[p]}(\theta_{j_1,n_1})h_\alpha^\alpha \end{cases}, \tag{IV.27}$$

Note that this way of computing approximations for $c_1^{[p]}(\theta_{j_1,n_1}), \ldots, c_\alpha^{[p]}(\theta_{j_1,n_1})$ is completely analogous to the Richardson extrapolation procedure that is employed in the context of Romberg integration to accelerate the convergence of the trapezoidal rule [132, Section 3.4]. In this regard, the asymptotic expansion (IV.18) plays here the same role as the Euler–Maclaurin summation formula [132, Section 3.3]. For more advanced studies on extrapolation methods, we refer the reader to Brezinski and Redivo-Zaglia [23]. The next theorem shows that the approximation error $|c_k^{[p]}(\theta_{j_1,n_1}) - \tilde{c}_k^{[p]}(\theta_{j_1,n_1})|$ is $O(h_1^{\alpha-k+1})$.

Theorem IV.4.1. *There exists a constant $A_\alpha^{[p]}$ depending only on α and p such that, for $j_1 = 1, \dots, n_1^{[p]}$ and $k = 1, \dots, \alpha$,*

$$|c_k^{[p]}(\theta_{j_1,n_1}) - \tilde{c}_k^{[p]}(\theta_{j_1,n_1})| \le A_\alpha^{[p]} h_1^{\alpha-k+1}. \tag{IV.28}$$

Proof. It is a straightforward adaptation of the proof of [61, Theorem 1]. □

The improvement of the algorithm is performed by using an *interpolation procedure*. This has been designed following the idea in [61].

We fix an index $j \in \{1, \dots, n^{[p]}\}$. To compute an approximation of $\lambda_j(n^{-2}L_n^{[p]})$ through the expansion (IV.18) we would need the value $c_k^{[p]}(\theta_{j,n})$ for each $k = 1, \dots, \alpha$. Of course, $c_k^{[p]}(\theta_{j,n})$ is not available in practice, but we can approximate it by interpolating in some way the values $\tilde{c}_k^{[p]}(\theta_{j_1,n_1})$, $j_1 = 1, \dots, n_1^{[p]}$. For example, we may define $\tilde{c}_k^{[p]}(\theta)$ as the interpolation polynomial of the data $(\theta_{j_1,n_1}, \tilde{c}_k^{[p]}(\theta_{j_1,n_1}))$, $j_1 = 1, \dots, n_1^{[p]}$, — so that $\tilde{c}_k^{[p]}(\theta)$ is expected to be an approximation of $c_k^{[p]}(\theta)$ over the whole interval $[0, \pi]$ — and take $\tilde{c}_k^{[p]}(\theta_{j,n})$ as an approximation to $c_k^{[p]}(\theta_{j,n})$. It is known, however, that interpolation over a large number of uniform nodes is not advisable as it may give rise to spurious oscillations (Runge's phenomenon). It is therefore better to adopt another kind of approximation. An alternative could be the following: we approximate $c_k^{[p]}(\theta)$ by the spline function $\tilde{c}_k^{[p]}(\theta)$ which is linear on each interval $[\theta_{j_1,n_1}, \theta_{j_1+1,n_1}]$ and takes the value $\tilde{c}_k^{[p]}(\theta_{j_1,n_1})$ at θ_{j_1,n_1} for all $j_1 = 1, \dots, n_1^{[p]}$. This strategy usually removes any spurious oscillation, yet it is not accurate. In particular, it does not preserve the accuracy of approximation at the nodes θ_{j_1,n_1} established in Theorem IV.4.1, i.e., there is no guarantee that $|c_k^{[p]}(\theta) - \tilde{c}_k^{[p]}(\theta)| \le B_\alpha^{[p]} h_1^{\alpha-k+1}$ for $\theta \in [0, \pi]$ or $|c_k^{[p]}(\theta_{j,n}) - \tilde{c}_k^{[p]}(\theta_{j,n})| \le B_\alpha^{[p]} h_1^{\alpha-k+1}$ for $j = 1, \dots, n^{[p]}$, with $B_\alpha^{[p]}$ being a constant depending only on α and p. As proved in Theorem IV.4.2, a local approximation strategy that preserves the accuracy (IV.28), at least if $c_k^{[p]}(\theta)$ is sufficiently smooth, is the following: let $\theta^{(1)}, \dots, \theta^{(\alpha-k+1)}$ be $\alpha - k + 1$ points of the grid $\{\theta_{1,n_1}, \dots, \theta_{n_1^{[p]},n_1}\}$ which are closest to the point $\theta_{j,n}$,[2] and let $\tilde{c}_{k,j}^{[p]}(\theta)$ be the interpolation polynomial of the data $(\theta^{(1)}, \tilde{c}_k^{[p]}(\theta^{(1)})), \dots, (\theta^{(\alpha-k+1)}, \tilde{c}_k^{[p]}(\theta^{(\alpha-k+1)}))$; then, we approximate $c_k^{[p]}(\theta_{j,n})$ by $\tilde{c}_{k,j}^{[p]}(\theta_{j,n})$. Note that, by selecting $\alpha - k + 1$ points from $\{\theta_{1,n_1}, \dots, \theta_{n_1^{[p]},n_1}\}$, we are implicitly assuming that $n_1^{[p]} \ge \alpha - k + 1$.

Theorem IV.4.2. *Let $p \ge 3$ and $1 \le k \le \alpha$, and suppose $n_1^{[p]} \ge \alpha - k + 1$ and $c_k^{[p]} \in C^{\alpha-k+1}[0, \pi]$. For $j = 1, \dots, n^{[p]}$, if $\theta^{(1)}, \dots, \theta^{(\alpha-k+1)}$ are $\alpha - k + 1$ points of $\{\theta_{1,n_1}, \dots, \theta_{n_1^{[p]},n_1}\}$ which are closest to $\theta_{j,n}$, and if $\tilde{c}_{k,j}^{[p]}(\theta)$ is the interpolation polynomial of the data*

$$(\theta^{(1)}, \tilde{c}_k^{[p]}(\theta^{(1)})), \dots, (\theta^{(\alpha-k+1)}, \tilde{c}_k^{[p]}(\theta^{(\alpha-k+1)})),$$

then

$$|c_k^{[p]}(\theta_{j,n}) - \tilde{c}_{k,j}^{[p]}(\theta_{j,n})| \le B_\alpha^{[p]} h_1^{\alpha-k+1} \tag{IV.29}$$

for some constant $B_\alpha^{[p]}$ depending only on α and p.

Proof. It is a straightforward adaptation of the proof of [58, Theorem 2]. □

[2] These $\alpha - k + 1$ points are uniquely determined by $\theta_{j,n}$ except in the following two cases: (a) $\theta_{j,n}$ coincides with a grid point θ_{j_1,n_1} and $\alpha - k + 1$ is even; (b) $\theta_{j,n}$ coincides with the midpoint between two consecutive grid points $\theta_{j_1,n_1}, \theta_{j_1+1,n_1}$ and $\alpha - k + 1$ is odd.

We are now ready to formulate our algorithm for computing the eigenvalues of $L_n^{[p]}$, excluding the outliers. Note that this algorithm permits a parallel implementation, as in the case of [58, Algorithm 1]; see [58, Remark 4].

Algorithm 1. *Given $p \geq 3$ and $n, n_1, \alpha \in \mathbb{N}$ with $n_1^{[p]} \geq \alpha$, we compute approximations of the eigenvalues $\lambda_j(L_n^{[p]})$ for $j = 1, \ldots, n^{[p]}$ as follows.*

1. *For $j_1 = 1, \ldots, n_1^{[p]}$ compute $\tilde{c}_1^{[p]}(\theta_{j_1,n_1}), \ldots, \tilde{c}_\alpha^{[p]}(\theta_{j_1,n_1})$ by solving* (IV.27).
2. *For $j = 1, \ldots, n^{[p]}$*
   - *for $k = 1, \ldots, \alpha$*
     - *determine $\alpha - k + 1$ points $\theta^{(1)}, \ldots, \theta^{(\alpha-k+1)} \in \{\theta_{1,n_1}, \ldots, \theta_{n_1^{[p]},n_1}\}$ which are closest to $\theta_{j,n}$;*
     - *compute $\tilde{c}_{k,j}^{[p]}(\theta_{j,n})$, where $\tilde{c}_{k,j}^{[p]}(\theta)$ is the interpolation polynomial of the data*

$$(\theta^{(1)}, \tilde{c}_k^{[p]}(\theta^{(1)})), \ \ldots, \ (\theta^{(\alpha-k+1)}, \tilde{c}_k^{[p]}(\theta^{(\alpha-k+1)}));$$

   - *compute $\tilde{\lambda}_j(n^{-2}L_n^{[p]}) = e_p(\theta_{j,n}) + \sum_{k=1}^{\alpha} \tilde{c}_{k,j}^{[p]}(\theta_{j,n})h^k$ and $\tilde{\lambda}_j(L_n^{[p]}) = n^2\tilde{\lambda}_j(n^{-2}L_n^{[p]})$.*
3. *Return $(\tilde{\lambda}_1(L_n^{[p]}), \ldots, \tilde{\lambda}_{n^{[p]}}(L_n^{[p]}))$ as an approximation to $(\lambda_1(L_n^{[p]}), \ldots, \lambda_{n^{[p]}}(L_n^{[p]}))$.*

Remark 8. *Algorithm 1 is specifically designed for computing the eigenvalues of $L_n^{[p]}$ in the case where n is quite large. When applying this algorithm, it is implicitly assumed that n_1 and α are small (much smaller than n), so that each $n_k = 2^{k-1}n_1$ is small as well and the computation of the eigenvalues of $L_{n_k}^{[p]}$ — which is required in the first step — can be efficiently performed by any standard eigensolver (e.g., the solver used by the function `eig` of* MATLAB*).*

The last theorem of this section provides an estimate for the approximation error made by Algorithm 1.

Theorem IV.4.3. *Let $p \geq 3$, $n^{[p]} \geq n_1^{[p]} \geq \alpha$ and $c_k^{[p]} \in C^{\alpha-k+1}[0, \pi]$ for $k = 1, \ldots, \alpha$. Let $(\tilde{\lambda}_1(L_n^{[p]}), \ldots, \tilde{\lambda}_{n^{[p]}}(L_n^{[p]}))$ be the approximation of $(\lambda_1(L_n^{[p]}), \ldots, \lambda_{n^{[p]}}(L_n^{[p]}))$ computed by Algorithm 1. Then, there exists a constant $D_\alpha^{[p]}$ depending only on α and p such that, for $j = 1, \ldots, n^{[p]}$,*

$$|\lambda_j(L_n^{[p]}) - \tilde{\lambda}_j(L_n^{[p]})| \leq D_\alpha^{[p]} n h_1^\alpha. \tag{IV.30}$$

Proof. By (IV.18) and Theorem IV.4.2,

$$\begin{aligned}
|\lambda_j(n^{-2}L_n^{[p]}) - \tilde{\lambda}_j(n^{-2}L_n^{[p]})| &= \left| e_p(\theta_{j,n}) + \sum_{k=1}^{\alpha} c_k^{[p]}(\theta_{j,n})h^k + E_{j,n,\alpha}^{[p]} - e_p(\theta_{j,n}) - \sum_{k=1}^{\alpha} \tilde{c}_{k,j}^{[p]}(\theta_{j,n})h^k \right| \\
&\leq \sum_{k=1}^{\alpha} |c_k^{[p]}(\theta_{j,n}) - \tilde{c}_{k,j}^{[p]}(\theta_{j,n})|h^k + |E_{j,n,\alpha}^{[p]}| \\
&\leq B_\alpha^{[p]} \sum_{k=1}^{\alpha} h_1^{\alpha-k+1}h^k + C_\alpha^{[p]}h^{\alpha+1} \leq D_\alpha^{[p]} h_1^\alpha h,
\end{aligned}$$

where $D_\alpha^{[p]} = (\alpha + 1)\max(B_\alpha^{[p]}, C_\alpha^{[p]})$. Multiplying both sides by n^2 we get the thesis. □

n_1	200	300	400	500	600
$\theta_{n_1}^{(\varepsilon)}(3,1)$	$\frac{86\pi}{200} \approx 1.3509$	$\frac{129\pi}{300} \approx 1.3509$	$\frac{172\pi}{400} \approx 1.3509$	$\frac{214\pi}{500} \approx 1.3446$	$\frac{257\pi}{600} \approx 1.3456$
$\theta_{n_1}^{(\varepsilon)}(3,2)$	$\frac{115\pi}{200} \approx 1.8064$	$\frac{172\pi}{300} \approx 1.8012$	$\frac{229\pi}{400} \approx 1.7986$	$\frac{286\pi}{500} \approx 1.7970$	$\frac{343\pi}{600} \approx 1.7959$
$\theta_{n_1}^{(\varepsilon)}(3,3)$	$\frac{126\pi}{200} \approx 1.9792$	$\frac{188\pi}{300} \approx 1.9687$	$\frac{251\pi}{400} \approx 1.9713$	$\frac{313\pi}{500} \approx 1.9666$	$\frac{377\pi}{600} \approx 1.9740$

Table IV.1: Example 6, $p = 3$: values $\theta_{n_1}^{(\varepsilon)}(3,k)$ for $k = 1,2,3$ and $n_1 = 200, 300, 400, 500, 600$, computed with the threshold $\varepsilon = 0.0005$.

Note that the error estimate provided by Theorem IV.4.3 seems disappointing, because of the presence of the large factor n in the right-hand side of (IV.30). However, one should take into account that (IV.30) is an absolute error estimate which, moreover, is uniform in j. Considering that the largest non-outlier eigenvalue of $L_n^{[p]}$, namely $\lambda_{n^{[p]}}(L_n^{[p]})$, diverges to ∞ with the same asymptotic speed as n^2, from (IV.30) we obtain the approximate inequality

$$\frac{|\lambda_{n^{[p]}}(L_n^{[p]}) - \tilde{\lambda}_{n^{[p]}}(L_n^{[p]})|}{|\lambda_{n^{[p]}}(L_n^{[p]})|} \leq D_\alpha^{[p]} h_1^\alpha h,$$

which is a good relative error estimate. We refer the reader to Subsection IV.5.2 for several numerical illustrations of the actual performance of Algorithm 1.

IV.5 Numerical experiments

This section is composed of two subsections. In Subsection IV.5.1 we implement the program described in Items 3 and 4 of Subsection IV.1. In other words, we validate through numerical experiments the expansion (IV.18) for $p \geq 3$; we numerically show, for $p \geq 3$ and $k \geq 1$, the existence of a point $\theta(p,k) \in (0,\pi)$ such that $c_k^{[p]}(\theta)$ vanishes over $[0, \theta(p,k)]$; and we provide numerical evidence of the fact that the infimum $y_p = \inf_{k\geq 1} \theta(p,k)$ is strictly positive and the equation $\lambda_j(n^{-2}L_n^{[p]}) = e_p(\theta_{j,n})$ holds numerically whenever $\theta_{j,n} < \theta(p)$, with $\theta(p)$ being a point in $(0, y_p]$. In Subsection IV.5.2 we illustrate the numerical performance of Algorithm 1.

IV.5.1 Numerical experiments in support of the eigenvalue expansion

Fix $p \geq 3$ and $\alpha \in \mathbb{N}$. As in Section IV.4, for every $n_1 \in \mathbb{N}$ we set

$$\begin{aligned} n_k &= 2^{k-1} n_1, \qquad k = 1, \dots, \alpha, \\ j_k &= 2^{k-1} j_1, \qquad k = 1, \dots, \alpha, \qquad j_1 = 1, \dots, n_1. \end{aligned}$$

In the hypothesis that the expansion (IV.18) holds, we can follow the derivation of Section IV.4 until Theorem IV.4.1 and we conclude that, for each $k = 1, \dots, \alpha$ and $j_1 = 1, \dots, n_1^{[p]}$, the value $\tilde{c}_k^{[p]}(\theta_{j_1,n_1})$ computed by solving the linear system (IV.27) converges to the value $c_k^{[p]}(\theta_{j_1,n_1})$ as $n_1 \to \infty$ with the same asymptotic speed as $h_1^{\alpha-k+1}$. In other words, in the hypothesis that the expansion (IV.18) holds, if we plot the values $\tilde{c}_k^{[p]}(\theta_{j_1,n_1})$ versus the points θ_{j_1,n_1} for

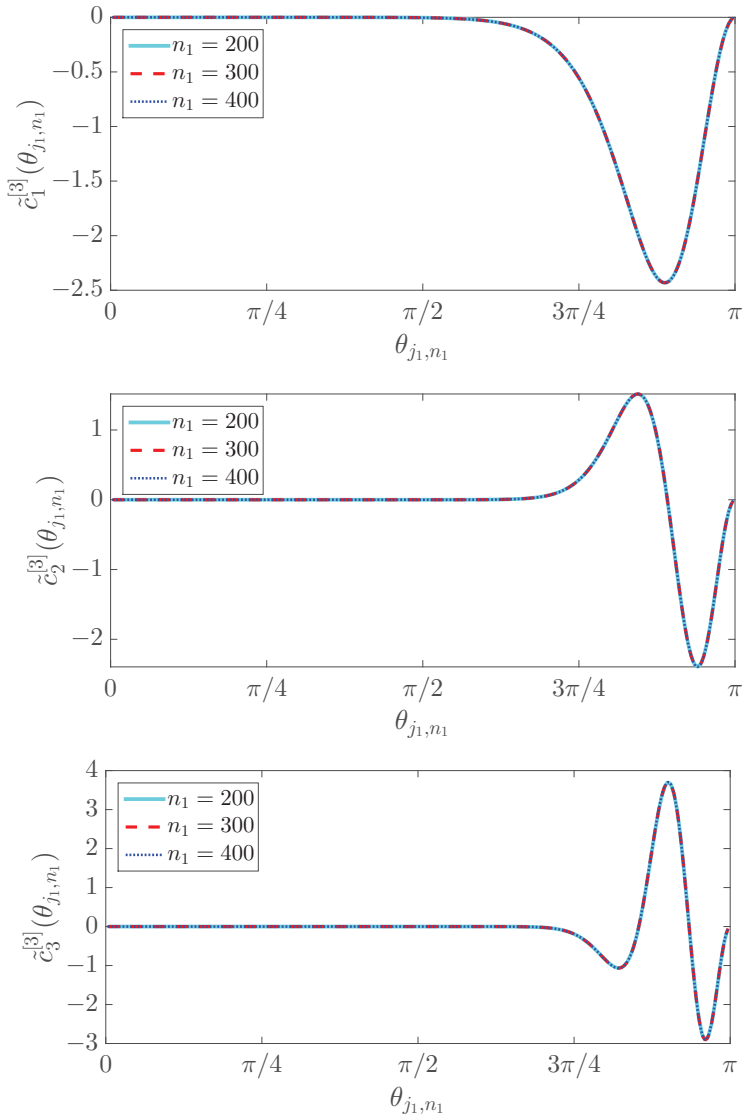


Figure IV.4: Example 6, $p = 3$: graph of the pairs $(\theta_{j_1,n_1}, \tilde{c}_k^{[3]}(\theta_{j_1,n_1}))$, $j_1 = 1, \ldots, n_1 - 1$, for $n_1 = 200, 300, 400$ and $k = 1, 2, 3$.

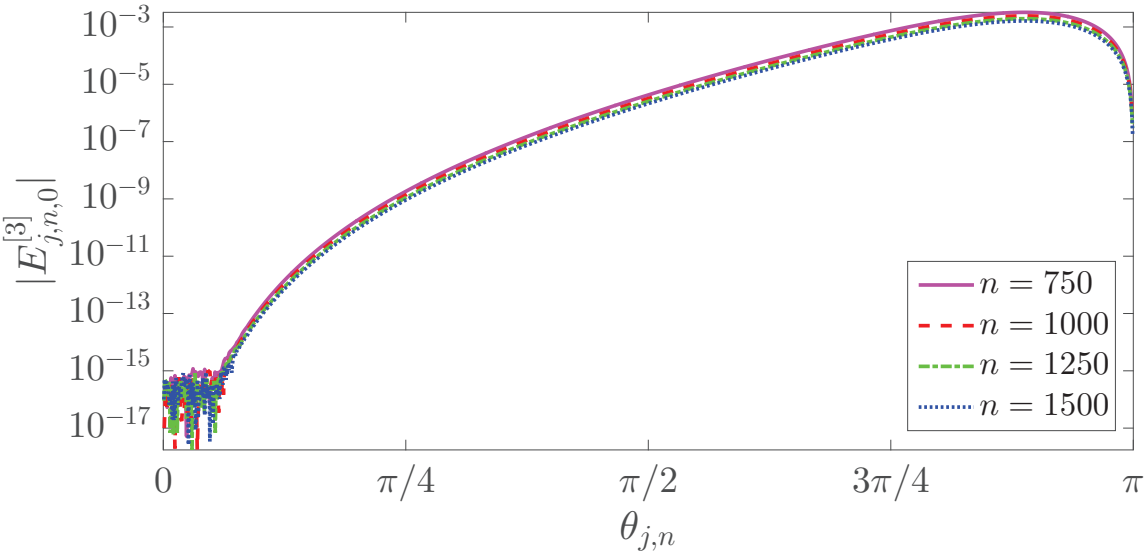


Figure IV.5: Example 6, $p = 3$: errors $|E^{[3]}_{j,n,0}|$ versus $\theta_{j,n}$ for $j = 1, \dots, n-1$ and $n = 750, 1000, 1250, 1500$.

n	750	1000	1250	1500
j	58	80	101	123
$\theta_{j,n}$	0.2429	0.2513	0.2538	0.2576

Table IV.2: Example 6, $p = 3$: first index j such that $|E^{[3]}_{j,n,0}| > 10^{-14}$ and corresponding grid point $\theta_{j,n}$, for $n = 750, 1000, 1250, 1500$.

$j_1 = 1, \dots, n_1^{[p]}$, the resulting picture should converge as $n_1 \to \infty$ to the graph of a function from $[0, \pi]$ to $\mathbb{R}$, which is, by definition, $c_k^{[p]}(\theta)$. The next examples show that this is in fact the case, thus providing a validation of the expansion (IV.18). The examples also support the following conjectures:

- the limit function $c_k^{[p]}(\theta)$ vanishes over an interval $[0, \theta(p,k)]$ with $\theta(p,k) \in (0, \pi)$;
- $y_p = \inf_{k \geq 1} \theta(p,k) > 0$;
- $\lambda_j(n^{-2} L_n^{[p]}) = e_p(\theta_{j,n})$ numerically whenever $\theta_{j,n} < \theta(p)$, where $\theta(p)$ is a point in $(0, y_p]$ which grows with p.

Example 6. *Fix $p = 3$ and let $\alpha = 3$. In Figure IV.4 we plot the pairs*

$$(\theta_{j_1,n_1}, \tilde{c}_k^{[3]}(\theta_{j_1,n_1})), \qquad j_1 = 1, \dots, n_1^{[3]} = n_1 - 1, \tag{IV.31}$$

for $n_1 = 200, 300, 400$ and $k = 1, 2, 3$. We note that, for each fixed k, the graph of the pairs (IV.31) *is essentially the same for all the considered values of n_1. In other words, this graph converges to the graph of a function $c_k^{[3]}(\theta)$ as $n_1 \to \infty$, and the convergence is essentially reached already for $n_1 = 200$, at least from the point of view of graphical visualization. Moreover, the limit function $c_k^{[3]}(\theta)$ is apparently zero over an interval $[0, \theta(3,k)]$, where $\theta(3,k) \in (0, \pi)$. An ε-approximation of $\theta(3,k)$ is obtained as the limit of $\theta_{n_1}^{(\varepsilon)}(3,k)$ for $n_1 \to \infty$, where*

$$\theta_{n_1}^{(\varepsilon)}(3,k) = \max\{\theta_{j_1,n_1} : \ 1 \leq j_1 \leq n_1 - 1, \ |\tilde{c}_k^{[3]}(\theta_{i_1,n_1})| \leq \varepsilon \text{ for all } i_1 < j_1\}$$

n_1	200	300	400	500	600
$\theta^{(\varepsilon)}_{n_1}(4,1)$	$\frac{97\pi}{200} \approx 1.5237$	$\frac{146\pi}{300} \approx 1.5289$	$\frac{194\pi}{400} \approx 1.5237$	$\frac{242\pi}{500} \approx 1.5205$	$\frac{291\pi}{600} \approx 1.5237$
$\theta^{(\varepsilon)}_{n_1}(4,2)$	$\frac{129\pi}{200} \approx 2.0263$	$\frac{194\pi}{300} \approx 2.0316$	$\frac{258\pi}{400} \approx 2.0263$	$\frac{322\pi}{500} \approx 2.0232$	$\frac{387\pi}{600} \approx 2.0263$
$\theta^{(\varepsilon)}_{n_1}(4,3)$	$\frac{145\pi}{200} \approx 2.2777$	$\frac{217\pi}{300} \approx 2.2724$	$\frac{289\pi}{400} \approx 2.2698$	$\frac{362\pi}{500} \approx 2.2745$	$\frac{434\pi}{600} \approx 2.2724$

Table IV.3: Example 7, $p = 4$: values $\theta^{(\varepsilon)}_{n_1}(4,k)$ for $k = 1,2,3$ and $n_1 = 200, 300, 400, 500, 600$, computed with the threshold $\varepsilon = 0.0005$.

n	750	1000	1250	1500
j	71	97	123	152
$\theta_{j,n}$	0.2974	0.3047	0.3091	0.3183

Table IV.4: Example 7, $p = 4$: first index j such that $|E^{[4]}_{j,n,0}| > 10^{-14}$ and corresponding grid point $\theta_{j,n}$, for $n = 750, 1000, 1250, 1500$.

and ε is a fixed threshold. Table IV.1 shows the values $\theta^{(\varepsilon)}_{n_1}(3,k)$ computed for $k = 1,2,3$ and $n_1 = 200, 300, 400, 500, 600$ with the fixed threshold $\varepsilon = 0.0005$. Both Figure IV.4 and Table IV.1 suggest that $\theta(3,k)$ grows with k. In particular, we may expect that

$$y_3 = \inf_{k \geq 1} \theta(3,k) = \theta(3,1) > 0.$$

In Figure IV.5 we plot the errors $|E^{[3]}_{j,n,0}| = |\lambda_j(n^{-2}L^{[3]}_n) - e_3(\theta_{j,n})|$ versus the points $\theta_{j,n}$ for $j = 1, \ldots, n^{[3]} = n - 1$ and $n = 750, 1000, 1250, 1500$. For the same values of n, in Table IV.2 we record the first index j such that $|E^{[3]}_{j,n,0}| > 10^{-14}$ and the corresponding grid point $\theta_{j,n}$. From Figure IV.5 and Table IV.2 we immediately see that a nontrivial portion of the spectrum of $n^{-2}L^{[3]}_n$ is exactly approximated, at least from a numerical viewpoint, by the spectral symbol $e_3(\theta)$. Moreover, the points $\theta_{j,n}$ shown in Table IV.2 apparently form a monotone increasing sequence; the limit of this sequence as $n \to \infty$, say $\theta(3) \approx 0.2576$, is a point such that the equation $\lambda_i(n^{-2}L^{[3]}_n) = e_3(\theta_{i,n})$ holds numerically whenever $\theta_{i,n} < \theta(3)$. In other words, the ratio $\theta(3)/\pi \approx 0.082$ represents the portion of the spectrum of $n^{-2}L^{[3]}_n$ which is exactly described by $e_3(\theta)$, at least numerically.

Example 7. *In this example we verbatim repeat for the case $p = 4$ what we have done in Example 6 for $p = 3$. For the sake of brevity, we do not include here any comment and we limit to report the exact analogs of Figure IV.4, Table IV.1, Figure IV.5, and Table IV.2 in Figure IV.6, Table IV.3, Figure IV.7, and Table IV.4.*

Example 8. *A comparison between Table IV.2 and Table IV.4 shows that the portion of the spectrum of $n^{-2}L^{[p]}_n$ which is exactly described by $e_p(\theta)$, at least from a numerical viewpoint, grows from $\theta(3)/\pi \approx 0.082$ for $p = 3$ to $\theta(4)/\pi \approx 0.101$ for $p = 4$. Actually, this spectrum portion increases more and more with p, i.e., $\theta(p)$ grows with p; see Figure IV.8.*

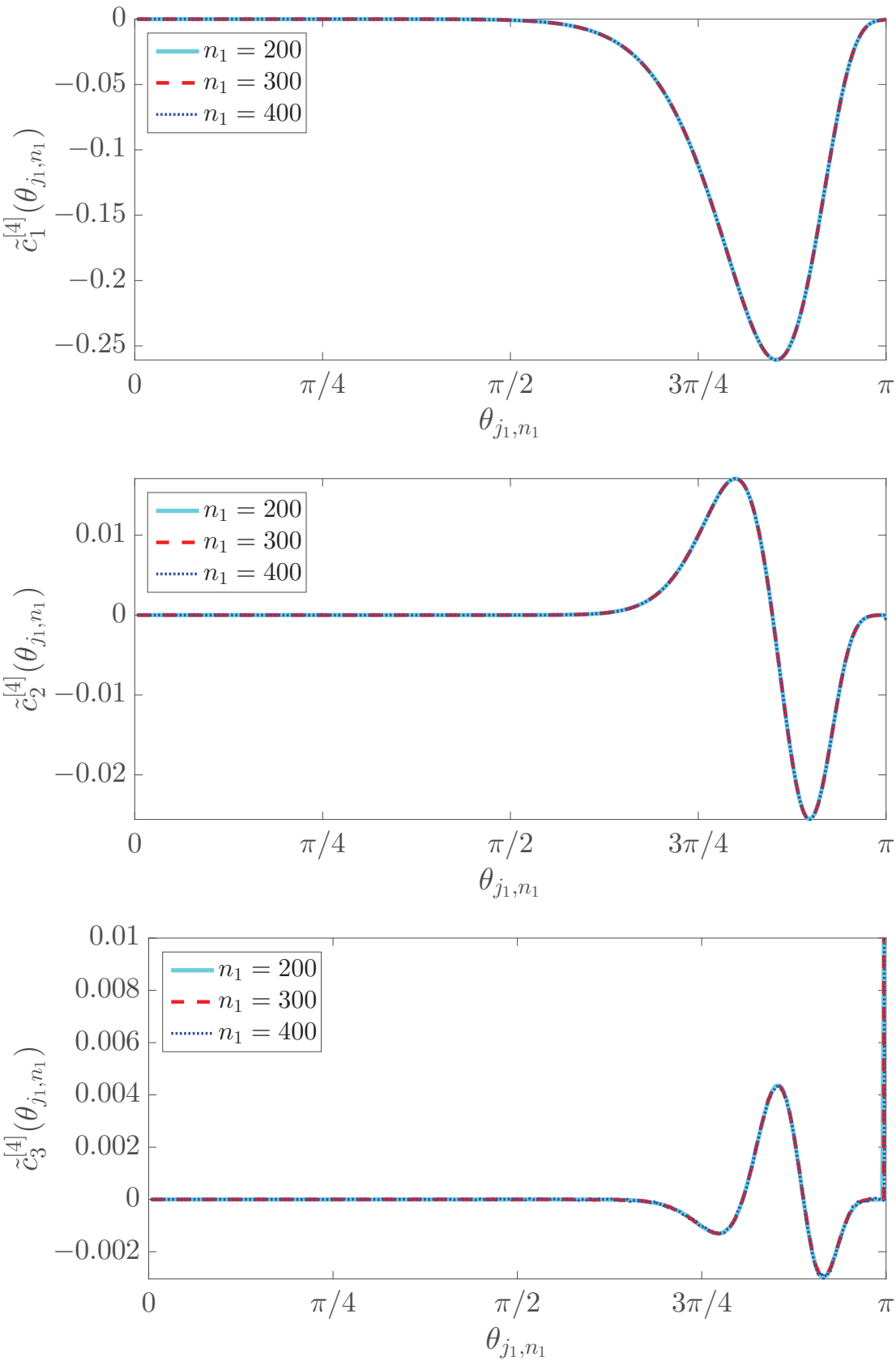


Figure IV.6: Example 7, $p = 4$: graph of the pairs $(\theta_{j_1,n_1}, \tilde{c}_k^{[4]}(\theta_{j_1,n_1}))$, $j_1 = 1, \ldots, n_1$, for $n_1 = 200, 300, 400$ and $k = 1, 2, 3$.

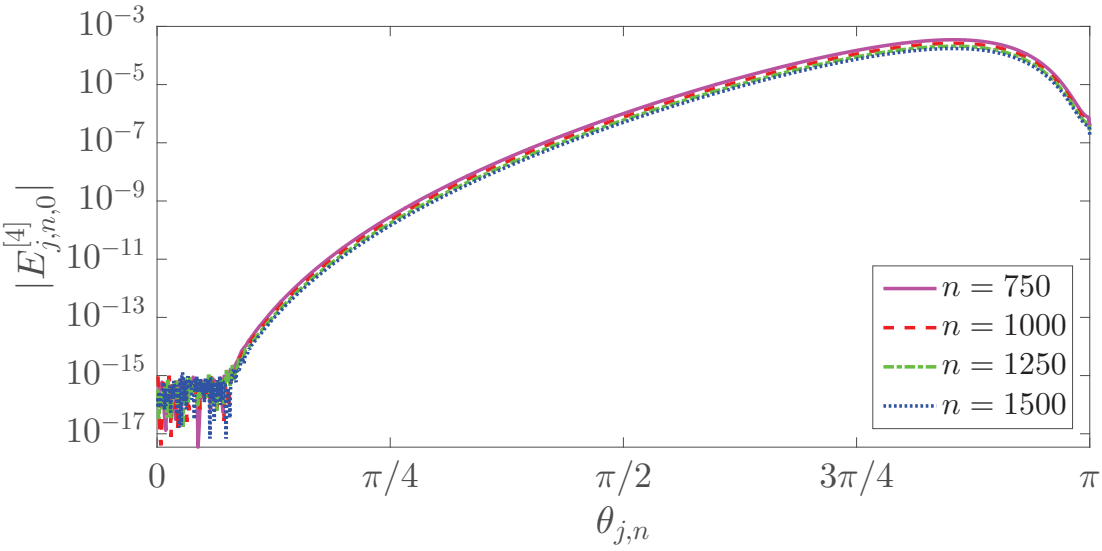


Figure IV.7: Example 7, $p = 4$: errors $|E^{[4]}_{j,n,0}|$ versus $\theta_{j,n}$ for $j = 1, \ldots, n$ and $n = 750, 1000, 1250, 1500$.

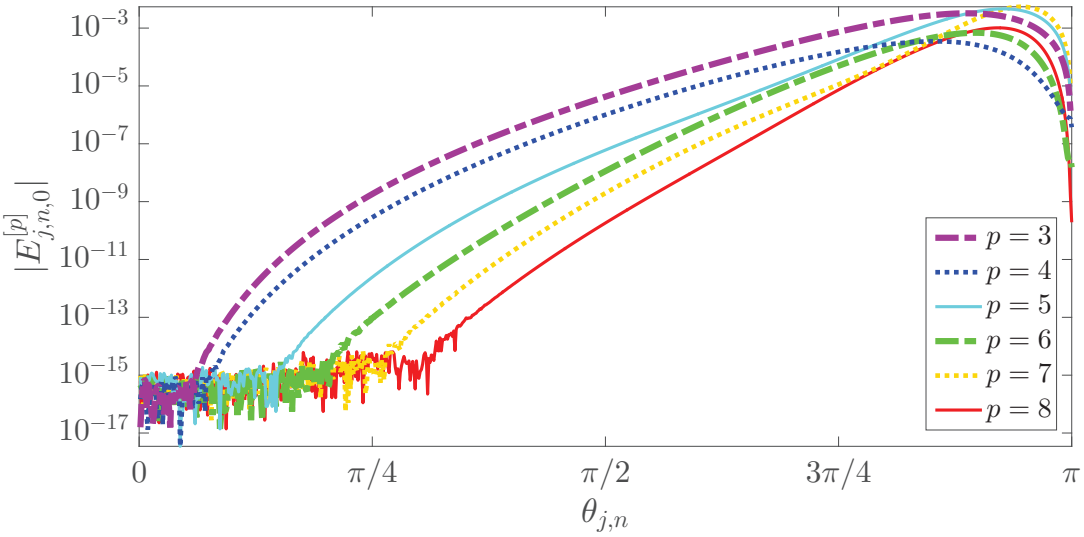


Figure IV.8: Example 8: errors $|E^{[p]}_{j,n,0}|$ versus $\theta_{j,n}$ for $j = 1, \ldots, n - \mathrm{mod}(p, 2)$ and $p = 3, \ldots, 8$, with $n = 750$.

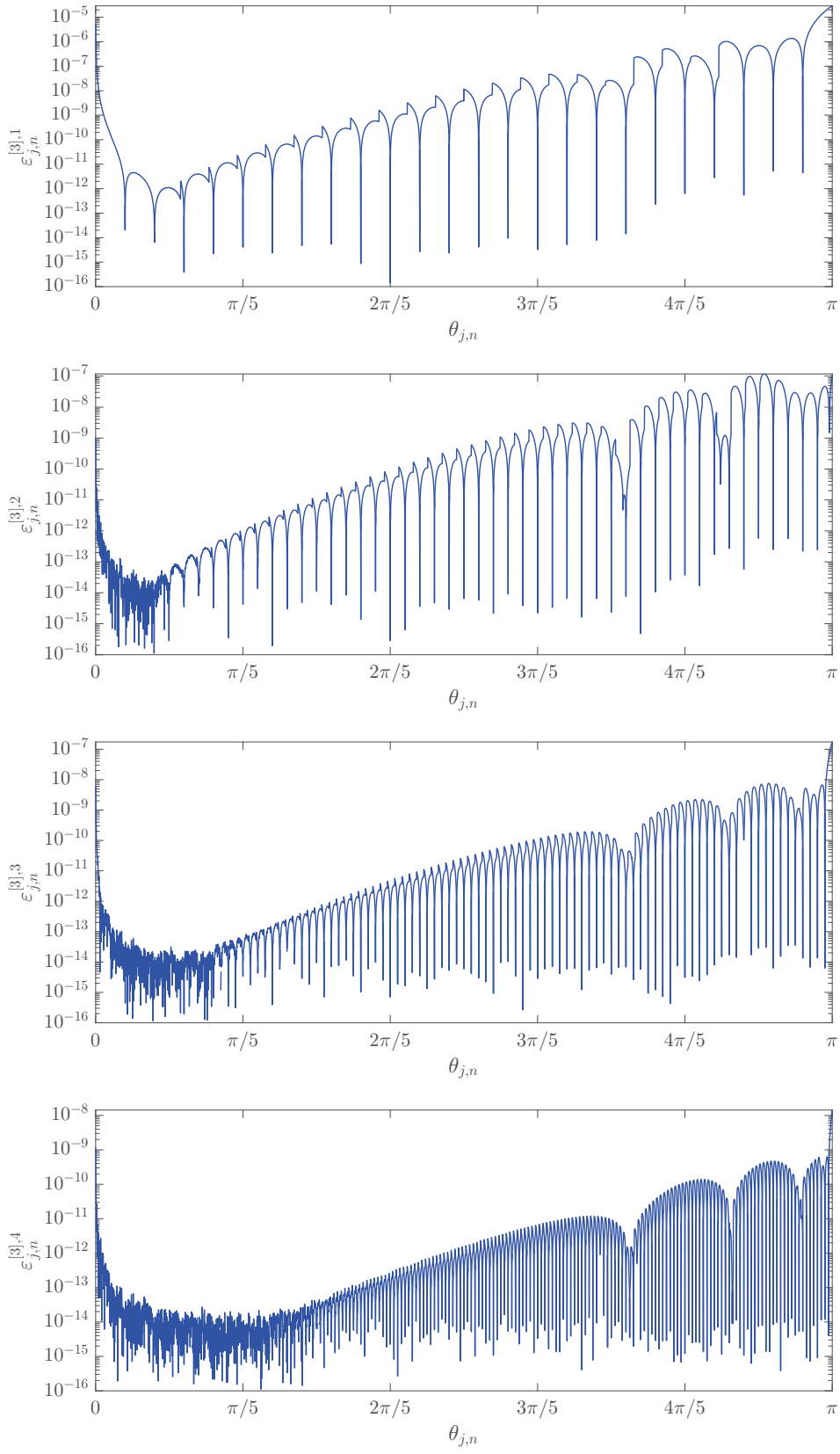


Figure IV.9: Example 9, $p = 3$: errors $\varepsilon_{j,n}^{[3],m}$ versus $\theta_{j,n}$ for $j = 1, \ldots, n-1$, in the case where $n = 5000$, $n_1 = 25 \cdot 2^{m-1}$, and $\alpha = 4$.

IV.5.2 Numerical experiments illustrating the performance of algorithm 1

Example 9. *Let $p = 3$. Suppose we want to approximate the eigenvalues of $L_n^{[3]}$ (excluding the $n_3^{\text{out}} = 2$ outliers) for $n = 5000$. Let $\tilde{\lambda}_j^{(m)}(L_n^{[3]})$ be the approximation of $\lambda_j(L_n^{[3]})$ obtained by applying Algorithm 1 with $n_1 = 25 \cdot 2^{m-1}$ and $\alpha = 4$. In Figure IV.9 we plot the relative errors*

$$\varepsilon_{j,n}^{[3],m} = \frac{|\lambda_j(L_n^{[3]}) - \tilde{\lambda}_j^{(m)}(L_n^{[3]})|}{|\lambda_j(L_n^{[3]})|}$$

versus $\theta_{j,n}$ for $j = 1, \ldots, n^{[3]} = n - 1$ and $m = 1, \ldots, 4$. We see from the figure that the errors decrease rather quickly as m increases. A careful consideration of Figure IV.9 also reveals that, aside from the exceptional minima attained in a neighborhood of $\theta = 0$,[3] the local minima of $\varepsilon_{j,n}^{[3],m}$ are attained when $\theta_{j,n}$ is approximately equal to some of the coarse grid points θ_{j_1,n_1}, $j_1 = 1, \ldots, n_1$. This is no surprise, because for $\theta_{j,n} = \theta_{j_1,n_1}$ we have $\tilde{c}_{k,j}^{[3]}(\theta_{j,n}) = \tilde{c}_k^{[3]}(\theta_{j_1,n_1})$ and $c_k^{[3]}(\theta_{j,n}) = c_k^{[3]}(\theta_{j_1,n_1})$, which means that the error of the approximation $\tilde{c}_{k,j}^{[3]}(\theta_{j,n}) \approx c_k^{[3]}(\theta_{j,n})$ reduces to the error of the approximation $\tilde{c}_k^{[3]}(\theta_{j_1,n_1}) \approx c_k^{[3]}(\theta_{j_1,n_1})$; that is, we are not introducing further error due to the interpolation process.

Example 10. *Let $p = 4$. Suppose we want to approximate the eigenvalues of $L_n^{[4]}$ (excluding the $n_4^{\text{out}} = 2$ outliers) for $n = 5000$. Let $\tilde{\lambda}_j^{(m)}(L_n^{[4]})$ be the approximation of $\lambda_j(L_n^{[4]})$ obtained by applying Algorithm 1 with $n_1 = 10 \cdot 2^{m-1}$ and $\alpha = 5$. In Figure IV.10 we plot the relative errors*

$$\varepsilon_{j,n}^{[4],m} = \frac{|\lambda_j(L_n^{[4]}) - \tilde{\lambda}_j^{(m)}(L_n^{[4]})|}{|\lambda_j(L_n^{[4]})|},$$

versus $\theta_{j,n}$ for $j = 1, \ldots, n^{[4]} = n$ and $m = 1, \ldots, 4$. Considerations analogous to those of Example 9 apply also in this case.

IV.6 Extension to the multidimensional setting

We present in this section the extension to the multidimensional setting of the analysis carried out in the previous sections. In what follows, we will systematically use the multi-index notation and the properties of tensor products as described in [76, Subsections 2.1.1 and 2.6.1]. If $w_i : D_i \to \mathbb{C}$, $i = 1, \ldots, k$, are arbitrary functions, we will denote by $w_1 \otimes \cdots \otimes w_k : D_1 \times \cdots \times D_k \to \mathbb{C}$ the tensor-product function

$$(w_1 \otimes \cdots \otimes w_k)(\xi_1, \ldots, \xi_k) = \prod_{i=1}^{k} w_i(\xi_i), \qquad (\xi_1, \ldots, \xi_k) \in D_1 \times \cdots \times D_k.$$

Problem setting

Consider the k-dimensional Laplacian eigenvalue problem

$$\begin{cases} -\Delta u(\mathbf{x}) = \lambda u(\mathbf{x}), & \mathbf{x} \in (0,1)^k, \\ u(\mathbf{x}) = 0, & \mathbf{x} \in \partial((0,1)^k). \end{cases} \tag{IV.32}$$

[3] These minima, as well as the highly oscillatory behavior of the error around $\theta = 0$, are probably due to the fact that $e_3(\theta)$ provides a numerically exact description of the spectrum of $n^{-2}L_n^{[3]}$ around $\theta = 0$; see also Example 6.

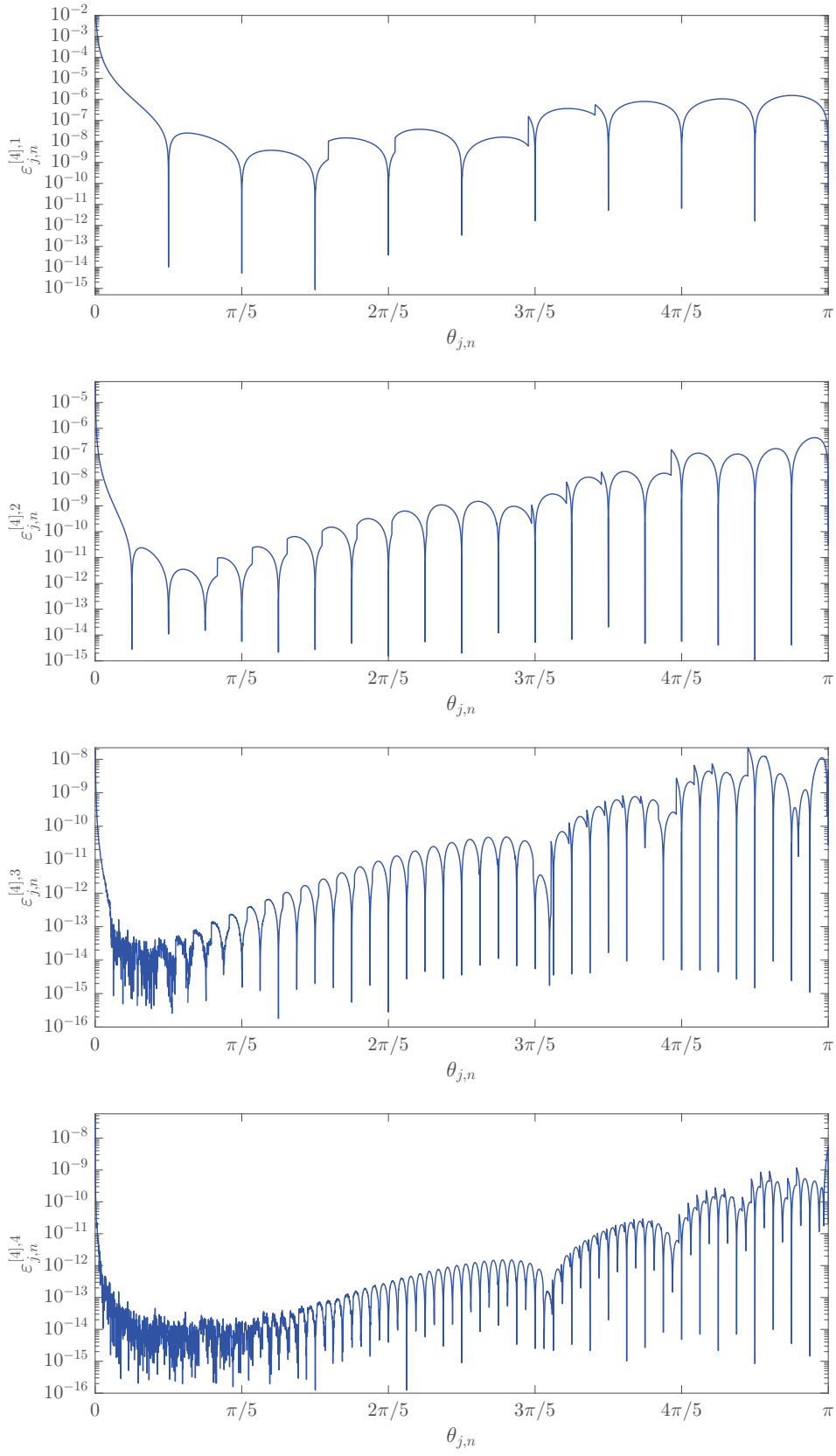


Figure IV.10: Example 10, $p = 4$: errors $\varepsilon_{j,n}^{[4],m}$ versus $\theta_{j,n}$ for $j = 1, \ldots, n$, in the case where $n = 5000$, $n_1 = 10 \cdot 2^{m-1}$, and $\alpha = 5$.

The corresponding weak formulation reads as follows: find eigenvalues $\lambda \in \mathbb{R}^+$ and eigenfunctions $u \in H_0^1((0,1)^k)$ such that, for all $v \in H_0^1((0,1)^k)$,

$$\mathrm{a}(u,v) = \lambda(u,v),$$

where

$$\mathrm{a}(u,v) = \int_{(0,1)^k} \nabla u(\mathbf{x}) \cdot \nabla v(\mathbf{x}) \mathrm{d}\mathbf{x}, \qquad (u,v) = \int_{(0,1)^k} u(\mathbf{x}) v(\mathbf{x}) \mathrm{d}\mathbf{x}.$$

In the "tensor-product version" of Galerkin's method, we choose k finite-dimensional vector spaces $\mathscr{W}_1, \dots, \mathscr{W}_k \subset H_0^1(0,1)$ and we set

$$\mathscr{W} = \mathscr{W}_1 \otimes \cdots \otimes \mathscr{W}_k = \mathrm{span}(w_1 \otimes \cdots \otimes w_k : \ w_1 \in \mathscr{W}_1, \dots, w_k \in \mathscr{W}_k) \subset H_0^1((0,1)^k).$$

Then, we define $N_s = \dim \mathscr{W}_s$ for $s = 1, \dots, k$ and $\boldsymbol{N} = (N_1, \dots, N_k)$, and we look for approximations of the exact eigenpairs

$$\lambda_{\boldsymbol{j}} = \sum_{i=1}^k j_i^2 \pi^2, \qquad u_{\boldsymbol{j}}(\mathbf{x}) = \prod_{i=1}^k \sin(j_i \pi x_i), \qquad \boldsymbol{j} = (j_1, \dots, j_k) \in \mathbb{N}^k, \tag{IV.33}$$

by solving the following Galerkin problem: find $\lambda_{\boldsymbol{j},\mathscr{W}} \in \mathbb{R}^+$ and $u_{\boldsymbol{j},\mathscr{W}} \in \mathscr{W}$, for $\boldsymbol{j} = \mathbf{e}, \dots, \boldsymbol{N}$, such that, for all $v \in \mathscr{W}$,

$$\mathrm{a}(u_{\boldsymbol{j},\mathscr{W}}, v) = \lambda_{\boldsymbol{j},\mathscr{W}}(u_{\boldsymbol{j},\mathscr{W}}, v). \tag{IV.34}$$

If $\{\varphi_{1,[s]}, \dots, \varphi_{N_s,[s]}\}$ is a basis of $\mathscr{W}_s$ for $s = 1, \dots, k$, then

$$\varphi_{\boldsymbol{i}} = \varphi_{i_1,[1]} \otimes \cdots \otimes \varphi_{i_k,[d]}, \qquad \boldsymbol{i} = \mathbf{e}, \dots, \boldsymbol{N},$$

is a basis of $\mathscr{W}$, and in view of the canonical identification between each $v \in \mathscr{W}$ and its coefficient vector with respect to $\{\varphi_{\mathbf{e}}, \dots, \varphi_{\boldsymbol{N}}\}$, solving the Galerkin problem (IV.34) is equivalent to solving the generalized eigenvalue problem

$$K\mathbf{u}_{\boldsymbol{j},\mathscr{W}} = \lambda_{\boldsymbol{j},\mathscr{W}} M \mathbf{u}_{\boldsymbol{j},\mathscr{W}}, \tag{IV.35}$$

where $\mathbf{u}_{\boldsymbol{j},\mathscr{W}}$ is the coefficient vector of $u_{\boldsymbol{j},\mathscr{W}}$ with respect to $\{\varphi_{\mathbf{e}}, \dots, \varphi_{\boldsymbol{N}}\}$,

$$K = [\mathrm{a}(\varphi_{\boldsymbol{j}}, \varphi_{\boldsymbol{i}})]_{\boldsymbol{i},\boldsymbol{j}=\mathbf{e}}^{\boldsymbol{N}} = \left[\int_{(0,1)^k} \nabla \varphi_{\boldsymbol{j}}(\mathbf{x}) \cdot \nabla \varphi_{\boldsymbol{i}}(\mathbf{x}) \mathrm{d}\mathbf{x} \right]_{\boldsymbol{i},\boldsymbol{j}=\mathbf{e}}^{\boldsymbol{N}} = \tag{IV.36}$$

$$\sum_{r=1}^k \left(\bigotimes_{s=1}^{r-1} M^{(s)} \right) \otimes K^{(r)} \otimes \left(\bigotimes_{s=r+1}^{k} M^{(s)} \right), \tag{IV.37}$$

$$M = [(\varphi_{\boldsymbol{j}}, \varphi_{\boldsymbol{i}})]_{\boldsymbol{i},\boldsymbol{j}=\mathbf{e}}^{\boldsymbol{N}} = \left[\int_{(0,1)^k} \varphi_{\boldsymbol{j}}(\mathbf{x}) \varphi_{\boldsymbol{i}}(\mathbf{x}) \mathrm{d}\mathbf{x} \right]_{\boldsymbol{i},\boldsymbol{j}=\mathbf{e}}^{\boldsymbol{N}} = \bigotimes_{s=1}^{k} M^{(s)}, \tag{IV.38}$$

and

$$K^{(s)} = \left[\int_0^1 \varphi'_{j,[s]}(x) \varphi'_{i,[s]}(x) \mathrm{d}x \right]_{i,j=1}^{N_s}, \qquad s = 1, \dots, k,$$

$$M^{(s)} = \left[\int_0^1 \varphi_{j,[s]}(x) \varphi_{i,[s]}(x) \mathrm{d}x \right]_{i,j=1}^{N_s}, \qquad s = 1, \dots, k.$$

The matrices K and M are, respectively, the stiffness matrix and the mass matrix. Both K and M are always symmetric positive definite, regardless of the basis functions $\varphi_{\mathbf{e}}, \ldots, \varphi_{\boldsymbol{N}}$. Moreover, it is clear from (IV.35) that the numerical eigenvalues $\lambda_{\boldsymbol{j},\not\!\!/}$, $\boldsymbol{j} = \mathbf{e}, \ldots, \boldsymbol{N}$, are just the eigenvalues of the matrix

$$L = M^{-1}K = \sum_{r=1}^{k} \Bigg(\bigotimes_{s=1}^{r-1} I_{N_s} \Bigg) \otimes (M^{(r)})^{-1} K^{(r)} \otimes \Bigg(\bigotimes_{s=r+1}^{k} I_{N_s} \Bigg). \tag{IV.39}$$

In the IgA approximation of (IV.32) based on uniform tensor-product B-splines of degree $\boldsymbol{p} = (p_1, \ldots, p_k)$, we look for approximations of the exact eigenpairs (IV.33) by using the tensor-product version of the Galerkin method described above, in which the basis functions $\varphi_{1,[s]}, \ldots, \varphi_{N_s,[s]}$ are chosen as the B-splines $N_{2,[p_s]}, \ldots, N_{n_s+p_s-1,[p_s]}$ for $s = 1, \ldots, k$, where the functions $N_{i_s+1,[p_s]}$, $i_s = 1, \ldots, n_s + p_s - 2$, are defined in (IV.8) for $n = n_s$ and $p = p_s$. Setting $\boldsymbol{n} = (n_1, \ldots, n_k)$, the resulting stiffness and mass matrices (IV.37)–(IV.38) are given by

$$K_{\boldsymbol{n}}^{[\boldsymbol{p}]} = \sum_{r=1}^{k} \Bigg(\bigotimes_{s=1}^{r-1} M_{n_s}^{[p_s]} \Bigg) \otimes K_{n_r}^{[p_r]} \otimes \Bigg(\bigotimes_{s=r+1}^{k} M_{n_s}^{[p_s]} \Bigg), \tag{IV.40}$$

$$M_{\boldsymbol{n}}^{[\boldsymbol{p}]} = \bigotimes_{s=1}^{k} M_{n_s}^{[p_s]}, \tag{IV.41}$$

and the numerical eigenvalues $\lambda_{\boldsymbol{j},\boldsymbol{n}}^{[\boldsymbol{p}]}$, $\boldsymbol{j} = \mathbf{e}, \ldots, \boldsymbol{n} + \boldsymbol{p} - \mathbf{2}$, are the eigenvalues of the matrix

$$L_{\boldsymbol{n}}^{[\boldsymbol{p}]} = (M_{\boldsymbol{n}}^{[\boldsymbol{p}]})^{-1} K_{\boldsymbol{n}}^{[\boldsymbol{p}]} = \sum_{r=1}^{k} \Bigg(\bigotimes_{s=1}^{r-1} I_{n_s+p_s-2} \Bigg) \otimes L_{n_r}^{[p_r]} \otimes \Bigg(\bigotimes_{s=r+1}^{k} I_{n_s+p_s-2} \Bigg), \tag{IV.42}$$

where the matrices $K_n^{[p]}, M_n^{[p]}, L_n^{[p]}$ are defined in (IV.10)–(IV.12) for all $p, n \geq 1$.

IV.6.1 Eigenvalue–eigenvector structure of $L_{\boldsymbol{n}}^{[\boldsymbol{p}]}$

We now show that the eigenvalue–eigenvector structure of $L_{\boldsymbol{n}}^{[\boldsymbol{p}]}$ is determined by the eigenvalue–eigenvector structure of the matrices $L_n^{[p]}$ for $p \in \{p_1, \ldots, p_k\}$. It will immediately follow that the eigenvalues and eigenvectors of $L_{\boldsymbol{n}}^{[\boldsymbol{p}]}$ are explicitly known for $\mathbf{e} \leq \boldsymbol{p} \leq \mathbf{2}$, because of the results of Section IV.3. Moreover, the parallel interpolation–extrapolation algorithm devised in Section IV.4 for computing the eigenvalues of $L_n^{[p]}$ also allows the computation of the eigenvalues of $L_{\boldsymbol{n}}^{[\boldsymbol{p}]}$.

For $p, n \geq 1$, let

$$L_n^{[p]} = V_n^{[p]} D_n^{[p]} (V_n^{[p]})^{-1}, \qquad D_n^{[p]} = \operatorname*{diag}_{j=1,\ldots,n+p-2} \lambda_j(L_n^{[p]}), \tag{IV.43}$$

be a spectral decomposition of $L_n^{[p]}$. Note that such a decomposition exists because $L_n^{[p]}$ is diagonalizable, because of the similarity equation

$$L_n^{[p]} = (M_n^{[p]})^{-1} K_n^{[p]} = (M_n^{[p]})^{-1/2} \big[(M_n^{[p]})^{-1/2} K_n^{[p]} (M_n^{[p]})^{-1/2} \big] (M_n^{[p]})^{1/2}.$$

It follows from (IV.43) and the properties of tensor products that

$$\begin{aligned} L_{\boldsymbol{n}}^{[\boldsymbol{p}]} &= \sum_{r=1}^{k} \left(\bigotimes_{s=1}^{r-1} I_{n_s+p_s-2} \right) \otimes L_{n_r}^{[p_r]} \otimes \left(\bigotimes_{s=r+1}^{k} I_{n_s+p_s-2} \right), \\ &= \left(\bigotimes_{s=1}^{k} V_{n_s}^{[p_s]} \right) \left[\sum_{r=1}^{k} \left(\bigotimes_{s=1}^{r-1} I_{n_s+p_s-2} \right) \otimes D_{n_r}^{[p_r]} \otimes \left(\bigotimes_{s=r+1}^{k} I_{n_s+p_s-2} \right) \right] \left(\bigotimes_{s=1}^{k} V_{n_s}^{[p_s]} \right)^{-1}, \end{aligned} \tag{IV.44}$$

which is a spectral decomposition of $L_{\boldsymbol{n}}^{[\boldsymbol{p}]}$. More explicitly, let $\mathbf{v}_{1,n}^{[p]}, \ldots, \mathbf{v}_{n+p-2,n}^{[p]}$ be the columns of $V_n^{[p]}$, i.e., the eigenvectors of $L_n^{[p]}$,

$$L_n^{[p]} \mathbf{v}_{j,n}^{[p]} = \lambda_j(L_n^{[p]}) \mathbf{v}_{j,n}^{[p]}, \qquad j = 1, \ldots, n+p-2,$$

and let

$$\mathbf{v}_{\boldsymbol{j},\boldsymbol{n}}^{[\boldsymbol{p}]} = \bigotimes_{s=1}^{k} \mathbf{v}_{j_s,n_s}^{[p_s]}, \qquad \boldsymbol{j} = \mathbf{e}, \ldots, \boldsymbol{n}+\boldsymbol{p}-\mathbf{2}. \tag{IV.45}$$

Then, we can rewrite (IV.44) as

$$L_{\boldsymbol{n}}^{[\boldsymbol{p}]} \mathbf{v}_{\boldsymbol{j},\boldsymbol{n}}^{[\boldsymbol{p}]} = \lambda_{\boldsymbol{j}}(L_{\boldsymbol{n}}^{[\boldsymbol{p}]}) \mathbf{v}_{\boldsymbol{j},\boldsymbol{n}}^{[\boldsymbol{p}]}, \qquad \boldsymbol{j} = \mathbf{e}, \ldots, \boldsymbol{n}+\boldsymbol{p}-\mathbf{2},$$

where

$$\lambda_{\boldsymbol{j}}(L_{\boldsymbol{n}}^{[\boldsymbol{p}]}) = \sum_{r=1}^{k} \lambda_{j_r}(L_{n_r}^{[p_r]}), \qquad \boldsymbol{j} = \mathbf{e}, \ldots, \boldsymbol{n}+\boldsymbol{p}-\mathbf{2}. \tag{IV.46}$$

In other words, the eigenvalue–eigenvector pairs of $L_{\boldsymbol{n}}^{[\boldsymbol{p}]}$ are

$$(\lambda_{\boldsymbol{j}}(L_{\boldsymbol{n}}^{[\boldsymbol{p}]}), \mathbf{v}_{\boldsymbol{j},\boldsymbol{n}}^{[\boldsymbol{p}]}), \qquad \boldsymbol{j} = \mathbf{e}, \ldots, \boldsymbol{n}+\boldsymbol{p}-\mathbf{2},$$

with $\mathbf{v}_{\boldsymbol{j},\boldsymbol{n}}^{[\boldsymbol{p}]}$ and $\lambda_{\boldsymbol{j}}(L_{\boldsymbol{n}}^{[\boldsymbol{p}]})$ defined as in (IV.45) and (IV.46), respectively.

We have considered the B-spline IgA approximation of the k-dimensional Laplacian eigenvalue problem (IV.32). Through tensor-product arguments, we have shown that the eigenvalue–eigenvector structure of the resulting discretization matrix $L_{\boldsymbol{n}}^{[\boldsymbol{p}]}$ is completely determined by the eigenvalue–eigenvector structure of the matrix $L_n^{[p]}$ arising from the B-spline IgA approximation of the unidimensional eigenproblem (IV.1). As for the matrix $L_n^{[p]}$, we implemented the program detailed in Items 1 to 5 of Subsection IV.1.

Chapter V

Asymptotic Expansion: extension to the block case

A substantial step forward in order to ascertain the existence of an asymptotic eigenvalue expansion for several PDE discretization matrices has been the generalization of the proposed theory to the block and preconditioned block context.

Special attention is dedicated to the generalization of the results of **Chapters III-IV** under the assumptions that $\mathbf{f}$ of is an $s \times s$ matrix-valued trigonometric polynomial with $s \geq 1$, and $T_n(\mathbf{f})$ is the associated block Toeplitz matrix, whose size is $N(n,s) = sn$.

Main contributions

The main contributions of the Chapter can be summarized as follows.

1. First we derive the conditions (either local or global) which ensure the existence of an asymptotic expansion for the eigenvalues of $T_n(\mathbf{f})$, generalizing those for the scalar-valued setting, $s = 1$.

2. We provide numerical evidence of a precise asymptotic expansion for the eigenvalues of $T_n(\mathbf{f})$, under the specific conditions derived in the first item. In particular, we conjecture on the basis of numerical experiments that for every integer $\alpha \geq 0$, every $s \geq 1$, and every $q \in \{1, \ldots, s\}$, the following asymptotic expansion holds: for all $n \in \mathbb{N}$ and $j = 1, \ldots, n$,

$$\lambda_\gamma(T_n(\mathbf{f})) = \lambda^{(q)}(\mathbf{f}(\theta_{j,n})) + \sum_{k=1}^{\alpha} c_k^{(q)}(\theta_{j,n}) h^k + E_{j,n,\alpha}^{(q)}\,, \tag{V.1}$$

   where:

   - $\gamma = \gamma(q,j) = (q-1)n + j$;
   - $\lambda_k(T_n(\mathbf{f}))$, $k \in \{1, \ldots, N(n,s)\}$, are the eigenvalues of $T_n(\mathbf{f})$, which are sorted so that, for each fixed $\bar{q} \in \{1, \ldots, s\}$, the eigenvalues $\lambda_{(\bar{q}-1)n+j}(T_n(\mathbf{f}))$, $j = 1, \ldots, n$, are arranged in non decreasing or non increasing order, depending on whether $\lambda^{(\bar{q})}(\mathbf{f})$ is increasing or decreasing (this can be seen using the local or the global condition below);

- $\{c_k^{(q)}\}_{k=1,2,\ldots,\alpha}$ is a sequence of functions from $[0,\pi]$ to $\mathbb{R}$ which depends only on $\mathbf{f}$;
- $h = \frac{1}{n+1}$ and $\theta_{j,n} = \frac{j\pi}{n+1} = j\pi h$, $j = 1,\ldots,n$;
- $E_{j,n,\alpha}^{(q)} = O(h^{\alpha+1})$ is the remainder (the error), which satisfies the inequality $|E_{j,n,\alpha}^{(q)}| \leq C_\alpha h^{\alpha+1}$ for some constant C_α depending only on α and $\mathbf{f}$.

We refer the reader to the **Chapter VI** Section VI.6 for a proof of the expansion (IV.18) for $\alpha = 0$.

3. Following the proposal for $s = 1$ of the previous chapters, we devise an interpolation–extrapolation algorithm for computing the eigenvalues of banded symmetric block Toeplitz matrices, with a high level of accuracy and a low computational cost, and we present several examples of practical interest.

4. We provide the exact formulae for the eigenvalues of the mass and stiffness matrices coming from the one dimensional $\mathbb{Q}_p$ Lagrangian Finite Element approximation of a second order elliptic differential problem and the preconditioned block matrices coming from the classical Lagrangian Finite Element approximation of the classical eigenvalue problem for the Laplacian operator in one dimension.

V.1 Conditions for the existence of block asymptotic expansion

We recall that an nth block Toeplitz matrix generated by a matrix-valued function $\phi : [-\pi,\pi] \to \mathbb{C}^{s\times s}$ is defined as

$$T_n(\phi) = \left[\hat{\phi}_{i-j}\right]_{i,j=1}^{n},$$

where the quantities $\hat{\phi}_l \in \mathbb{C}^{s\times s}$ are the Fourier coefficients of ϕ, that is,

$$\hat{\phi}_l = \frac{1}{2\pi}\int_{-\pi}^{\pi} \phi(\theta)\,\mathrm{e}^{-\iota l\theta}\mathrm{d}\theta, \qquad l \in \mathbb{Z}. \tag{V.2}$$

We refer to $\{T_n(\phi)\}_n$ as the block Toeplitz sequence generated by ϕ, which in turn is called the generating function or the symbol of $\{T_n(\phi)\}_n$. Such type of matrix sequences have been studied, especially for $s = 1$, by many authors including Szegő, Avram, Böttcher, Parter, Sibermann, Tilli, and Tyrtyshnikov (see, e.g., [77, 140] and references therein).

Furthermore, if ϕ is Hermitian almost everywhere then, by (V.2), $\hat{\phi}_{-k} = \hat{\phi}_k^*$ for every $k \in \mathbb{Z}$ and therefore each $T_n(\phi)$ is Hermitian. As a consequence, the spectrum of $T_n(\phi)$ is real. Moreover, the analytical properties of ϕ decide many delicate features of the eigenvalues of $T_n(\phi)$ such as *distribution*, *clustering*, and *localization*, as we briefly describe below without entering into technical details.

Distribution. In [140] it was proved that $\{T_n(\phi)\}_n$ has an asymptotic spectral distribution, in the Weyl sense, described by $\phi(\theta)$, under the assumption that $\phi(\theta)$ is a Lebesgue integrable matrix-valued function which is Hermitian almost everywhere. An extension to the non-Hermitian case was given in [53], by adapting the tools introduced by Tilli in [142] for complex-valued generating functions.

When the symbol ϕ is also continuous, i.e., each component $\phi_{i,j}$ is continuous, the present distribution result can be described as follows: for sufficiently large n, up to a small number of possible outliers, the eigenvalues of $T_n(\phi)$ can be grouped into s "branches" having approximate cardinality n and for each $q = 1, \ldots, s$ the eigenvalues belonging to the qth branch are approximately given by the samples over a certain uniform grid in $[-\pi, \pi]$ of the qth eigenvalue function $\lambda^{(q)}(\phi)$.

Clustering. For any $\epsilon > 0$, take an ϵ-neighborhood of the set $\mathcal{R}_\phi$, which is defined as the union of the essential ranges of the eigenvalue functions $\lambda^{(q)}(\phi)$. Then the spectrum of $\{T_n(\phi)\}_n$ is clustered at $\mathcal{R}_\phi$ in the sense that the number of the eigenvalues of $T_n(\phi)$ that do not belong to the ϵ-neighborhood of $\mathcal{R}_\phi$ is $o(n)$ as n tends to infinity. If ϕ is a Hermitian-valued trigonometric polynomial, then the number of such outliers is $O(1)$ and it is at most linearly depending on s and on the degree of the polynomial. Such clustering results are consequences of the distribution result.

Localization. Assume that $\lambda^{(q)}(\phi)$, $q = 1, \ldots, s$, are sorted in non decreasing order, that is, $\lambda^{(1)}(\phi) \le \lambda^{(2)}(\phi) \le \cdots \le \lambda^{(s)}(\phi)$. Then, for all n, the eigenvalues of $T_n(\phi)$ belong to the interval $[m_\phi, M_\phi]$, where $m_\phi = \operatorname{ess\,inf}_{\theta\in[-\pi,\pi]} \lambda^{(1)}(\phi)$ and $M_\phi = \operatorname{ess\,sup}_{\theta\in[-\pi,\pi]} \lambda^{(s)}(\phi)$. Moreover, if the function $\lambda^{(1)}(\phi)$ is not essentially constant, then the eigenvalues of $T_n(\phi)$ belong to $(m_\phi, M_\phi]$, and, if the function $\lambda^{(s)}(\phi)$ is not essentially constant, then the eigenvalues of $T_n(\phi)$ belong to $[m_\phi, M_\phi)$. For such results refer to [117, 121].

Remark 9. Part 1. *When the symbol ϕ is continuous, then each eigenvalue function $\lambda^{(q)}(\phi)$, $q = 1, \ldots, s$, is continuous and therefore the essential infimum becomes a minimum and the essential supremum becomes a maximum (because the interval $[-\pi, \pi]$ is a compact set), while the essential range is the standard range.* **Part 2.** *Finally the interval $[-\pi, \pi]$ can be replaced by the interval $[0, \pi]$ when $\phi(-\theta) = \phi(\theta)^T$: this is precisely the case we consider, see (V.4).*

In this chapter we focus on the case where the symbol is a Hermitian matrix-valued trigonometric polynomial (HTP) $\mathbf{f}$ with Fourier coefficients $\hat{\mathbf{f}}_0, \hat{\mathbf{f}}_1, \ldots, \hat{\mathbf{f}}_m \in \mathbb{R}^{s\times s}$, that is, a function of the form

$$\mathbf{f}(\theta) = \sum_{k=-m}^{m} \hat{\mathbf{f}}_k \mathrm{e}^{\iota k\theta} = \hat{\mathbf{f}}_0 + \sum_{k=1}^{m} \left(\hat{\mathbf{f}}_k \mathrm{e}^{\iota k\theta} + \hat{\mathbf{f}}_k^T \mathrm{e}^{-\mathbf{i}k\theta} \right), \qquad m = \deg(\mathbf{f}(\theta)) \in \mathbb{N},$$

where we set

$$\hat{\mathbf{f}}_{-k} = \hat{\mathbf{f}}_k^T, \quad k = 0, \ldots, m. \tag{V.3}$$

The assumptions on $\mathbf{f}(\theta)$ imply that $T_n(\mathbf{f})$ is a real symmetric block banded matrix with "block

bandwidth" $2m+1$, of the form

$$T_n(\mathbf{f}) = \begin{bmatrix}
\hat{\mathbf{f}}_0 & \hat{\mathbf{f}}_1^T & \cdots & \hat{\mathbf{f}}_m^T & & & & & & \\
\hat{\mathbf{f}}_1 & \ddots & \ddots & & \ddots & & & & & \\
\vdots & \ddots & \ddots & \ddots & & \ddots & & & & \\
\hat{\mathbf{f}}_m & & \ddots & \ddots & \ddots & & \ddots & & & \\
& \ddots & & \ddots & \ddots & \ddots & & \ddots & & \\
& & \boxed{\hat{\mathbf{f}}_m} & \cdots & \hat{\mathbf{f}}_1 & \hat{\mathbf{f}}_0 & \hat{\mathbf{f}}_1^T & \cdots & \hat{\mathbf{f}}_m^T & \\
& & & \ddots & & \ddots & \ddots & \ddots & & \ddots \\
& & & & \ddots & & \ddots & \ddots & \ddots & & \hat{\mathbf{f}}_m^T \\
& & & & & \ddots & & \ddots & \ddots & \ddots & \vdots \\
& & & & & & \ddots & & \ddots & \ddots & \hat{\mathbf{f}}_1^T \\
& & & & & & & \hat{\mathbf{f}}_m & \cdots & \hat{\mathbf{f}}_1 & \hat{\mathbf{f}}_0
\end{bmatrix}.$$

Note that from (V.3) we have

$$\mathbf{f}(\theta)^T = \hat{\mathbf{f}}_0 + \sum_{k=1}^{m} \left(\hat{\mathbf{f}}_k^T \mathrm{e}^{\iota k\theta} + \hat{\mathbf{f}}_{-k}^T \mathrm{e}^{-\mathbf{i}k\theta} \right) = \hat{\mathbf{f}}_0 + \sum_{k=1}^{m} \left(\hat{\mathbf{f}}_{-k} \mathrm{e}^{\iota k\theta} + \hat{\mathbf{f}}_k \mathrm{e}^{-\mathbf{i}k\theta} \right) = \mathbf{f}(-\theta), \tag{V.4}$$

$$\mathbf{f}(\theta)^* = \hat{\mathbf{f}}_0 + \sum_{k=1}^{m} \left(\hat{\mathbf{f}}_k^T \mathrm{e}^{-\iota k\theta} + \hat{\mathbf{f}}_{-k}^T \mathrm{e}^{\iota k\theta} \right) = \hat{\mathbf{f}}_0 + \sum_{k=1}^{m} \left(\hat{\mathbf{f}}_{-k} \mathrm{e}^{-\mathbf{i}k\theta} + \hat{\mathbf{f}}_k \mathrm{e}^{\iota k\theta} \right) = \mathbf{f}(\theta), \tag{V.5}$$

and hence $\mathbf{f}(\theta)$ has the same eigenvalues as $\mathbf{f}(-\theta)$. Thus, each eigenvalue function $\lambda^{(q)}(\mathbf{f})$ is even and we can therefore simply focus on its restriction $\lambda^{(q)}(\mathbf{f}) : [0,\pi] \to \mathbb{R}^{s\times s}$ (in accordance with the second part of Remark 9).

In view of the above *distribution*, *clustering*, and *localization* results, up to $O(1)$ possible outliers, the eigenvalues of the symmetric matrix $T_n(\mathbf{f})$ can be partitioned in s subsets (or "branches") of approximately the same cardinality n; and the eigenvalues belonging to the qth branch are approximately equal to the samples of the qth eigenvalue function $\lambda^{(q)}(\mathbf{f})$ over a uniform grid in $[0,\pi]$.

In this chapter we show that the different branches have a much finer structure and that, under mild restrictions, there exists a hierarchy of symbols which allow us to design extremely economical procedures for the computation of the eigenvalues of the matrices $T_n(\mathbf{f})$. In particular, we conjecture on the basis of numerical experiments that for every integer $\alpha \geq 0$, every $s \geq 1$, and every $q \in \{1,\ldots,s\}$, the following asymptotic expansion holds under the specific *local condition* and *global condition* that will be discussed below: for all $n \in \mathbb{N}$ and $j = 1,\ldots,n$,

$$\lambda_\gamma(T_n(\mathbf{f})) = \lambda^{(q)}(\mathbf{f}(\theta_{j,n})) + \sum_{k=1}^{\alpha} c_k^{(q)}(\theta_{j,n}) h^k + E_{j,n,\alpha}^{(q)}\,, \tag{V.6}$$

where:

- $\gamma = \gamma(q, j) = (q-1)n + j$;
- $\lambda_k(T_n(\mathbf{f}))$, $k \in \{1, \ldots, N(n,s)\}$, are the eigenvalues of $T_n(\mathbf{f})$, which are sorted so that, for each fixed $\bar{q} \in \{1, \ldots, s\}$, the eigenvalues $\lambda_{(\bar{q}-1)n+j}(T_n(\mathbf{f}))$, $j = 1, \ldots, n$, are arranged in non decreasing or non increasing order, depending on whether $\lambda^{(\bar{q})}(\mathbf{f})$ is increasing or decreasing (this can be seen using the local or the global condition below);
- $\{c_k^{(q)}\}_{k=1,2,\ldots,\alpha}$ is a sequence of functions from $[0, \pi]$ to $\mathbb{R}$ which depends only on $\mathbf{f}$;
- $h = \frac{1}{n+1}$ and $\theta_{j,n} = \frac{j\pi}{n+1} = j\pi h$, $j = 1, \ldots, n$;
- $E_{j,n,\alpha}^{(q)} = O(h^{\alpha+1})$ is the remainder (the error), which satisfies the inequality $|E_{j,n,\alpha}^{(q)}| \leq C_\alpha h^{\alpha+1}$ for some constant C_α depending only on α and $\mathbf{f}$.

We note that in the scalar-valued case $s = 1$, several theoretical and computational results are available in support of the above expansion [7, 16, 17, 19, 58, 62, 63], including also extensions to preconditioned matrices and matrices arising in a differential context [1, 58].

Unfortunately, as already shown in [7, 62, 63], the expansion (V.6) is not always satisfied even for $s = 1$. Below we give two conditions which ensure that the expansion holds.

Local condition. The eigenvalue $\lambda_\gamma(T_n(\mathbf{f}))$ can be expanded as in (V.6) if there exists $\bar{\epsilon} > 0$ such that, for all $\epsilon \in (0, \bar{\epsilon})$ and all $y \in (\lambda_\gamma(T_n(\mathbf{f})) - \epsilon, \lambda_\gamma(T_n(\mathbf{f})) + \epsilon)$, there exists a unique $q \in \{1, \ldots, s\}$ and a unique $\bar{\theta} \in [0, \pi]$ for which

$$y = \lambda^{(q)}(\mathbf{f}(\bar{\theta})). \tag{V.7}$$

Global condition. A trivial global condition is obtained by imposing that the local condition is satisfied for every eigenvalue which is not an outlier (if the eigenvalue $\lambda_\gamma(T_n(\mathbf{f}))$ is an outlier, then, by definition, it does not belong to the range of $\mathbf{f}$ and consequently relation (V.7) cannot be satisfied). A simple general assumption, which is equivalent to the trivial global condition, is that each $\lambda^{(q)}(\mathbf{f})$, $q = 1, \ldots, s$, is monotone (non increasing or non decreasing) over the interval $[0, \pi]$ and

$$\max_{\theta \in [0,\pi]} \lambda^{(q)}(\mathbf{f}) < \min_{\theta \in [0,\pi]} \lambda^{(q+1)}(\mathbf{f})$$

for $q = 1, \ldots, s-1$. In other words, the global condition can be summarized as follows: strict monotonicity of every eigenvalue function and the intersection of the ranges of two eigenvalue functions $\lambda^{(j)}(\mathbf{f})$ and $\lambda^{(k)}(\mathbf{f})$ is empty for every pair of indices $j, k \in \{1, \ldots, s\}$ such that $j \neq k$. This version of the *global condition* is of course much simpler to verify. Moreover, in the case $s = 1$ it reduces to the monotonicity condition already used in the literature; see [7, 16, 17, 19, 62, 63] and references therein.

In previous chapters we employed the asymptotic expansion (V.6) with $s = 1$ for computing an accurate approximation of $\lambda_j(T_n(f))$ for very large n, if the values $\lambda_{j_1}(T_{n_1}(f)), \ldots, \lambda_{j_k}(T_{n_k}(f))$ are available for moderately sized $n_1, \ldots, n_k$ such that $\theta_{j_1,n_1} = \cdots = \theta_{j_k,n_k} = \theta_{j,n}$. We stress that the preliminary version of the algorithm was developed in [62] and then improved in [1, 57, 58],

while the mathematical foundations of the considered expansions and few numerical tests were already present in [17].

The purpose of this chapter is to carry out this idea and to support it by numerical experiments accompanied by an appropriate error analysis in the more general case where $s > 1$. In particular, we devise an algorithm to compute $\lambda_j(T_n(\mathbf{f}))$ with a high level of accuracy and a relatively low computational cost. The algorithm is completely analogous to the extrapolation procedure [132, Section 3.4], which is employed in the context of Romberg integration to obtain high precision approximations of an integral from a few coarse trapezoidal approximations. In this regard, the asymptotic expansion (V.6) plays here the same role as the Euler-Maclaurin summation formula [132, Section 3.3].

The chapter is organized as follows. Assuming the asymptotic eigenvalue expansion (V.6), in Section IV.4, we present our extrapolation algorithm for computing the eigenvalues of the $s \times s$ block matrix $T_n(\mathbf{f})$ for $s > 1$. In Section V.3 we provide numerical experiments in support of the asymptotic eigenvalue expansion (V.6) in different cases. Furthermore, we derive exact formulae for the eigenvalues in some practical examples and for matrices coming from order p Lagrangian Finite Element approximations of a second order elliptic differential problem, which are denoted as $\mathbb{Q}_p$. Finally we provide exact formulae for the eigenvalues of the preconditioned block matrices coming from the classical Lagrangian Finite Element approximation of the classical eigenvalue problem for the Laplacian operator in one dimension. In the Section VI.6 of **Chapter VI** we formally prove (V.6) in the basic case $\alpha = 0$, and we report in detail the mass and stiffness $\mathbb{Q}_p$ elements for $p = 2, 3, 4$.

V.2 Algorithm for computing the eigenvalues of $T_n(\mathbf{f})$ for $s > 1$

Assuming that the expansion (V.6) holds and taking inspiration from [58], we propose in the present section an interpolation–extrapolation algorithm for computing the eigenvalues of $T_n(\mathbf{f})$. In what follows, for each positive integer $n \in \mathbb{N} = \{1, 2, 3, \ldots\}$ and each $s > 1$ we define $N(n, s) = sn$. Moreover, with each positive integer n we associate the stepsize $h = 1/(n+1)$ and the grid points $\theta_{j,n} = j\pi h$, $j = 1, \ldots, n$. For notational convenience, unless otherwise stated, we will always denote a positive integer and the associated stepsize in a strongly related way. For example, if the positive integer is n, then the associated stepsize is h; if the positive integer is n_1, then the associated stepsize is h_1; if the positive integer is $\bar{n}$, then the associated stepsize is $\bar{h}$; etc. Throughout this section, we make the following assumptions.

- $s > 1$ and $n, n_1, \alpha \in \mathbb{N}$ are fixed parameters.
- $n_k = 2^{k-1}(n_1 + 1) - 1$ for $k = 1, \ldots, \alpha$.
- $j_k = 2^{k-1} j_1$ where $j_1 = \{1, \ldots, n_1\}$ and $k = 1, \ldots, \alpha$; j_k are the indices such that $\theta_{j_k,n_k} = \theta_{j_1,n_1}$.

A graphical representation of the grids $\theta_{[n_k]} = \{\theta_{j_k,n_k} : \ j_k = 1, \ldots, n_k\}$, $k = 1, \ldots, \alpha$, is shown in Figure V.1 for $n_1 = 5$ and $\alpha = 4$.

For each fixed $j_1 = \{1, \ldots, n_1\}$ we apply α times the expansion (V.6) with $n = n_1, n_2, \ldots, n_\alpha$ and $j = j_1, j_2, \ldots, j_\alpha$. Since $\theta_{j_1,n_1} = \theta_{j_2,n_2} = \ldots = \theta_{j_\alpha,n_\alpha}$ (by definition of $j_2, \ldots, j_\alpha$), we obtain,

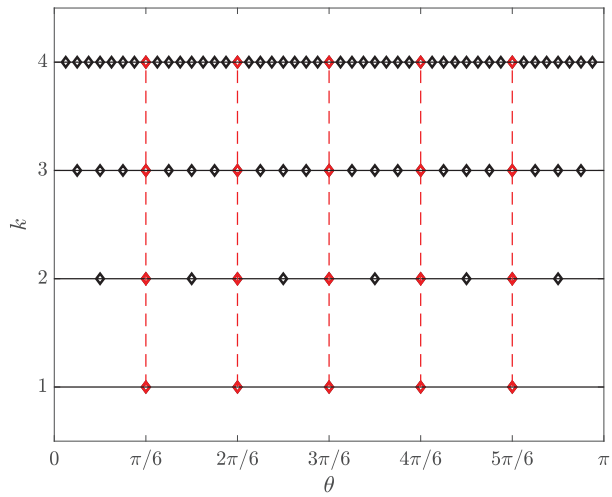


Figure V.1: Representation of the grids $\theta_{[n_k]}$, $k = 1, \ldots, \alpha$, for $n_1 = 5$ and $\alpha = 4$. The red diamonds represent the grid points θ_{j_k,n_k} and the black ones represent the rest of the grid points of $\theta_{[n_k]}$.

for $q = 1, \ldots, s$,

$$\begin{cases} E^{(q)}_{j_1,n_1,0} = c^{(q)}_1(\theta_{j_1,n_1})h_1 + c^{(q)}_2(\theta_{j_1,n_1})h_1^2 + \ldots + c^{(q)}_\alpha(\theta_{j_1,n_1})h_1^\alpha + E^{(q)}_{j_1,n_1,\alpha}, \\ E^{(q)}_{j_2,n_2,0} = c^{(q)}_1(\theta_{j_1,n_1})h_2 + c^{(q)}_2(\theta_{j_1,n_1})h_2^2 + \ldots + c^{(q)}_\alpha(\theta_{j_1,n_1})h_2^\alpha + E^{(q)}_{j_2,n_2,\alpha}, \\ \quad \vdots \\ E^{(q)}_{j_\alpha,n_\alpha,0} = c^{(q)}_1(\theta_{j_1,n_1})h_\alpha + c^{(q)}_2(\theta_{j_1,n_1})h_\alpha^2 + \ldots + c^{(q)}_\alpha(\theta_{j_1,n_1})h_\alpha^\alpha + E^{(q)}_{j_\alpha,n_\alpha,\alpha}, \end{cases} \tag{V.8}$$

where

$$E^{(q)}_{j_k,n_k,0} = \lambda_{\gamma_k}(T_{n_k}(\mathbf{f})) - \lambda^{(q)}(\mathbf{f}(\theta_{j_1,n_1})), \quad k = 1, \ldots, \alpha, \quad \gamma_k = (q-1)n_k + j_k$$

and

$$\left| E^{(q)}_{j_k,n_k,\alpha} \right| \leq C^{(q)}_\alpha h_k^{\alpha+1}, \quad k = 1, \ldots, \alpha. \tag{V.9}$$

For $q = 1, \ldots, s$, let

$$\tilde{c}^{(q)}_1(\theta_{j_1,n_1}), \ldots, \tilde{c}^{(q)}_\alpha(\theta_{j_1,n_1})$$

be the approximations of

$$c^{(q)}_1(\theta_{j_1,n_1}), \ldots, c^{(q)}_\alpha(\theta_{j_1,n_1})$$

obtained by removing all the errors $E^{(q)}_{j_1,n_1,\alpha}, \ldots, E^{(q)}_{j_\alpha,n_\alpha,\alpha}$ in (V.8) and by solving the resulting linear system:

$$\begin{cases} E^{(q)}_{j_1,n_1,0} = \tilde{c}^{(q)}_1(\theta_{j_1,n_1})h_1 + \tilde{c}^{(q)}_2(\theta_{j_1,n_1})h_1^2 + \ldots + \tilde{c}^{(q)}_\alpha(\theta_{j_1,n_1})h_1^\alpha, \\ E^{(q)}_{j_2,n_2,0} = \tilde{c}^{(q)}_1(\theta_{j_1,n_1})h_2 + \tilde{c}^{(q)}_2(\theta_{j_1,n_1})h_2^2 + \ldots + \tilde{c}^{(q)}_\alpha(\theta_{j_1,n_1})h_2^\alpha, \\ \quad \vdots \\ E^{(q)}_{j_\alpha,n_\alpha,0} = \tilde{c}^{(q)}_1(\theta_{j_1,n_1})h_\alpha + \tilde{c}^{(q)}_2(\theta_{j_1,n_1})h_\alpha^2 + \ldots + \tilde{c}^{(q)}_\alpha(\theta_{j_1,n_1})h_\alpha^\alpha. \end{cases} \tag{V.10}$$

Note that this way of computing approximations for $c^{(q)}_1(\theta_{j_1,n_1}), \ldots, c^{(q)}_\alpha(\theta_{j_1,n_1})$ is completely analogous to the Richardson extrapolation procedure that is employed in the context of Romberg

integration to accelerate the convergence of the trapezoidal rule [132, Section 3.4], with the asymptotic expansion (V.6) playing here the same role as the Euler–Maclaurin summation formula [132, Section 3.3]. For more advanced studies on extrapolation methods, we refer the reader to the classical book by Brezinski and Redivo-Zaglia [23]. The next theorem shows that, for $q = 1, \dots, s$, the approximation error $\left| c_k^{(q)}(\theta_{j_1,n_1}) - \tilde{c}_k^{(q)}(\theta_{j_1,n_1}) \right|$ is $O(h_1^{\alpha-k+1})$.

Theorem V.2.1. *There exists a constant $A_\alpha^{(q)}$ depending only on α and $q = 1, \dots, s$ such that, for $j_1 = 1, \dots, n_1$ and $k = 1, \dots, \alpha$,*

$$\left| c_k^{(q)}(\theta_{j_1,n_1}) - \tilde{c}_k^{(q)}(\theta_{j_1,n_1}) \right| \le A_\alpha^{(q)} h_1^{\alpha-k+1}, \qquad q = 1, \dots, s. \tag{V.11}$$

Proof. It is a straightforward adaptation of the proof given in [58, Theorem 1]. □

Take an $n \gg n_1$ and fix an index $j \in \{1, \dots, n\}$. We henceforth assume that $q \in \{1, 2, \dots, s\}$. To compute an approximation of $\lambda_\gamma(T_n(\mathbf{f}))$, $\gamma = (q-1)n + j$, through the expansion (V.6) we need the value $c_k^{(q)}(\theta_{j,n})$ for each $k = 1, \dots, \alpha$. Of course, $c_k^{(q)}(\theta_{j,n})$ is not available in practice, but we can approximate it by interpolating and extrapolating the values $\tilde{c}_k^{(q)}(\theta_{j_1,n_1})$, $j_1 = 1, \dots, n_1$. For example, we may define $\tilde{c}_k^{(q)}(\theta)$ as the interpolation polynomial of the data $(\theta_{j_1,n_1}, \tilde{c}_k^{(q)}(\theta_{j_1,n_1}))$, $j_1 = 1, \dots, n_1$, — so that $\tilde{c}_k^{(q)}(\theta)$ is expected to be an approximation of $c_k^{(q)}(\theta)$ over the whole interval $[0, \pi]$ — and take $\tilde{c}_k^{(q)}(\theta_{j,n})$ as an approximation to $c_k^{(q)}(\theta_{j,n})$. It is known, however, that interpolating over a large number of uniform nodes is not advisable, as it may give rise to spurious oscillations (Runge's phenomenon). It is therefore better to adopt another kind of approximation. An alternative could be the following: we approximate $c_k^{(q)}(\theta)$ by the spline function $\tilde{c}_k^{(q)}(\theta)$ which is linear on each interval $[\theta_{j_1,n_1}, \theta_{j_1+1,n_1}]$ and takes the value $\tilde{c}_k^{(q)}(\theta_{j_1,n_1})$ at θ_{j_1,n_1} for all $j_1 = 1, \dots, n_1$. This strategy removes for sure any spurious oscillation, yet it is not accurate. In particular, it does not preserve the accuracy of approximation at the nodes θ_{j_1,n_1} established in Theorem V.2.1, i.e., there is no guarantee that $|c_k^{(q)}(\theta) - \tilde{c}_k^{(q)}(\theta)| \le B_\alpha^{(q)} h_1^{\alpha-k+1}$ for $\theta \in [0, \pi]$ or $|c_k^{(q)}(\theta_{j,n}) - \tilde{c}_k^{(q)}(\theta_{j,n})| \le B_\alpha^{(q)} h_1^{\alpha-k+1}$ for $j = 1, \dots, n$, with $B_\alpha^{(q)}$ being a constant depending only on α and q. As proved in Theorem IV.4.2, a local approximation strategy that preserves the accuracy (V.11), at least if $c_k^{(q)}(\theta)$ is sufficiently smooth, is the following: let $\theta^{(1)}, \dots, \theta^{(\alpha-k+1)}$ be $\alpha - k + 1$ points of the grid $\{\theta_{1,n_1}, \dots, \theta_{n_1,n_1}\}$ which are closest to the point $\theta_{j,n}$,[1] and let $\tilde{c}_{k,j}^{(q)}(\theta)$ be the interpolation polynomial of the data $(\theta^{(1)}, \tilde{c}_k^{(q)}(\theta^{(1)})), \dots, (\theta^{(\alpha-k+1)}, \tilde{c}_k^{(q)}(\theta^{(\alpha-k+1)}))$; then, we approximate $c_k^{(q)}(\theta_{j,n})$ by $\tilde{c}_{k,j}^{(q)}(\theta_{j,n})$. Note that, by selecting $\alpha - k + 1$ points from $\{\theta_{1,n_1}, \dots, \theta_{n_1,n_1}\}$, we are implicitly assuming that $n_1 \ge \alpha - k + 1$.

Theorem V.2.2. *Let $1 \le k \le \alpha$, and suppose $n_1 \ge \alpha - k + 1$ and $c_k^{(q)} \in C^{\alpha-k+1}[0, \pi]$. For $j = 1, \dots, n$, if $\theta^{(1)}, \dots, \theta^{(\alpha-k+1)}$ are $\alpha - k + 1$ points of $\{\theta_{1,n_1}, \dots, \theta_{n_1,n_1}\}$ which are closest to $\theta_{j,n}$, and if $\tilde{c}_{k,j}^{(q)}(\theta)$ is the interpolation polynomial of the data*

$$(\theta^{(1)}, \tilde{c}_k^{(q)}(\theta^{(1)})), \dots, (\theta^{(\alpha-k+1)}, \tilde{c}_k^{(q)}(\theta^{(\alpha-k+1)})),$$

[1] These $\alpha - k + 1$ points are uniquely determined by $\theta_{j,n}$ except in the following two cases: (a) $\theta_{j,n}$ coincides with a grid point θ_{j_1,n_1} and $\alpha - k + 1$ is even; (b) $\theta_{j,n}$ coincides with the midpoint between two consecutive grid points $\theta_{j_1,n_1}, \theta_{j_1+1,n_1}$ and $\alpha - k + 1$ is odd.

then

$$\left| c_k^{(q)}(\theta_{j,n}) - \tilde{c}_{k,j}^{(q)}(\theta_{j,n}) \right| \leq B_\alpha^{(q)} h_1^{\alpha-k+1} \tag{V.12}$$

for some constant $B_\alpha^{(q)}$ depending only on α and q.

Proof. It is a straightforward adaptation of the proof of [58, Theorem 2]. □

We are now ready to formulate our algorithm for computing the eigenvalues of $T_n(\mathbf{f})$.

Algorithm 2. *Given $n, n_1, \alpha \in \mathbb{N}$ with $n_1 \geq \alpha$, we compute approximations of $\lambda_\gamma(T_n(\mathbf{f}))$, $\gamma = (q-1)n + j$, for $j = 1, \ldots, n$ and $q = 1, \ldots, s$ as follows.*

1. *For $j_1 = \{1, \ldots, n_1\}$, compute $\tilde{c}_k^{(q)}(\theta_{j_1,n_1})$, for $k = 1, \ldots, \alpha$, by solving* (V.10).
2. *For $j = 1, \ldots, n$,*
   - *for $k = 1, \ldots, \alpha$*
     - *determine $\alpha - k + 1$ points $\theta^{(1)}, \ldots, \theta^{(\alpha-k+1)} \in \{\theta_{1,n_1}, \ldots, \theta_{n_1,n_1}\}$ which are closest to $\theta_{j,n}$;*
     - *compute $\tilde{c}_{k,j}^{(q)}(\theta_{j,n})$, where $\tilde{c}_{k,j}^{(q)}(\theta)$ is the interpolation polynomial of the data*

       $$(\theta^{(1)}, \tilde{c}_k^{(q)}(\theta^{(1)})), \ \ldots, \ (\theta^{(\alpha-k+1)}, \tilde{c}_k^{(q)}(\theta^{(\alpha-k+1)}));$$

   - *compute $\tilde{\lambda}_\gamma(T_n(\mathbf{f})) = \lambda^{(q)}(\mathbf{f}(\theta_{j,n})) + \sum_{k=1}^{\alpha} \tilde{c}_{k,j}^{(q)}(\theta_{j,n}) h^k$.*
3. *Return the vector $(\tilde{\lambda}_{(q-1)n+1}(T_n(\mathbf{f})), \tilde{\lambda}_{(q-1)n+2}(T_n(\mathbf{f})), \ldots, \tilde{\lambda}_{qn}(T_n(\mathbf{f}))$ as an approximation to the vector $(\lambda_{(q-1)n+1}(T_n(\mathbf{f})), \lambda_{(q-1)n+2}(T_n(\mathbf{f})) \ldots, \lambda_{qn}(T_n(\mathbf{f}))$.*

Remark 10. *Algorithm 2 is specifically designed for computing $\lambda_\gamma(T_n(\mathbf{f}))$ in the case where n is quite large. When applying this algorithm, it is implicitly assumed that n_1 and α are small (much smaller than n), so that each $n_k = 2^{k-1}(n_1 + 1) - 1$ is small as well and the computation of the eigenvalues $\tilde{\lambda}_\gamma(T_n(\mathbf{f}))$ — which is required in the first step — can be efficiently performed by any standard eigensolver (e.g., the* MATLAB `eig` *function).*

The last theorem of the current section provides an estimate for the approximation error made by Algorithm 2.

Theorem V.2.3. *Let $n \geq n_1 \geq \alpha$ and $c_k^{(q)} \in C^{\alpha-k+1}[0, \pi]$ for $k = 1, \ldots, \alpha$. Let*

$$(\tilde{\lambda}_{(q-1)n+1}(T_n(\mathbf{f})), \tilde{\lambda}_{(q-1)n+2}(T_n(\mathbf{f})) \ldots, \tilde{\lambda}_{qn}(T_n(\mathbf{f}))$$

be the approximation of $(\lambda_{(q-1)n+1}(T_n(\mathbf{f})), \lambda_{(q-1)n+2}(T_n(\mathbf{f})) \ldots, \lambda_{qn}(T_n(\mathbf{f}))$ computed by Algorithm 2. Then, there exists a constant $D_\alpha^{(q)}$ depending only on α and s such that, for $j = 1, \ldots, n$, $\gamma = (q-1)n + j$,

$$\left| \lambda_\gamma(T_n(\mathbf{f})) - \tilde{\lambda}_\gamma(T_n(\mathbf{f})) \right| \leq D_\alpha^{(q)} h h_1^\alpha. \tag{V.13}$$

Proof. By (V.6) and Theorem V.2.2,

$$\begin{aligned}
&\left|\lambda_\gamma(T_n(\mathbf{f})) - \tilde{\lambda}_\gamma(T_n(\mathbf{f}))\right| = \\
&\left|\lambda^{(q)}(\mathbf{f}(\theta_{j,n})) + \sum_{k=1}^{\alpha} c_k^{(q)}(\theta_{j,n})h^k + E_{j,n,\alpha}^{(q)} - \lambda^{(q)}(\mathbf{f}(\theta_{j,n})) - \sum_{k=1}^{\alpha} \tilde{c}_{k,j}^{(q)}(\theta_{j,n})h^k\right| \leq \\
&\sum_{k=1}^{\alpha} \left|c_k^{(q)}(\theta_{j,n}) - \tilde{c}_{k,j}^{(q)}(\theta_{j,n})\right| h^k + \left|E_{j,n,\alpha}^{(q)}\right| \leq B_\alpha^{(q)} \sum_{k=1}^{\alpha} h_1^{\alpha-k+1}h^k + C_\alpha^{(q)}h^{\alpha+1} \leq \\
&h\left(\alpha B_\alpha^{(q)}h_1^\alpha + C_\alpha^{(q)}h_1^\alpha\right) \leq D_\alpha^{(q)}h_1^\alpha h,
\end{aligned}$$

where $D_\alpha^{(q)} = (\alpha+1)\max\left(B_\alpha^{(q)}, C_\alpha^{(q)}\right)$. □

V.3 Numerical experiments

In the current section we present a selection of numerical experiments to validate the algorithms based on the asymptotic expansion (V.6) in different cases where $\mathbf{f}$ is matrix-valued, and we give exact formulae for the eigenvalues in some examples of practical interest.

We test the asymptotic expansion and the interpolation–extrapolation algorithm in Section V.2 in order to obtain an approximation of the eigenvalues $\lambda_\gamma(T_n(\mathbf{f}))$, $\gamma = 1, \ldots, sn$, for large n.

Example 1. We show that the expansion and the associated interpolation–extrapolation algorithm can be applied to the whole spectrum, since the symbol satisfies the *global condition*.

Example 2. We show that the expansion and the interpolation–extrapolation algorithm can be *locally* applied for computing the approximation of the eigenvalues verifying the *local condition*. In this particular case, the *global condition* does not hold because the intersection of ranges of two eigenvalue functions is a nontrivial interval and in addition there exists an index $q \in \{1, \ldots, s\}$ such that $\lambda^{(q)}(\mathbf{f})$ is non-monotone.

Example 3. We show that the expansion and interpolation–extrapolation algorithm can be *locally* applied for the computation of the eigenvalues satisfying the *local condition*. For the specific example, the *global condition* does not hold since there exists an index $q \in \{1, \ldots, s\}$ such that $\lambda^{(q)}(\mathbf{f})$ is non-monotone either globally on $[0, \pi]$ or just on a subinterval contained in $[0, \pi]$.

Example 4. We show how to bypass the *local condition* in a few special cases: in fact, using different sampling grids, we can recover exact formulas for parts of the spectrum, where the assumption of monotonicity is violated.

Example 5. We give a closed formula for the eigenvalues of matrices arising from the discretization of a second order elliptic differential problem by the rectangular Lagrange Finite Element method with polynomials of degree $p > 1$, usually denoted as $\mathbb{Q}_p$ elements. Moreover we provide the exact formulae for the eigenvalues of the preconditioned block matrices stemming from the $\mathbb{Q}_p$ Lagrangian FEM of the classical eigenvalue problem for the Laplacian operator in one dimension.

The number of the eigenvalue functions, which verify the *global condition*, depends on the order of the $\mathbb{Q}_p$ elements. In this specific setting we have $s = p$.

In Examples 1–3 we do not compute analytically the eigenvalue functions of $\mathbf{f}$, but, for $q = 1, \dots s$, we are able to provide an "exact" evaluation of $\lambda^{(q)}(\mathbf{f})$ at θ_{j_k,n_k}, $j_k = 1, \dots, n_k$, by exploiting the following procedure:

- sample $\mathbf{f}$ at θ_{j_k,n_k}, $j_k = 1, \dots, n_k$, obtaining n_k $s \times s$ matrices, M_{j_k}, $j_k = 1, \dots, n_k$;
- for each $j_k = 1, \dots, n_k$, compute the s eigenvalues of M_{j_k}, $\lambda_q(M_{j_k})$, $q = 1, \dots, s$;
- for a fixed $q = 1, \dots s$, the evaluation of $\lambda^{(q)}(\mathbf{f})$ at θ_{j_k,n_k}, $j_k = 1, \dots, n_k$, is given by $\lambda_q(M_{j_k})$, $j_k = 1, \dots, n_k$.

This procedure is justified by the fact that here $\mathbf{f}$ is a trigonometric polynomial and, denoting by $C_{n_k}(\mathbf{f})$ the circulant matrix generated by $\mathbf{f}$, the eigenvalues of $C_{n_k}(\mathbf{f})$ are given by the evaluations of $\lambda^{(q)}(\mathbf{f})$ at the grid points $\theta_{r,n_k} = 2\pi\frac{r}{n_k}$, $r = 0, \dots, n_k - 1$, since

$$C_{n_k}(\mathbf{f}) = (F_{n_k} \otimes I_s) D_{n_k}(\mathbf{f}) (F_{n_k} \otimes I_s)^*,$$

where

$$D_{n_k}(\mathbf{f}) = \mathrm{diag}_{0 \le r \le n_k - 1} \left(\mathbf{f}\left(\theta_{r,n_k}\right) \right), \quad \theta_{r,n_k} = 2\pi \frac{r}{n_k}, \quad F_{n_k} = \frac{1}{\sqrt{n_k}} \left(\mathrm{e}^{-\iota 2\pi \frac{jr}{n_k}} \right)_{j,r=0}^{n_k-1},$$

and I_s the $s \times s$ identity matrix [78]. Furthermore, by exploiting the localization results [117, 121] stated in the introduction, we know that each eigenvalue of $T_n(\mathbf{f})$, for each n, belongs to the interval

$$\left(\min_{\theta \in [0,\pi]} \lambda^{(1)}(\mathbf{f}), \max_{\theta \in [0,\pi]} \lambda^{(s)}(\mathbf{f}) \right).$$

V.3.1 Global condition example

Example 1.

In this example we have block size $s = 3$, and each eigenvalue function $\lambda^{(q)}(\mathbf{f}), q = 1, 2, 3$, is strictly monotone over $[0, \pi]$. The eigenvalue functions satisfy

$$\max_{\theta \in [0,\pi]} \lambda^{(1)}(\mathbf{f}) < \min_{\theta \in [0,\pi]} \lambda^{(2)}(\mathbf{f}),$$
$$\max_{\theta \in [0,\pi]} \lambda^{(2)}(\mathbf{f}) < \min_{\theta \in [0,\pi]} \lambda^{(3)}(\mathbf{f}).$$

In Figure V.2 the graphs of the three eigenvalue functions are shown.

The Toeplitz matrix generated by $\mathbf{f}$ is a pentadiagonal block matrix, $T_n(\mathbf{f}) \in \mathbb{R}^{N\times N}$, where $N = 3n$, and all the blocks belong to $\mathbb{R}^{3\times 3}$, that is

$$T_n(\mathbf{f}) = \begin{bmatrix} \hat{\mathbf{f}}_0 & \hat{\mathbf{f}}_1 & \hat{\mathbf{f}}_2 & & & \\ \hat{\mathbf{f}}_1 & \ddots & \ddots & \ddots & & \\ \hat{\mathbf{f}}_2 & \ddots & \ddots & \ddots & \ddots & \\ & \ddots & \ddots & \ddots & \ddots & \hat{\mathbf{f}}_2 \\ & & \ddots & \ddots & \ddots & \hat{\mathbf{f}}_1 \\ & & & \hat{\mathbf{f}}_2 & \hat{\mathbf{f}}_1 & \hat{\mathbf{f}}_0 \end{bmatrix}, \tag{V.14}$$

$$\hat{\mathbf{f}}_0 = \begin{bmatrix} 50 & 2 & 0 \\ 2 & -55 & 2 \\ 0 & 2 & 10 \end{bmatrix}, \quad \hat{\mathbf{f}}_1 = \begin{bmatrix} 11 & -1 & 0 \\ -1 & -6 & -1 \\ 0 & -1 & 9 \end{bmatrix}, \quad \hat{\mathbf{f}}_2 = \begin{bmatrix} 1 & 0 & 2 \\ 0 & 1 & 0 \\ 2 & 0 & 1 \end{bmatrix}.$$

Here $\mathbf{f}$ is such that the *global condition* is satisfied. Hence we can use the asymptotic expansion and Algorithm 1 to get an accurate approximation of the eigenvalues of $T_n(\mathbf{f})$ for a large n. Solving system (V.10) with $\alpha = 4$ and $n_1 = 100$, we obtain the approximation of $c_k^{(q)}(\theta_{j_1,n_1})$, $k = 1,\dots,\alpha$. In Figure V.3 the approximated expansion functions $\tilde{c}_k^{(q)}(\theta_{j_1,n_1})$, $k = 1,\dots,\alpha$, $q = 1,\dots,s$ are shown for each eigenvalue function. Once that, for a fixed $q = 1,\dots,s$, the values $\tilde{c}_k^{(q)}(\theta_{j_1,n_1})$, $k = 1,\dots,\alpha$, $j_1 = 1,\dots,n_1$ are known, we can finally compute $\tilde{\lambda}_\gamma(T_n(\mathbf{f}))$ for $n = 10000$, by using (V.6). For simplicity we plot the eigenvalue functions and also the expansion errors, $E_{j_1,n_1,0}^{(q)}$, for $q = 1,2,3$. In the top panel of Figure V.4 (in black) we show the errors, $E_{j,n,0}^{(q)}$, $q = 1,\dots,3$, versus γ, from direct calculation of

$$\lambda_\gamma(T_n(\mathbf{f})) - \lambda^{(q)}(\mathbf{f}(\theta_{j,n})),$$

for $j = 1,\dots,n$, $q = 1,\dots,3$. As expected, with $\alpha = 0$, the errors $E_{j,n,0}^{(q)}$, $q = 1,\dots,3$, are rather large. In the top panel of Figure V.4, comparing $E_{j,n,0}^{(q)}$ with errors $\tilde{E}_{j,n,\alpha}^{(q)}$, $q = 1,\dots,3$, we see the errors are significantly reduced if we calculate $\tilde{\lambda}_\gamma(T_n(\mathbf{f}))$, $\gamma = 1,\dots,3n$, shown in the bottom panel of Figure V.4, using Algorithm 1, with $\alpha = 4$, $n_1 = 100$, and $n = 10000$. Furthermore, a careful study of the top panel of Figure V.4 (coloured) also reveals that, for $q = 1,\dots,s$, $\tilde{E}_{j,n,\alpha}^{(q)}$ have local minima, attained when $\theta_{j,n}$ is approximately equal to some of the coarse grid points θ_{j_1,n_1}, $j_1 = 1,\dots,n_1$. This is no surprise, because for $\theta_{j,n} = \theta_{j_1,n_1}$ we have $\tilde{c}_{k,j}^{(q)}(\theta_{j,n}) = \tilde{c}_k^{(q)}(\theta_{j_1,n_1})$ and $c_k^{(q)}(\theta_{j,n}) = c_k^{(q)}(\theta_{j_1,n_1})$, which means that the error of the approximation $\tilde{c}_{k,j}^{(q)}(\theta_{j,n}) \approx c_k^{(q)}(\theta_{j,n})$ reduces to the error of the approximation $\tilde{c}_k^{(q)}(\theta_{j_1,n_1}) \approx c_k^{(q)}(\theta_{j_1,n_1})$. The latter implies that we are not introducing further errors due to the interpolation process.

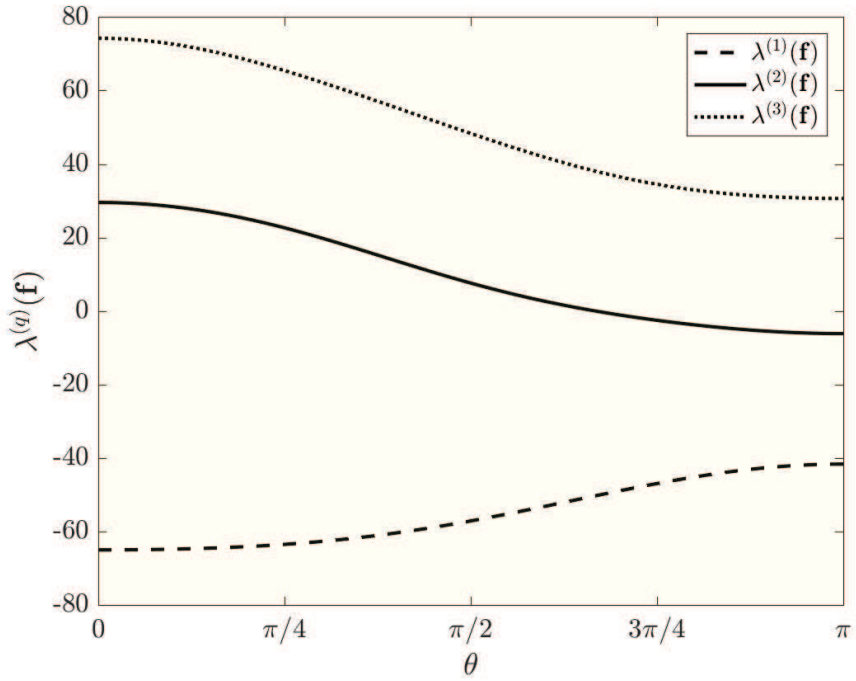


Figure V.2: Example 1: The three eigenvalue functions, $\lambda^{(q)}(\mathbf{f}), q = 1, 2, 3$.

We point out that the results in the presented example, and those in the following ones, have been provided using MATLAB. Proper timing experiments have not been conducted but in [61] the authors show, for the scalar setting, that for $n = 10^6$, LAPACK takes approximatively 10 hours of computations, whereas the matrix-less method takes approximatively 10 minutes.

We test the accuracy of the algorithm also for a more demanding computation with a matrix size n of order $O(10^5)$.

In Figure V.5 we report the error curves $E^{(q)}_{j,n,0}$ and $\tilde{E}^{(q)}_{j,n,\alpha}$, $q = 1, \ldots, 3$, $\alpha = 4$ and $n_1 = 100$, for a more costly size $n = 2 \cdot 10^5$, respect to that in V.4.

V.3.2 Local condition: intersection of the ranges

Example 2.

In the present example we choose block size $s = 3$, with eigenvalue functions $\lambda^{(1)}(\mathbf{f})$ and $\lambda^{(3)}(\mathbf{f})$ being strictly monotone on $[0, \pi]$. The second eigenvalue function, $\lambda^{(2)}(\mathbf{f})$, is non-monotone on a small subinterval of $[0, \pi]$. Furthermore the range of $\lambda^{(2)}(\mathbf{f})$ intersects that of $\lambda^{(3)}(\mathbf{f})$, that is

$$\max_{\theta\in[0,\pi]} \lambda^{(1)}(\mathbf{f}) < \min_{\theta\in[0,\pi]} \lambda^{(2)}(\mathbf{f}),$$
$$\max_{\theta\in[0,\pi]} \lambda^{(2)}(\mathbf{f}) > \min_{\theta\in[0,\pi]} \lambda^{(3)}(\mathbf{f}).$$

When comparing with Example 1, the only difference in forming the matrix $T_n(\mathbf{f})$ consists

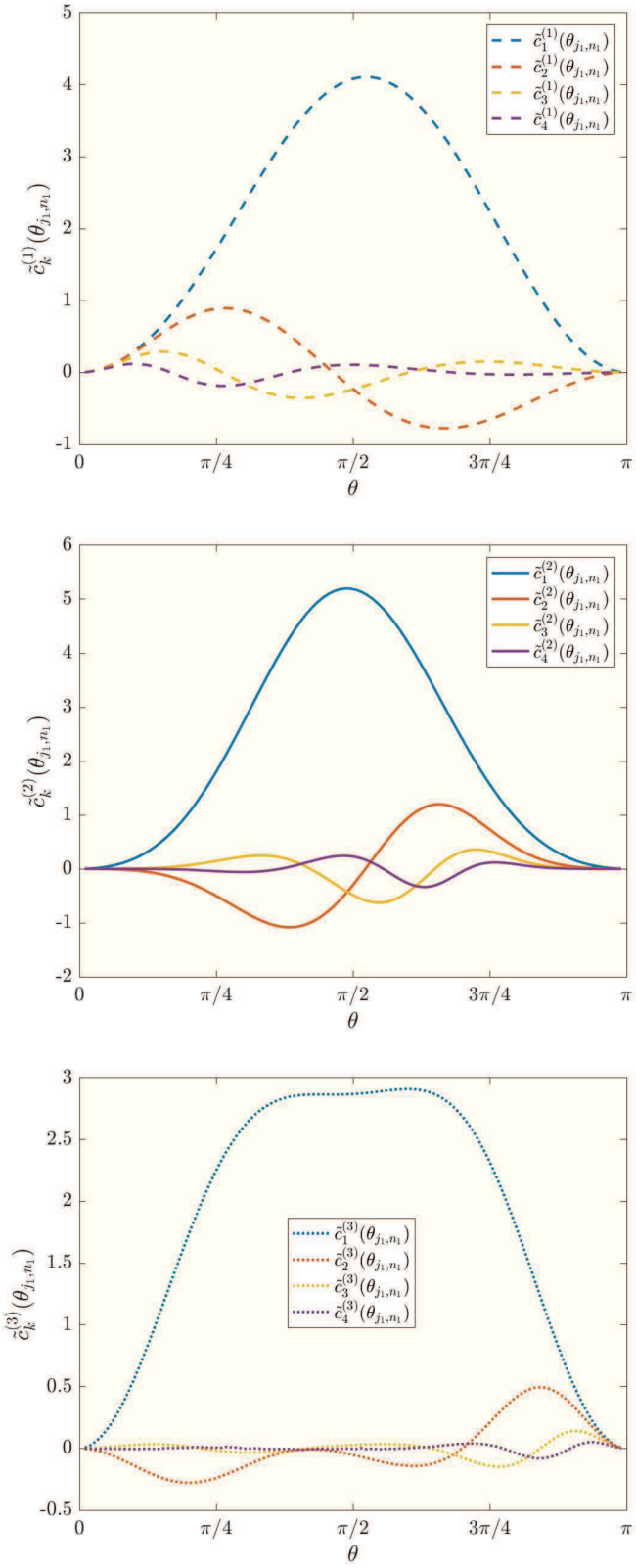


Figure V.3: Example 1: Computations made with $n_1 = 100$, $\alpha = 4$. From the top to the bottom panel the approximations $\tilde{c}_k^{(q)}(\theta_{j_1,n_1})$ for $\lambda^{(q)}(\mathbf{f})$, $q = 1, 2, 3$.

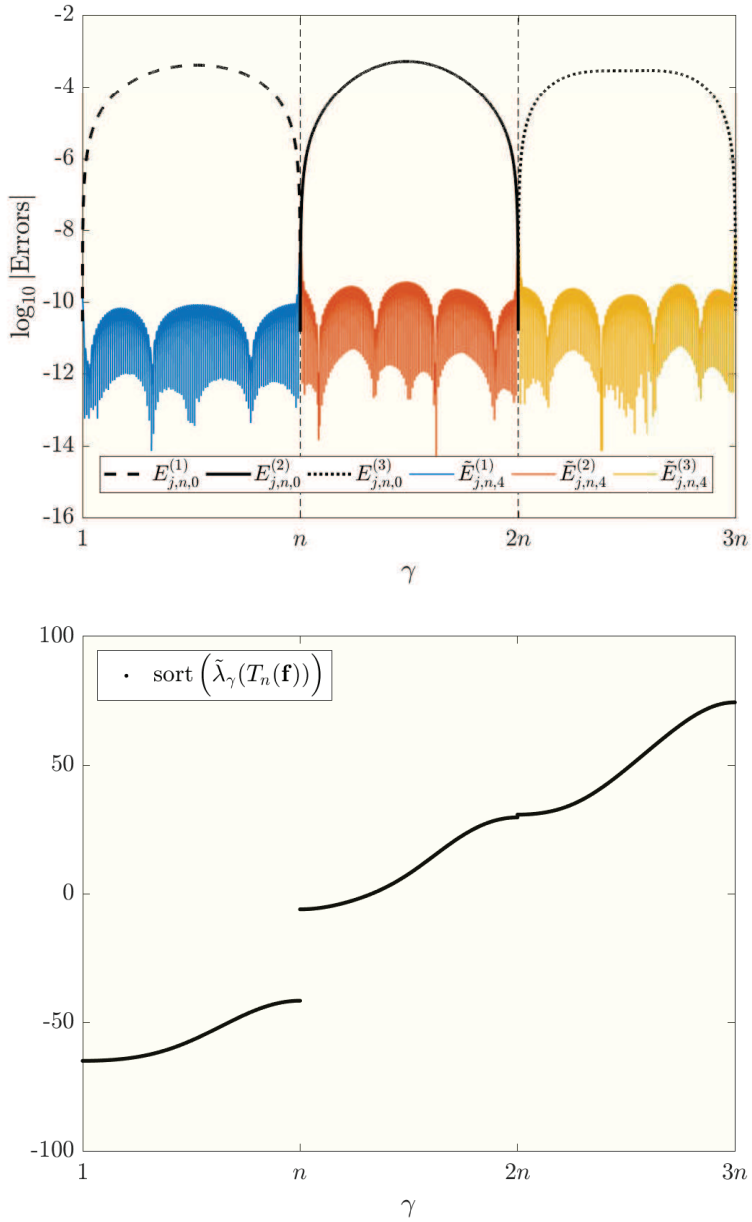


Figure V.4: Example 1: **Top:** Errors $\log_{10}|\tilde{E}^{(q)}_{j,n,\alpha}|$, with $\alpha = 4$, and errors $\log_{10}|E^{(q)}_{j,n,0}|$, $q = 1, 2, 3$, versus γ for $\gamma = 1, \ldots, 3n$. Computations made with $n_1 = 100$ and $n = 10000$. **Bottom:** Approximated eigenvalues $\tilde{\lambda}_\gamma(T_n(\mathbf{f}))$, sorted in non decreasing order. Computation made with the interpolation–extrapolation algorithm, with $\alpha = 4$, $n_1 = 100$ and $n = 10000$.

in the first Fourier coefficient which is defined as

$$\hat{\mathbf{f}}_0 = \begin{bmatrix} 12 & 2 & 0 \\ 2 & -55 & 2 \\ 0 & 2 & 10 \end{bmatrix}.$$

In this example we want to show that it is possible to give an approximation of the eigenvalues $\lambda_\gamma(T_n(\mathbf{f}))$, $n = 10000$, satisfying the *local condition*.

From the Figure V.6, where the graphs of the three eigenvalue functions are displayed, we

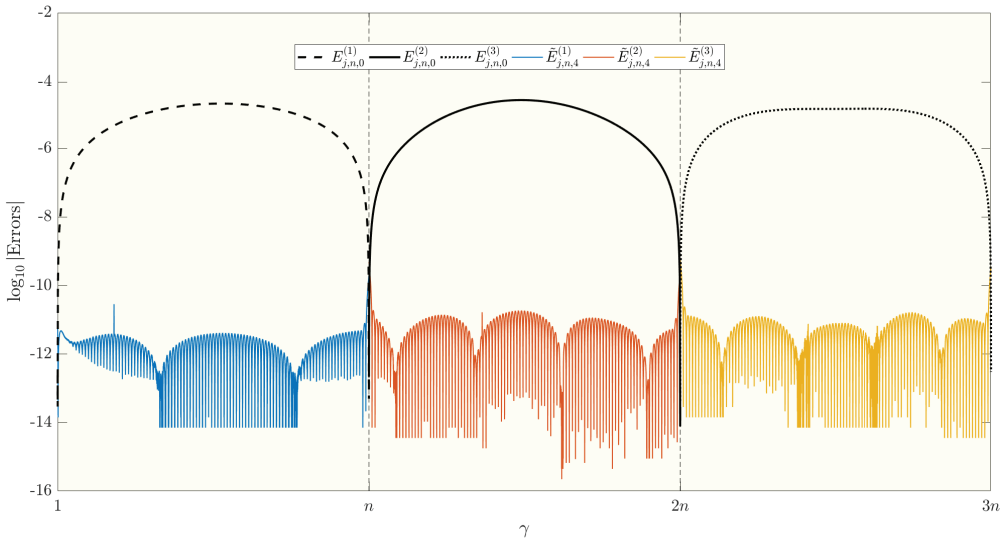


Figure V.5: Example 1: Errors $\log_{10}|\tilde{E}^{(q)}_{j,n,\alpha}|$, with $\alpha = 4$, and errors $\log_{10}|E^{(q)}_{j,n,0}|$, $q = 1, 2, 3$, versus γ for $\gamma = 1, \ldots, 3n$. Computations made with $n_1 = 100$ and $n = 2 \cdot 10^5$.

notice that

- $\lambda^{(1)}(\mathbf{f})$ is monotone non decreasing and its range does not intersect that of $\lambda^{(q)}(\mathbf{f})$, $q = 2, 3$. Hence, using the asymptotic expansion in (V.6), we expect that it is possible to give an approximation of the first n eigenvalues $\lambda_\gamma(T_n(\mathbf{f}))$, for $j = 1, \ldots, n$;
- $\lambda^{(3)}(\mathbf{f})$ is monotone non increasing and there exist $\hat{\theta}_1$, $\hat{\theta}_2 \in [0, \pi]$ such that, $\forall\ \theta \in [0, \hat{\theta}_1) \cup (\hat{\theta}_2, \pi]$,
$$\left(\lambda^{(3)}(\mathbf{f})\right)(\theta) \notin \text{Range}(\lambda^{(2)}(\mathbf{f})).$$
Hence, of the remaining $2n$ eigenvalues, we expect that it is possible to give a fast approximation just of those eigenvalues $\lambda_\gamma(T_n(\mathbf{f}))$ verifying *local condition*, that is those satisfying the relation below
$$\lambda_\gamma(T_n(\mathbf{f})) \in \left[\left(\lambda^{(3)}(\mathbf{f})\right)(\pi), \left(\lambda^{(3)}(\mathbf{f})\right)(\hat{\theta}_2)\right) \bigcup \left(\left(\lambda^{(3)}(\mathbf{f})\right)(\hat{\theta}_1), \left(\lambda^{(3)}(\mathbf{f})\right)(0)\right]. \tag{V.15}$$

We fix $\alpha = 4$, $n_1 = 100$ and we proceed to calculate the approximation of $c^{(q)}_k(\theta_{j_1,n_1}), k = 1, \ldots, \alpha$, as in the previous example. As expected, the graph of $\tilde{c}^{(1)}_k(\theta_{j_1,n_1})$, $k = 1, \ldots, 4$, shown in the top panel of Figure V.7, reveals that we can compute $\tilde{\lambda}_\gamma(T_n(\mathbf{f}))$, for $q = 1$ and $j = 1, \ldots, n$, using (V.6). In other words the first n eigenvalues of $T_n(\mathbf{f})$ can be computed using our matrix-less procedure.

For $q = 2$ no extrapolation procedure can be applied with $\tilde{c}^{(2)}_k(\theta_{j_1,n_1})$, $k = 1, \ldots, 4$, as we can see from the oscillating and irregular graph in the middle panel of Figure V.7. Concerning Figure V.8 the chaotic behavior of $\tilde{c}^{(2)}_k(\theta_{j_1,n_1})$, $k = 1, \ldots, 4$ corresponds to the rather large and oscillating errors $E^{(2)}_{j,n,0}$ and $\tilde{E}^{(2)}_{j,n,\alpha}$. On the other hand for $q = 3$ we can use the extrapolation procedure and the underlying asymptotic expansion with $\tilde{c}^{(3)}_k(\theta_{j_1,n_1}), k = 1, \ldots, 4$ for $\theta_{j_1,n_1} \in [0, \hat{\theta}_1) \cup (\hat{\theta}_2, \pi]$, $j_1 = 1, \ldots, n_1$.

As a consequence we compute the approximation of the first n eigenvalues $\lambda_\gamma(T_n(\mathbf{f}))$, for $\gamma = 1, \ldots, n$ and that of other $\hat{n}_1 + \hat{n}_2$, that verify (V.15). For simplicity, in the bottom panel of Figure V.8, we visualize them by using the non decreasing order instead of the computational one.
The good approximation of the $\hat{n}_1 + \hat{n}_2$ eigenvalues belonging to

$$\left[\left(\lambda^{(3)}(\mathbf{f})\right)(\pi), \left(\lambda^{(3)}(\mathbf{f})\right)(\hat{\theta}_2)\right) \bigcup \left(\left(\lambda^{(3)}(\mathbf{f})\right)(\hat{\theta}_1), \left(\lambda^{(3)}(\mathbf{f})\right)(0)\right]$$

is confirmed by the error $\tilde{E}^{(3)}_{j,n,\alpha}$ in the top panel of Figure V.8. In fact the error is quite high for $\gamma = 2n + \hat{n}_1 + 1, \ldots, 3n - \hat{n}_2$, but it becomes sufficiently small for $\gamma = 2n + 1, \ldots, 2n + \hat{n}_1$ and $\gamma = 3n - \hat{n}_2 + 1, \ldots, 3n$.

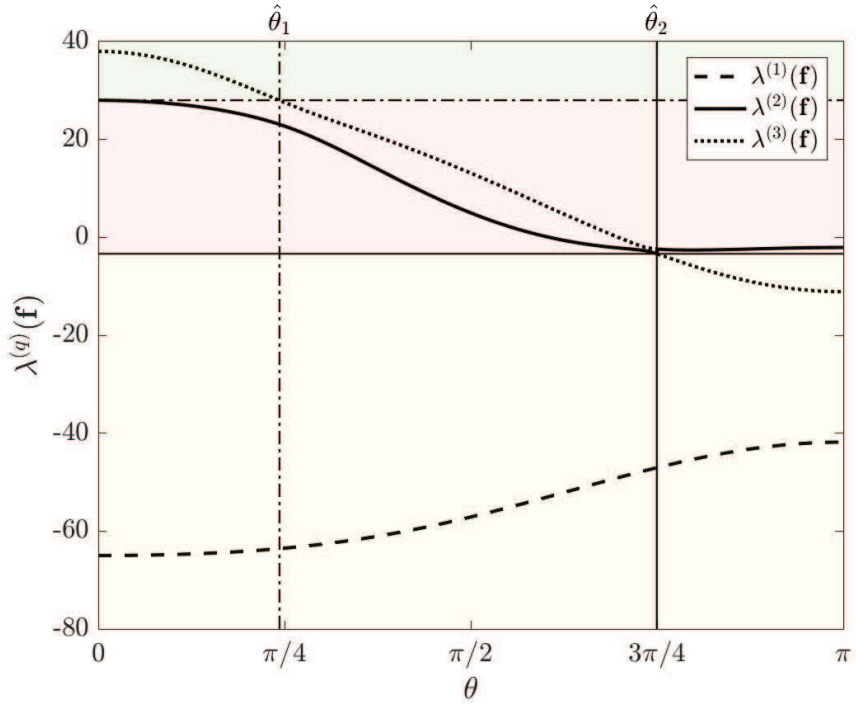


Figure V.6: Example 2: The three eigenvalue functions, $\lambda^{(q)}(\mathbf{f}), q = 1, 2, 3$.

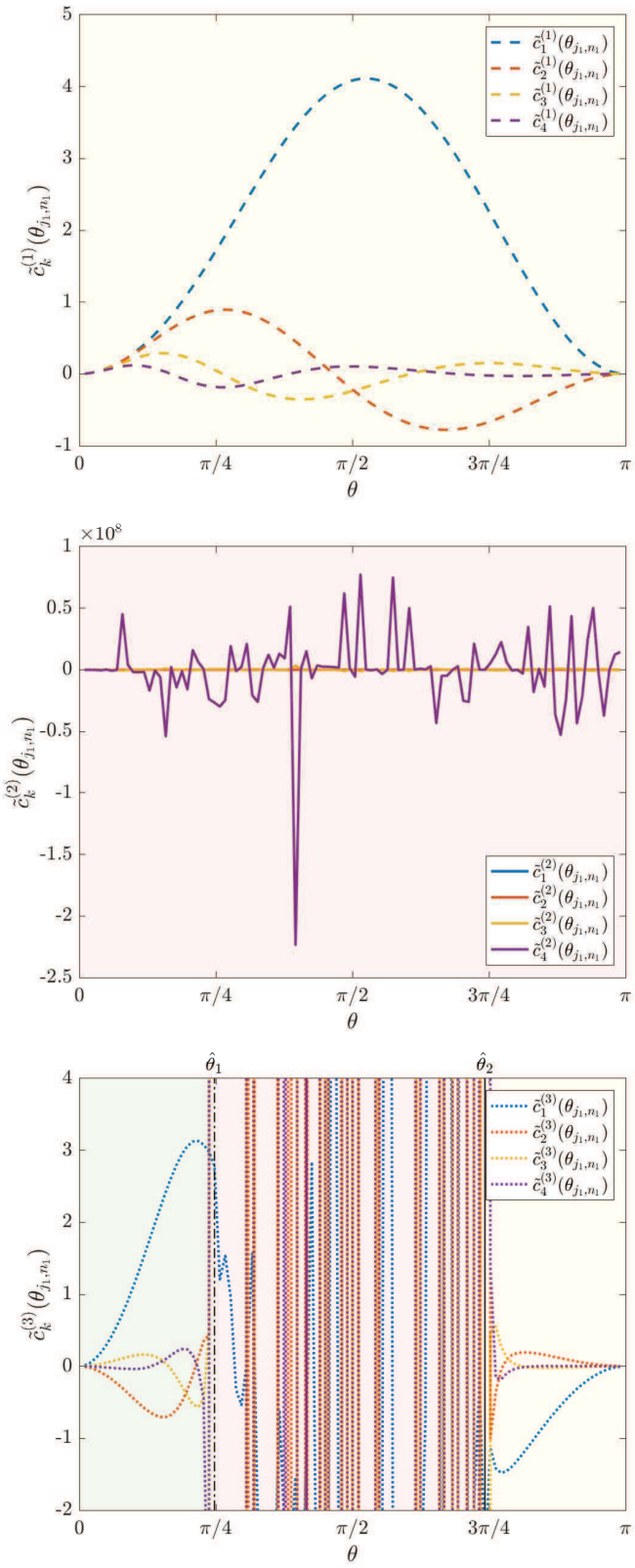


Figure V.7: Example 2: Computations made with $n_1 = 100$, $\alpha = 4$. From the top to the bottom panel the approximations $\tilde{c}_k^{(q)}(\theta_{j_1,n_1})$ for $\lambda^{(q)}(\mathbf{f})$, $q = 1, 2, 3$.

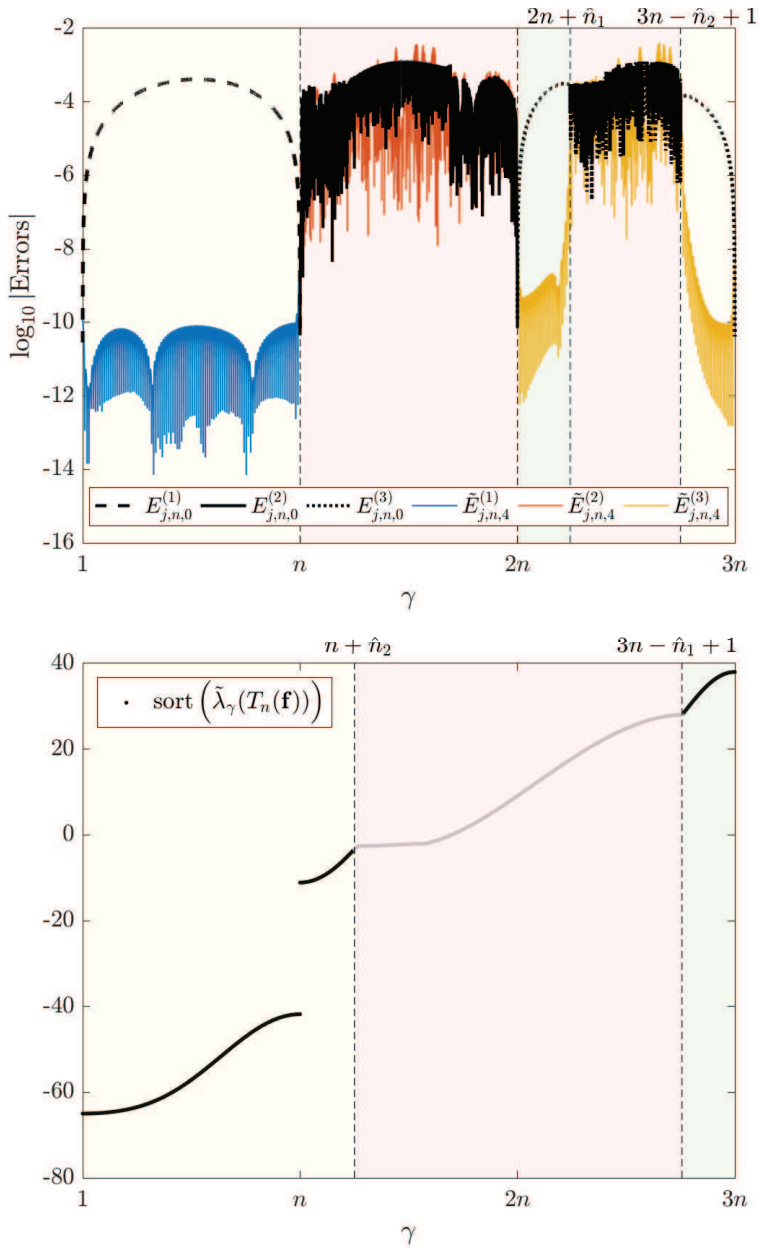


Figure V.8: Example 2: **Top:** Errors $\log_{10}|\tilde{E}^{(q)}_{j,n,\alpha}|$, with $\alpha = 4$, and errors $\log_{10}|E^{(q)}_{j,n,0}|$, $q = 1,2,3$, versus γ for $\gamma = 1,\dots,3n$. Computations made with $n_1 = 100$ and $n = 10000$. **Bottom:** Approximated eigenvalues $\tilde{\lambda}_\gamma(T_n(\mathbf{f}))$, sorted in non decreasing order, for $\gamma = 1,\dots,n$ and for γ such that $\lambda_\gamma(T_n(\mathbf{f}))$ verifies (V.15). Computation made with the interpolation–extrapolation algorithm, with $\alpha = 4$, $n_1 = 100$ and $n = 10000$.

V.3.3 Local condition: lack of the monotonicity

Example 3.

In this example we set the block size $s = 3$, and the eigenvalue functions $\lambda^{(q)}(\mathbf{f}), q = 1,2,3$, satisfy

$$\max_{\theta\in[0,\pi]} \lambda^{(1)}(\mathbf{f}) < \min_{\theta\in[0,\pi]} \lambda^{(2)}(\mathbf{f}),$$
$$\max_{\theta\in[0,\pi]} \lambda^{(2)}(\mathbf{f}) < \min_{\theta\in[0,\pi]} \lambda^{(3)}(\mathbf{f}).$$

See the Figure V.9 for the plot of $\lambda^{(q)}(\mathbf{f})$, $q = 1, 2, 3$.

The matrix $T_n(f) \in \mathbb{R}^{N\times N}$, $N = 3n$, shows a pentadiagonal block structure, and all the blocks belongs to $\mathbb{R}^{3\times 3}$, that is

$$T_n(\mathbf{f}) = \begin{bmatrix} \hat{\mathbf{f}}_0 & \hat{\mathbf{f}}_1^T & \hat{\mathbf{f}}_2^T & & & \\ \hat{\mathbf{f}}_1 & \ddots & \ddots & \ddots & & \\ \hat{\mathbf{f}}_2 & \ddots & \ddots & \ddots & \ddots & \\ & \ddots & \ddots & \ddots & \ddots & \hat{\mathbf{f}}_2^T \\ & & \ddots & \ddots & \ddots & \hat{\mathbf{f}}_1^T \\ & & & \hat{\mathbf{f}}_2 & \hat{\mathbf{f}}_1 & \hat{\mathbf{f}}_0 \end{bmatrix}, \hat{\mathbf{f}}_0 = \frac{1}{5}\begin{bmatrix} 16 & -12 & 5 \\ -12 & 34 & -10 \\ 5 & -10 & 100 \end{bmatrix},$$
$$\hat{\mathbf{f}}_1 = \frac{1}{10}\begin{bmatrix} -4 & 7 & 0 \\ 8 & -16 & 0 \\ 0 & 0 & -10 \end{bmatrix}, \hat{\mathbf{f}}_2 = \frac{1}{20}\begin{bmatrix} -12 & -12 & 0 \\ -16 & 12 & 1 \\ 0 & 2 & 0 \end{bmatrix}.$$

In analogy with the Example 2, we want to give an approximation of $\lambda_\gamma(T_n(\mathbf{f}))$, $n = 10000$, in case the *global condition* is not satisfied.
Although the intersection of the ranges of $\lambda^{(j)}(\mathbf{f})$ and $\lambda^{(k)}(\mathbf{f})$ is empty for every pair (j, k), $j \neq k$, $j, k \in \{1, 2, 3\}$, the assumption of monotonicity is violated either globally on $[0, \pi]$ or on a subinterval in $[0, \pi]$.
In detail:

- $\lambda^{(1)}(\mathbf{f})$, is fully non-monotone on $[0, \pi]$, hence we expect that no fast approximation can be given on the first n eigenvalues, $\lambda_\gamma(T_n(\mathbf{f}))$, for $\gamma = 1, \ldots, n$;
- $\lambda^{(3)}(\mathbf{f})$ is monotone non decreasing and its range does not intersect that of $\lambda^{(q)}(\mathbf{f})$, $q = 1, 2$. Hence we can provide an approximation, of the last n eigenvalues $\lambda_\gamma(T_n(\mathbf{f}))$ for $\gamma = 2n + 1, \ldots, 3n$, (analogously with what we did for treating the first n eigenvalues in Example 2);
- $\lambda^{(2)}(\mathbf{f})$ is non-monotone on a subinterval $[0, \hat{\theta}_1]$ in $[0, \pi]$ and monotone non decreasing on the remaining subinterval, $(\hat{\theta}_1, \pi]$. Hence we are able to efficiently compute also the eigenvalues that verify the following relation

$$\lambda_\gamma(T_n(\mathbf{f})) \in \left[\left(\lambda^{(2)}(\mathbf{f})\right)(\hat{\theta}_1), \left(\lambda^{(2)}(\mathbf{f})\right)(\pi)\right]. \tag{V.16}$$

We set $\alpha = 4$, $n_1 = 100$, for the computation and we proceed, as in the previous examples, to calculate first the approximation of $c_k^{(q)}(\theta_{j_1,n_1}), k = 1, \ldots, \alpha$.

In the top image of Figure V.10 we display the resulting chaotic graph of $\tilde{c}_k^{(1)}(\theta_{j_1,n_1}), k = 1, \ldots, 4$. The graph confirms that, for $q = 1$, the interpolation–extrapolation algorithm cannot be used and, consequently, the first n eigenvalues, $\lambda_\gamma(T_n(\mathbf{f}))$, $q = 1$, $j = 1, \ldots, n$, cannot be efficiently computed using (V.6): the latter is confirmed by the errors $\tilde{E}_{j,n,\alpha}^{(1)}$ and $E_{j,n,0}^{(1)}$, in Figure V.11.
The chaotic behaviour is also present in the values $\tilde{c}_k^{(2)}(\theta_{j_1,n_1}), k = 1, \ldots, 4$, see the middle panel of Figure V.10, in the subinterval $[0, \hat{\theta}_1]$ of $[0, \pi]$, that coincides with same subinterval where $\lambda^{(2)}(\mathbf{f})$ is non–monotone.
Hence, if we restrict to $[0, \hat{\theta}_1]$, the extrapolation procedure can be used again on $\tilde{c}_k^{(2)}(\theta_{j_1,n_1}), k = 1, \ldots, 4$, for $\theta_{j_1,n_1} \in (\hat{\theta}_1, \pi]$, $j_1 = 1, \ldots, n_1$. Consequently we obtain a good approximation of $\lambda_\gamma(T_n(\mathbf{f}))$, for $q = 2$, $j = \hat{j}, \ldots, n$. Notice that $\hat{j}$ is the first index in $\{1, \ldots, n\}$ such that $\frac{\hat{j}\pi}{n+1} \in (\hat{\theta}_1, \pi]$, that is we can compute the eigenvalues belonging to the interval reported in (V.16). This is reflected, in Figure V.11, in the gradual reduction of the errors $\tilde{E}_{j,n,\alpha}^{(2)}$ and $E_{j,n,0}^{(2)}$, for indices larger than $\hat{n}_1 = n + \hat{j}$.
Finally, the remaining n eigenvalues can be well reconstructed with a standard matrix-less procedure, using the values of $\tilde{c}_k^{(3)}(\theta_{j_1,n_1}), k = 1, \ldots, 4$, shown in the bottom panel of Figure V.10. The errors related to latter approximation, $\tilde{E}_{j,n,\alpha}^{(3)}$, are shown in Figure V.11.
In total, $3n - \hat{j} + 1$ eigenvalues of $T_n(\mathbf{f})$ can be computed and plotted (in non decreasing order) in the bottom of the Figure V.11.

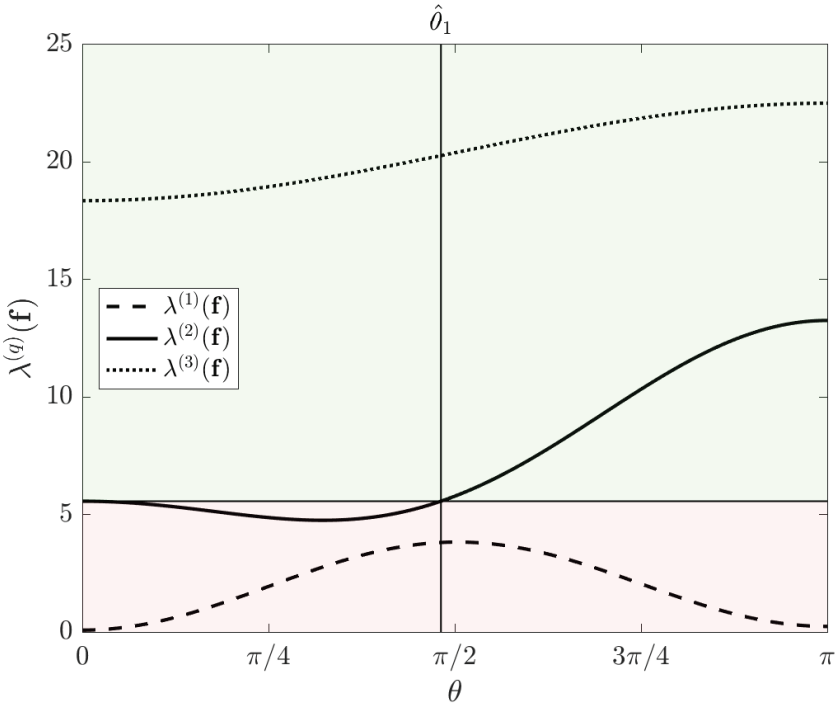


Figure V.9: Example 3: The three eigenvalue functions, $\lambda^{(q)}(\mathbf{f}), q = 1, 2, 3$.

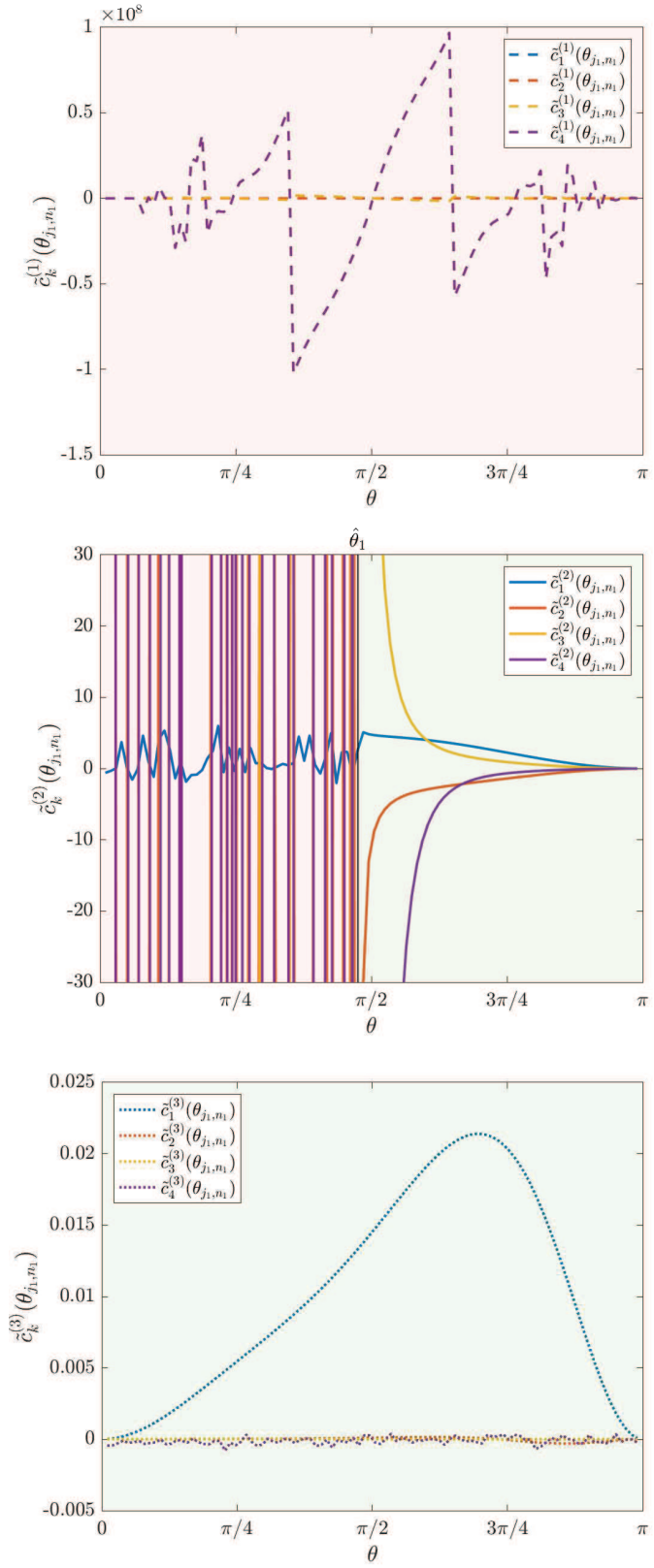


Figure V.10: Example 3: Computations made with $n_1 = 100$, $\alpha = 4$. From the top to the bottom panel the approximations $\tilde{c}_k(\theta_{j_1,n_1})$ for $\lambda^{(q)}(\mathbf{f})$, $q = 1, 2, 3$.

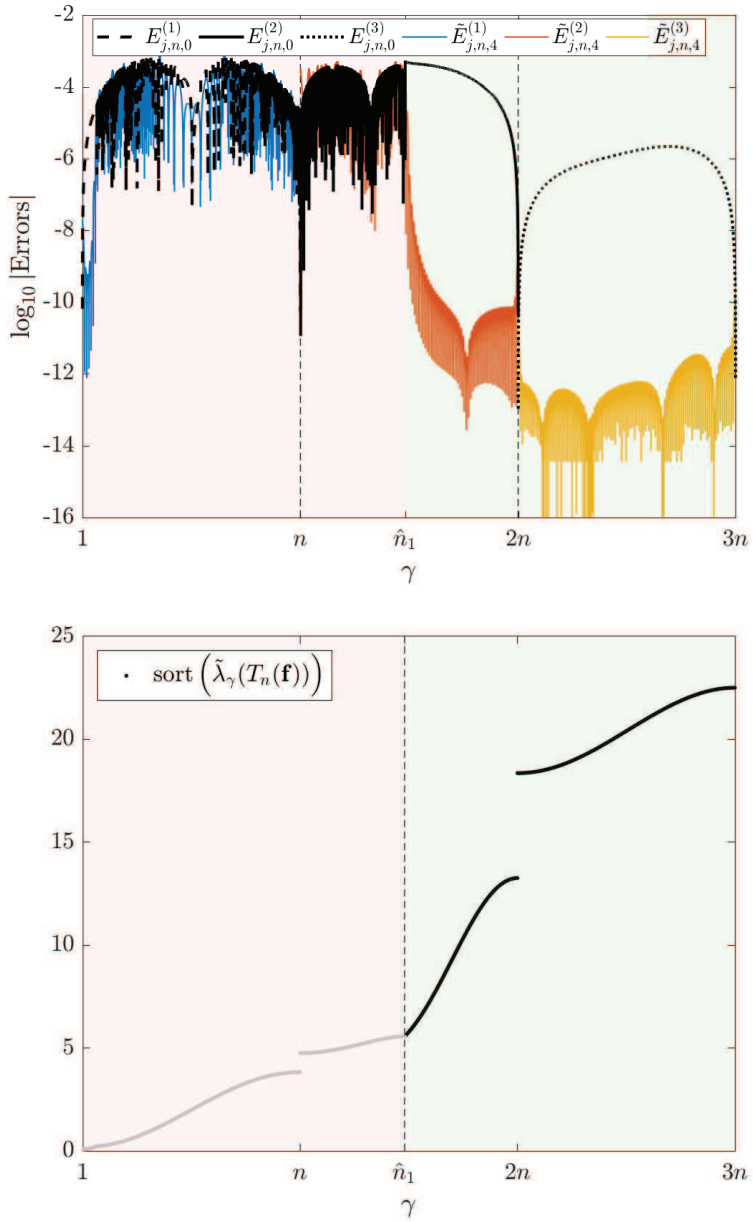


Figure V.11: Example 3: **Top:** Errors $\log_{10}|\tilde{E}^{(q)}_{j,n,\alpha}|$, with $\alpha = 4$, and errors $\log_{10}|E^{(q)}_{j,n,0}|$, $q = 1, 2, 3$, versus γ for $\gamma = 1, \ldots, 3n$. Computations made with $n_1 = 100$ and $n = 10000$. **Bottom:** Approximated eigenvalues $\tilde{\lambda}_\gamma(T_n(\mathbf{f}))$, sorted in non decreasing order, for $\gamma = 2n+1, \ldots, 3n$ and for γ such that $\lambda_\gamma(T_n(\mathbf{f}))$ verifies (V.16). Computation made with the interpolation–extrapolation algorithm, with $\alpha = 4$, $n_1 = 100$ and $n = 10000$.

V.3.4 Local condition: reduction from block to scalar.

Example 4.

In this example we consider three trigonometric polynomials,

$$p^{(1)}(\theta) = 2 - 2\cos(\theta),$$
$$p^{(2)}(\theta) = 7 - 2\cos(2\theta),$$
$$p^{(3)}(\theta) = 16 - 8\cos(\theta) + 2\cos(2\theta) = 10 + (p^{(1)}(\theta))^2,$$

with the aim of approximating the eigenvalues of a block banded Toeplitz matrix, with a matrix-valued generating function $\mathbf{f}(\theta)$, such that $\lambda^{(q)}(\mathbf{f}) = p^{(q)}$ for $q = 1, 2, 3$. We choose $s = 3$ but obviously the following procedure holds for any $s \in \mathbb{Z}_+$ and for any chosen s trigonometric polynomials, $p^{(1)}(\theta), p^{(2)}(\theta), \ldots, p^{(s)}(\theta)$, such that

$$\max_{\theta\in[0,\pi]} p^{(q)}(\theta) < \min_{\theta\in[0,\pi]} p^{(q+1)}(\theta),$$

for $q = 1, \ldots, s-1$. We can define

$$\mathbf{f}(\theta) = Q_3 \begin{bmatrix} p^{(1)}(\theta) & 0 & 0 \\ 0 & p^{(2)}(\theta) & 0 \\ 0 & 0 & p^{(3)}(\theta) \end{bmatrix} Q_3^{\mathrm{T}},$$

where Q_3 is any orthogonal matrix in $\mathbb{R}^{3\times 3}$. For the current example we choose

$$Q_3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos(\pi/3) & -\sin(\pi/3) \\ 0 & \sin(\pi/3) & \cos(\pi/3) \end{bmatrix} = \frac{1}{2}\begin{bmatrix} 2 & 0 & 0 \\ 0 & 1 & -\sqrt{3} \\ 0 & \sqrt{3} & 1 \end{bmatrix}.$$

Now we define the Fourier coefficients of $\mathbf{f}(\theta)$, that is

$$\hat{\mathbf{f}}_k = Q_3 \begin{bmatrix} \hat{p}_k^{(1)} & 0 & 0 \\ 0 & \hat{p}_k^{(2)} & 0 \\ 0 & 0 & \hat{p}_k^{(3)} \end{bmatrix} Q_3^{\mathrm{T}} = Q_3 \hat{D}_k Q_3^T, \tag{V.17}$$

where $\hat{p}_k^{(q)}$ is the kth Fourier coefficient of the eigenvalue function $p^{(q)}(\theta)$, and $k = -m, \ldots, m$, where $m = \max_{q=1,\ldots,s} \deg(p^{(q)}(\theta)$. In our example $m = 2$, for $p^{(2)}(\theta)$ and $p^{(3)}(\theta)$ and $m = 1$ for $p^{(1)}(\theta)$. Each $p^{(q)}(\theta)$ is a real cosine trigonometric polynomial (RCTP), so $\mathbf{f}(\theta)$ is a symmetric matrix-valued function with Fourier coefficients

$$\hat{\mathbf{f}}_0 = \frac{1}{4}\begin{bmatrix} 8 & 0 & 0 \\ 0 & 55 & -9\sqrt{3} \\ 0 & -9\sqrt{3} & 37 \end{bmatrix}, \quad \hat{\mathbf{f}}_1 = \begin{bmatrix} -1 & 0 & 0 \\ 0 & -3 & \sqrt{3} \\ 0 & \sqrt{3} & -1 \end{bmatrix}, \quad \hat{\mathbf{f}}_2 = \frac{1}{2}\begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & -\sqrt{3} \\ 0 & -\sqrt{3} & -1 \end{bmatrix},$$

where $\hat{\mathbf{f}}_{-k} = \hat{\mathbf{f}}_k^T = \hat{\mathbf{f}}_k$, $k = 0, 1, 2$.
The resulting block banded Toeplitz matrix is the following matrix

$$T_n(\mathbf{f}) = \begin{bmatrix} \hat{\mathbf{f}}_0 & \hat{\mathbf{f}}_{-1} & \hat{\mathbf{f}}_{-2} & & \\ \hat{\mathbf{f}}_1 & \ddots & \ddots & \ddots & \\ \hat{\mathbf{f}}_2 & \ddots & \ddots & \ddots & \hat{\mathbf{f}}_{-2} \\ & \ddots & \ddots & \ddots & \hat{\mathbf{f}}_{-1} \\ & & \hat{\mathbf{f}}_2 & \hat{\mathbf{f}}_1 & \hat{\mathbf{f}}_0 \end{bmatrix},$$

with symbol

$$\mathbf{f}(\theta) = \hat{\mathbf{f}}_0 + \sum_{k=1}^{2} \left(\hat{\mathbf{f}}_k e^{\iota k\theta} + \hat{\mathbf{f}}_{-k} e^{-\iota k\theta} \right) = \hat{\mathbf{f}}_0 + 2\hat{\mathbf{f}}_1 \cos(\theta) + 2\hat{\mathbf{f}}_2 \cos(2\theta).$$

We want to approximate the eigenvalues of $T_n(\mathbf{f})$, where $\mathbf{f}(\theta)$ is constructed from $p^{(q)}(\theta)$, $q = 1, 2, 3$. For the graph of the chosen polynomials see the top left panel of Figure V.12.
Due to the special structure of all $\hat{\mathbf{f}}_k$, see (V.17), we have

$$T_n(\mathbf{f}) = I_n \otimes Q_3 \begin{bmatrix} \hat{D}_0 & \hat{D}_{-1} & \hat{D}_{-2} & & \\ \hat{D}_1 & \ddots & \ddots & \ddots & \\ \hat{D}_2 & \ddots & \ddots & \ddots & \hat{D}_{-2} \\ & \ddots & \ddots & \ddots & \hat{D}_{-1} \\ & & \hat{D}_2 & \hat{D}_1 & \hat{D}_0 \end{bmatrix} I_n \otimes Q_3^T.$$

Therefore $T_n(\mathbf{f})$ is similar to the matrix

$$\begin{bmatrix} T_n(p^{(1)}(\theta)) & 0 & 0 \\ 0 & T_n(p^{(2)}(\theta)) & 0 \\ 0 & 0 & T_n(p^{(3)}(\theta)) \end{bmatrix},$$

and finally it is trivial to see that the block case, in this setting, is reduced to 3 different scalar problems, which can be treated separately.
Differently from previous examples, here the analytical expressions of the eigenvalue functions of $\mathbf{f}(\theta)$ are known, since they coincide, by construction, with $p^{(q)}(\theta)$, $q = 1, 2, 3$. So we will describe the spectrum of $T_n(\mathbf{f})$, approximating or calculating exactly the $3n$ eigenvalues, treating the 3 different scalar problems separately.

For the first n eigenvalues it is known that they can be calculated exactly, sampling $p^{(1)}$ with grid $\theta_{j,n} = \frac{j\pi}{n+1}$, $j = 1, \ldots, n$. Analogously, the n eigenvalues can be found exactly by sampling $p^{(2)}$ on a special grid defined in [63]. For the last n eigenvalues, the grid that gives exact eigenvalues is not known, but $p^{(3)}$ is monotone non decreasing and consequently we can use an asymptotic expansion in the scalar case.
We set the parameters as in previous cases: $n_1 = 100$ and $n = 10000$.
In the top right panel of Figure V.12 we report the expansion errors $E^{(q)}_{j_1,n_1,0}$, calculated using grid $\theta_{j_1,n_1} = \frac{\pi j_1}{n_1+1}$, $j_1 = 1, \ldots, n_1$, $q = 1, 2, 3$.

Obviously in the first region of the graph (green area) the error is zero, since the first n_1 eigenvalues are exactly given, sampling $p^{(1)}$ on the standard θ_{j_1,n_1} grid.

In the yellow area we see the result of the direct calculation of

$$\lambda_\gamma(T_{n_1}(\mathbf{f})) - \lambda^{(3)}(\mathbf{f}(\theta_{j_1,n_1})),$$

for $j_1 = 1, \ldots, n_1$, $q = 3$, as we are using the asymptotic expansion with $\alpha = 0$.
The green area, containing the errors related to $p^{(2)}(\theta)$, is obviously chaotic since $p^{(2)}(\theta)$ is non-monotone.

Following the notation and the analysis in [63], since $n_1 = 100$ and $p^{(2)} = 7 - 2\cos(2\theta)$, we have two changes of monotonicity which we collect in the parameter ω. As a consequence, in

accordance with the study in [63], we choose

$$\omega = 2, \quad \beta = \mathrm{mod}(n_1, \omega) = 0, \quad n_\omega = (n_1 - \beta)/\omega = 50,$$
$$\theta^{(1)}_{n_\omega} = \frac{j\pi}{n_\omega + 1}, \quad j = 1, \dots n_\omega,$$
$$\theta^{(2)}_{n_\omega+1} = \frac{j\pi}{n_\omega + 2}, \quad j = 1, \dots n_\omega + 1.$$

To map the two grids above to match the given symbol $\mathbf{f}(\theta)$ we construct θ_{n_1} by

$$\theta_{n_1} = \left\{ \frac{1}{2}\theta^{(1)}_{n_\omega}, \frac{1}{2}\theta^{(2)}_{n_\omega+1} + \frac{\pi}{2} \right\}.$$

A more general formula to match grids $\theta^{(1)}_{n_\omega}$ and $\theta^{(2)}_{n_\omega+1}$ to be evaluated on the standard symbol is

$$\theta_n = \frac{1}{\omega} \left\{ \bigcup_{r_1=1}^{\omega-\beta} \left(\theta^{(1)}_{n_\omega} + (r_1 - 1)\pi \right), \bigcup_{r_2=1}^{\beta} \left(\theta^{(2)}_{n_\omega+1} + (r_2 - 1)\pi + (\omega - \beta)\pi \right) \right\}. \tag{V.18}$$

In the left bottom panel of Figure V.12 we report the global expansion errors $E^{(q)}_{j_1,n_1,0}$, calculated using the grid described above. In this way the region where the error is 0 is the second (red area), since the eigenvalues are calculated exactly, by sampling $p^{(2)}(\theta)$. Furthermore, in the green and in the yellow areas we see the result of the direct calculation of

$$\lambda_\gamma(T_{n_1}(\mathbf{f})) - \lambda^{(q)}(\mathbf{f}(\theta_{j_1,n_1})),$$

for $j_1 = 1, \dots, n_1$, $q = 1, 3$, as we are using asymptotic expansion with $\alpha = 0$.
Hence, the first n eigenvalues of $T_n(\mathbf{f})$ can be calculated exactly sampling $p^{(1)}$ with grid $\theta_{j,n} = \frac{j\pi}{n+1}$, $j = 1, \dots, n$ and n exact eigenvalues can be found sampling $p^{(2)}$ on grid (V.18). For the computation of the last n eigenvalues, we use the matrix-less procedure in the scalar setting, passing through the approximation of $c^{(3)}_k(\theta_{j_1,n_1}), k = 1, \dots, \alpha$, for $\alpha = 4$, see the bottom right panel of Figure V.12.

For $\alpha = 4$ we ignore the first two evaluations of $c^{(3)}_4$ at the initial points $\theta_{1,n}$ and $\theta_{2,n}$, because their values behave in an erratic way. This problem has been emphasized in [7] and it is due to the fact that the first and second derivative of $p^{(3)}(\theta)$ at $\theta = 0$ vanish simultaneously. However, we have to make two observations for clarifying the situation

- The present pathology is not a counterexample to the asymptotic expansion (V.6) since we take θ fixed and all the pairs j, n such that $\theta_{j,n} = \theta$: in the current case and in that considered in [7] in the scalar-valued setting, we have j fixed and n grows so that the point θ is not well defined.

- There are simple ways to overcome the problem and then to compute reliable evaluations of $c^{(3)}_4$ at those bad points $\theta_{1,n}$ and $\theta_{2,n}$. One of them is described in [57] and consists in choosing a sufficiently large $\alpha > 4$ and in computing $c^{(3)}_k$, $k = 1, 2, 3, 4$. Using this trick, the $c^{(3)}_4$ at the initial points $\theta_{1,n}$ and $\theta_{2,n}$ have the expected behavior. In addition we stress the fact that this behavior has little impact on the numerically computed solution. Assuming double precision computations, the contribution to the error deriving from $c^{(3)}_4(\theta_{j,n})h^4$ will be numerically negligible, even for moderate n. Further discussions on the topic are presented in [57].

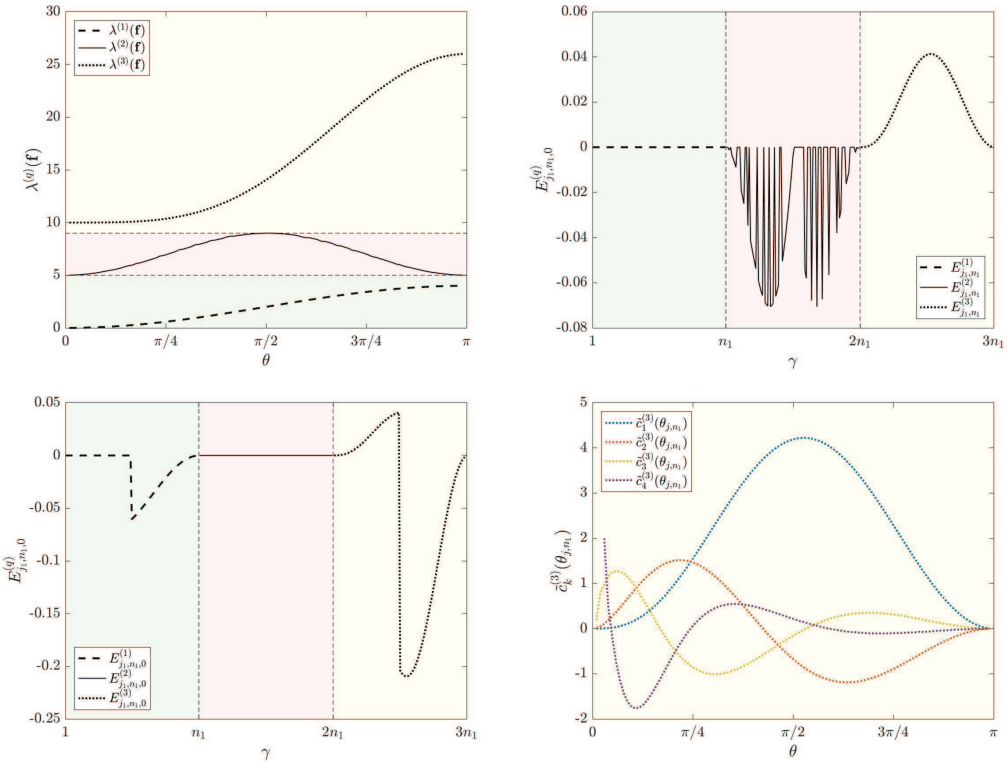


Figure V.12: Example 4: **Top Left:** Constructed eigenvalue functions. **Top Right:** Errors of the three eigenvalue functions, presented on global indices γ, when using grid $\theta_{j_1,n_1} = \frac{\pi j_1}{n_1+1}$, $j_1 = 1, \ldots, n_1$, $q = 1, 2, 3$. **Bottom Left:** Errors of the three eigenvalue functions, when using grid defined in (V.18). **Bottom Right:** Error expansion for the third eigenvalue function. Computations made with $n_1 = 100$ and $\alpha = 4$.

V.3.5 Exact formulae for $\mathbb{Q}_p$ Lagrangian FEM

Example 5.

Consider the $\mathbb{Q}_p$ Lagrangian Finite Element approximation, of the second order elliptic differential problem

$$\begin{cases} -\Delta u + \beta \cdot \nabla u + \gamma u = f, & \text{in } \Omega = (0,1)^k, \\ u = 0, & \text{on } \partial\Omega, \end{cases} \tag{V.19}$$

in one dimension with $\beta = \gamma = 0$, and $f \in L^2(\Omega)$. The resulting stiffness matrix is $A_n^{(p)} = nK_n^{(p)}$, where $K_n^{(p)}$ is a $(pn-1) \times (pn-1)$ block matrix. The construction of the matrix and the symbol is given in [78]. The $p \times p$ matrix-valued symbol of $K_n^{(p)}$ is

$$\mathbf{f}(\theta) = \hat{\mathbf{f}}_0 + \hat{\mathbf{f}}_1 e^{\iota\theta} + \hat{\mathbf{f}}_1^{\mathrm{T}} e^{-\iota\theta}.$$

We have

$$K_n^{(p)} = T_n(\mathbf{f})_-,$$

where the subscript $-$ denotes that the last row and column of $T_n(\mathbf{f})$ are removed. This is because of the homogeneous boundary conditions. For detailed expressions of $\hat{\mathbf{f}}_0$ and $\hat{\mathbf{f}}_1$ in the particular case $p = 2, 3, 4$, see the Section VI.5 of the **Chapter VI**.

In Table V.1, we list seven examples of uniform grids, with varying n. The general notation for a grid, where the type is defined by context, is $\theta_{j,n}$, where n is the number of grid points, and j is the indices $j = 1, \ldots, n$. The grid fineness parameter h, for the respective grids, is also presented in Table V.1. The names of the different grids are chosen in view of their relations with the τ-algebras [21] (see specifically equations (19), (22), and (23) therein).

Table V.1: Seven examples of uniform grids. Typically the τ_n-grid is the default choice, unless other grids provides more accurate, or even exact, eigenvalues when sampling the symbol.

Name	Grid	j	h	Description
τ_n	$j\pi/(n+1)$	$1, \ldots, n$	$1/(n+1)$	$\tau_n(0,0)$
τ_{n-1}	$j\pi/n$	$1, \ldots, n-1$	$1/n$	$\tau_{n-1}(0,0)$
τ_{n-2}	$j\pi/(n-1)$	$1, \ldots, n-2$	$1/(n-1)$	$\tau_{n-2}(0,0)$
τ_{n-1}^{0}	$(j-1)\pi/n$	$1, \ldots, n$	$1/n$	$\tau_n(1,1) = 0 \cup \tau_{n-1}(0,0)$
τ_{n-1}^{π}	$j\pi/n$	$1, \ldots, n$	$1/n$	$\tau_n(-1,-1) = \tau_{n-1}(0,0) \cup \pi$
$\tau_{n-2}^{0,\pi}$	$(j-1)\pi/(n-1)$	$1, \ldots, n$	$1/(n-1)$	$0 \cup \tau_{n-2}(0,0) \cup \pi$
$\tau_{n-1}^{0,\pi}$	$(j-1)\pi/n$	$1, \ldots, n+1$	$1/n$	$0 \cup \tau_{n-1}(0,0) \cup \pi$

In Example 1 of [78] the case $p = 2$ is considered, and explicit formulas for the two eigenvalue functions are given, with their notation,

$$\lambda_1(\mathbf{f}_2(\theta)) = 5 + \frac{1}{3}\cos(\theta) + \frac{1}{3}\sqrt{129 + 126\cos(\theta) + \cos^2(\theta)},$$
$$\lambda_2(\mathbf{f}_2(\theta)) = 5 + \frac{1}{3}\cos(\theta) - \frac{1}{3}\sqrt{129 + 126\cos(\theta) + \cos^2(\theta)}.$$

Here we present the two grids used to sample the two eigenvalue functions in order to attain exact eigenvalues,

$$\lambda_1(\mathbf{f}_2(\theta^{(1)}_{j_1,n-1})), \quad \theta^{(1)}_{j_1,n-1} = \frac{j_1\pi}{n}, \quad j_1 = 1, \ldots, n-1,$$
$$\lambda_2(\mathbf{f}_2(\theta^{(2)}_{j_2,n})), \quad \theta^{(2)}_{j_2,n} = \frac{j_2\pi}{n}, \quad j_2 = 1, \ldots, n.$$

With the notation in Table V.1, we use the grid τ_{n-1} for the first eigenvalue function, and grid τ_{n-1}^{π} for the second. Since for $p > 2$ the analytical expression of the eigenvalue functions can not be computed easily, the following four steps algorithm can be used to obtain the exact eigenvalues for any p.

Algorithm 3.

1. *Sample the matrix-valued symbol* $\mathbf{f}(\theta)$ *with the grid* $\tau_{n-1}^{0,\pi}$

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

Each sampling gives a matrix of size $p \times p$. *Use an eigensolver to get the* p *eigenvalues of the sampling, sorted in non decreasing order. This results in a total of* $p(n+1)$ *values:* $j_q = 1, \ldots, n+1$ *for all* p *eigenvalue functions, so we have to discard* $p+1$ *of samplings, since the total number of eigenvalues of the matrix* $K_n^{(p)}$ *is* $pn - 1$.

2. *For the eigenvalue function* $\lambda^{(1)}(\mathbf{f})$ *choose samplings with index* $j_1 = 2, \ldots, n$. *This corresponds to the choice of the grid* τ_{n-1}.

3. *For eigenvalue functions $\lambda^{(q)}(\mathbf{f})$, where q is even, choose samplings with index $j_q = 2, \ldots, n+1$. This corresponds to the choice of the grid τ^{π}_{n-1}.*

4. *For eigenvalue functions $\lambda^{(q)}(\mathbf{f})$, where q is odd, choose samplings with index $j_q = 1, \ldots, n$. This corresponds to the choice of the grid τ^{0}_{n-1}.*

The mass matrix, of the system (V.19) (that is, $\gamma = 1$), is $B_n^{(p)} = n^{-1} M_n^{(p)}$, where $M_n^{(p)} = T_n(\mathbf{g})_-$ is the scaled mass matrix.
The $p \times p$ matrix-valued symbol of $M_n^{(p)}$ is given by

$$\mathbf{g}(\theta) = \hat{\mathbf{g}}_0 + \hat{\mathbf{g}}_1 e^{\iota\theta} + \hat{\mathbf{g}}_1^{\mathrm{T}} e^{-\iota\theta}.$$

For detailed expressions of $\hat{\mathbf{g}}_0$ and $\hat{\mathbf{g}}_1$ in the particular case $p = 2, 3, 4$, see the Section VI.5 of the **Chapter VI**. The algorithm for writing the exact eigenvalues of $M_n^{(p)}$, for p even, is the same as the one described for $K_n^{(p)}$ above, just replacing $\mathbf{f}(\theta)$ with $\mathbf{g}(\theta)$. However, for $p \geq 3$ odd, we have a slight modification:

If $(p+1)/2$ is odd, that is $p = 5, 9, \ldots$, define $\hat{p} = p$. If $(p+1)/2$ is even, that is $p = 3, 7, \ldots$, define $\hat{p} = p-2$. In summary, to obtaining the exact eigenvalues of $M_n^{(p)}$, the algorithm becomes:

Algorithm 4.

- *Do steps 1. and 2. of Algorithm 3, just replacing $\mathbf{f}(\theta)$ with $\mathbf{g}(\theta)$.*
- *For $q = 2, \ldots, (\hat{p}+1)/2$,*
  - *For $\lambda^{(q)}(\mathbf{g})$, q is even, choose samplings with index $j_q = 1, \ldots, n$. This corresponds to the grid τ^{0}_{n-1}.*
  - *For $\lambda^{(q)}(\mathbf{g})$, q is odd, choose samplings with index $j_q = 2, \ldots, n+1$. This corresponds to the grid τ^{π}_{n-1}.*
- *Continue with steps 3. and 4. of Algorithm 3, for $q = (\hat{p}+1)/2 + 1, \ldots, p$, just replacing $\mathbf{f}(\theta)$ with $\mathbf{g}(\theta)$.*

In Figure V.13 we present the appropriate grids, defined in Table V.1, for the exact eigenvalues of $K_n^{(p)}$ and $M_n^{(p)}$ with $n = 6$ and $p = 5$.

By the use of high precision arithmetic computations we have found the following "exceptions" to the above procedures. Indeed testing the algorithms for $\mathbf{f}$ and $\mathbf{g}$, for $p = 1, \ldots, 20$, with 64 ($\mathcal{O}(\texttt{eps}) = 10^{-19}$), 128 ($\mathcal{O}(\texttt{eps}) = 10^{-39}$), and 256 ($\mathcal{O}(\texttt{eps}) = 10^{-77}$) bit precision, we have noticed that the grids which provide the exact values of eigenvalues are "switched " for some p and q. Precisely

- for $\mathbf{f}$ the exchange has to be performed for $p = 14, 15, 18, 20$:

$p = 14$: for $q = 9, 10$ choose the samplings of $\lambda^{(q)}(\mathbf{f})$ corresponding to the indices

$$j = 2 - \mathrm{mod}(q, 2), \ldots, n + \mathrm{mod}(q, 2);$$

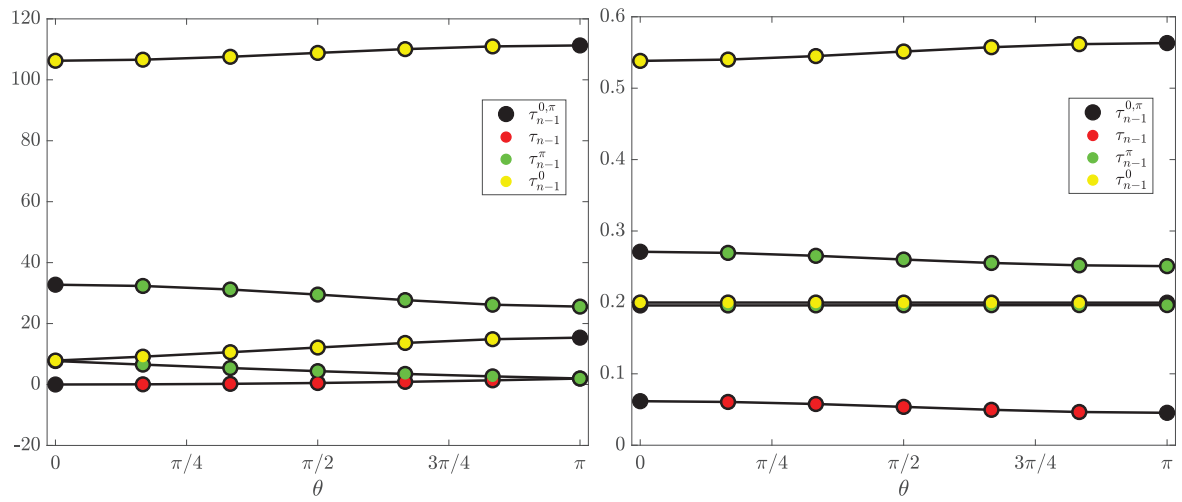


Figure V.13: Example 5: Grids for the exact eigenvalues of $K_n^{(p)}$ and $M_n^{(p)}$, with $n = 6$ and $p = 5$. **Left:** Grids chosen for each eigenvalue functions of $\mathbf{f}(\theta)$, for $q = 1, \ldots, 5$, according to Algorithm 3. **Right:** Grids chosen for each eigenvalue function of $\mathbf{g}(\theta)$, for $q = 1, \ldots, 5$, according to Algorithm 4.

$p = 15$: for $q = 10, 11$ choose the samplings of $\lambda^{(q)}(\mathbf{f})$ corresponding to the indices

$$j = 1 + \mathrm{mod}(q, 2), \ldots, n + 1 - \mathrm{mod}(q, 2);$$

$p = 18$: for $q = 12, 13$ choose the samplings of $\lambda^{(q)}(\mathbf{f})$ corresponding to the indices

$$j = 1 + \mathrm{mod}(q, 2), \ldots, n + 1 - \mathrm{mod}(q, 2);$$

$p = 20$: for $q = 13, 14$ choose the samplings of $\lambda^{(q)}(\mathbf{f})$ corresponding to the indices

$$j = 2 - \mathrm{mod}(q, 2), \ldots, n + \mathrm{mod}(q, 2);$$

– for $\mathbf{g}$ the exchange has to be performed for $p = 13, 14, 15, 19, 20$:

$p = 13$: for $q = 2, 3$ choose the samplings of $\lambda^{(q)}(\mathbf{f})$ corresponding to the indices

$$j = 1 + \mathrm{mod}(q, 2), \ldots, n + 1 - \mathrm{mod}(q, 2);$$

$p = 14$: for $q = 2, 3$ choose the samplings of $\lambda^{(q)}(\mathbf{g})$ corresponding to the indices

$$j = 1 + \mathrm{mod}(q, 2), \ldots, n + \mathrm{mod}(q, 2);$$

$p = 15$: for $q = 4, 5$ choose the samplings of $\lambda^{(q)}(\mathbf{g})$ corresponding to the indices

$$j = 2 - \mathrm{mod}(q, 2), \ldots, n + 1 - \mathrm{mod}(q, 2);$$

$p = 20$: for $q = 4, 5$ choose the samplings of $\lambda^{(q)}(\mathbf{g})$ corresponding to the indices

$$j = 1 + \mathrm{mod}(q, 2), \ldots, n + \mathrm{mod}(q, 2);$$

Despite in applications p is often less then 10, these patterns of exceptions warrant further research for $p > 20$.

Consider the one-dimensional Laplacian eigenvalue problem

$$\begin{cases} -u''(x) = \lambda u(x), & x \in (0,1), \\ u(0) = u(1) = 0. \end{cases} \tag{V.20}$$

The resulting discretized system, using the $\mathbb{Q}_p$ Lagrangian Finite Element approximation, is

$$K_n^{(p)}\mathbf{u}_n = \lambda M_n^{(p)}\mathbf{u}_n, \tag{V.21}$$

where the matrices $K_n^{(p)}$ and $M_n^{(p)}$ are the stiffness and the mass matrices previously defined.

Thus we have to solve the generalized eigenvalue problem

$$L_n^{(p)}\mathbf{u}_n = \lambda \mathbf{u}_n. \tag{V.22}$$

where

$$L_n^{(p)} = (M_n^{(p)})^{-1}K_n^{(p)}. \tag{V.23}$$

In [78] authors proved that

$$\{n^2 L_n^{(p)}\}_n \sim_{\text{GLT}} \mathbf{r} = \mathbf{g}^{-1}\mathbf{f}.$$

We here present closed formulae for the computing the eigenvalues of $n^2 L_n^{(p)}$ via the sampling on the symbol $\mathbf{r}$ on the exact grid.

Algorithm 5.

1. *Do step 1. of the Algorithms 3 (equivalently Algorithm 4), just replacing* $\mathbf{f}(\theta)$ *(equivalently* $\mathbf{g}(\theta)$*) with* $\mathbf{r}(\theta)$.
2. *If* p *is even, for* $q = 1, \dots, p$,
   - 2.1 *For the eigenvalue function* $\lambda^{(1)}(\mathbf{r})$*, choose samplings with index* $j_q = 2, \dots, n+1$. *This corresponds to the choice of the grid* τ_{n-1}^{π}.
   - 2.2. *For the eigenvalue functions* $\lambda^{(q)}(\mathbf{r})$*, where* q *is even, choose samplings with index* $j_q = 2, \dots, n$. *This corresponds to the choice of the grid* τ_{n-1}.
   - 2.3. *For the eigenvalue functions* $\lambda^{(q)}(\mathbf{r})$*, where* q *is odd, and* $q \neq 1$*, choose samplings with index* $j_q = 1, \dots, n+1$. *This corresponds to the choice of the grid* $\tau_{n-1}^{0,\pi}$.
3. *If* p *is odd, for* $q = 1, \dots, p$,
   - 3.2. *For the eigenvalue functions* $\lambda^{(q)}(\mathbf{r})$*, where* q *is even, choose samplings with index* $j_q = 1, \dots, n+1$. *This corresponds to the choice of the grid* $\tau_{n-1}^{0,\pi}$.
   - 3.3. *For the eigenvalue functions* $\lambda^{(q)}(\mathbf{r})$*, where* q *is odd, choose samplings with index* $j_q = 2, \dots, n$. *This corresponds to the choice of the grid* τ_{n-1}.

Chapter VI

Technical Results

VI.1 Staggered DG matrix symbol for $k = 2$ and $p = 2$

Recall that for the two-dimensional case ($k = 2$) the matrix symbol f is given according to (II.13) by

$$f(\theta_1, \theta_2) = \hat{f}_{(0,0)} + \hat{f}_{(-1,0)} e^{-\mathbf{i}\theta_1} + \hat{f}_{(0,-1)} e^{-\mathbf{i}\theta_2} + \hat{f}_{(1,0)} e^{\mathbf{i}\theta_1} + \hat{f}_{(0,1)} e^{\mathbf{i}\theta_2}.$$

For the special case $p = 2$, the matrices appearing in the above expression (see [65] for details concerning their definition) read

$$\hat{f}_{(0,0)} = \begin{pmatrix}
\frac{127}{360} & \frac{41}{480} & \frac{-43}{320} & \frac{41}{480} & \frac{-1}{360} & \frac{-2}{45} & \frac{-43}{320} & \frac{-2}{45} & \frac{13}{288} \\
\frac{41}{480} & \frac{103}{90} & \frac{41}{480} & \frac{-1}{360} & \frac{5}{24} & \frac{-1}{360} & \frac{-2}{45} & \frac{-113}{240} & \frac{-2}{45} \\
\frac{-43}{320} & \frac{41}{480} & \frac{127}{360} & \frac{-2}{45} & \frac{-1}{360} & \frac{41}{480} & \frac{13}{288} & \frac{-2}{45} & \frac{-43}{320} \\
\frac{41}{480} & \frac{-1}{360} & \frac{-2}{45} & \frac{103}{90} & \frac{5}{24} & \frac{-113}{240} & \frac{41}{480} & \frac{-1}{360} & \frac{-2}{45} \\
\frac{-1}{360} & \frac{5}{24} & \frac{-1}{360} & \frac{5}{24} & \frac{158}{45} & \frac{5}{24} & \frac{-1}{360} & \frac{5}{24} & \frac{-1}{360} \\
\frac{-2}{45} & \frac{-1}{360} & \frac{41}{480} & \frac{-113}{240} & \frac{5}{24} & \frac{103}{90} & \frac{-2}{45} & \frac{-1}{360} & \frac{41}{480} \\
\frac{-43}{320} & \frac{-2}{45} & \frac{13}{288} & \frac{41}{480} & \frac{-1}{360} & \frac{-2}{45} & \frac{127}{360} & \frac{41}{480} & \frac{-43}{320} \\
\frac{-2}{45} & \frac{-113}{240} & \frac{-2}{45} & \frac{-1}{360} & \frac{5}{24} & \frac{-1}{360} & \frac{41}{480} & \frac{103}{90} & \frac{41}{480} \\
\frac{13}{288} & \frac{-2}{45} & \frac{-43}{320} & \frac{-2}{45} & \frac{-1}{360} & \frac{41}{480} & \frac{-43}{320} & \frac{41}{480} & \frac{127}{360}
\end{pmatrix};$$

$$\hat{f}_{(-1,0)} = \begin{pmatrix} \frac{5}{288} & \frac{5}{576} & \frac{-5}{1152} & \frac{23}{720} & \frac{23}{1440} & \frac{-23}{2880} & \frac{-11}{1440} & \frac{-11}{2880} & \frac{11}{5760} \\ \frac{5}{576} & \frac{5}{72} & \frac{5}{576} & \frac{23}{1440} & \frac{23}{180} & \frac{23}{1440} & \frac{-11}{2880} & \frac{-11}{360} & \frac{-11}{2880} \\ \frac{-5}{1152} & \frac{5}{576} & \frac{5}{288} & \frac{-23}{2880} & \frac{23}{1440} & \frac{23}{720} & \frac{11}{5760} & \frac{-11}{2880} & \frac{-11}{1440} \\ \frac{-17}{144} & \frac{-17}{288} & \frac{17}{576} & \frac{-47}{360} & \frac{-47}{720} & \frac{47}{1440} & \frac{23}{720} & \frac{23}{1440} & \frac{-23}{2880} \\ \frac{-17}{288} & \frac{-17}{36} & \frac{-17}{288} & \frac{-47}{720} & \frac{-47}{90} & \frac{-47}{720} & \frac{23}{1440} & \frac{23}{180} & \frac{23}{1440} \\ \frac{17}{576} & \frac{-17}{288} & \frac{-17}{144} & \frac{47}{1440} & \frac{-47}{720} & \frac{-47}{360} & \frac{-23}{2880} & \frac{23}{1440} & \frac{23}{720} \\ \frac{-7}{288} & \frac{-7}{576} & \frac{7}{1152} & \frac{-17}{144} & \frac{-17}{288} & \frac{17}{576} & \frac{5}{288} & \frac{5}{576} & \frac{-5}{1152} \\ \frac{-7}{576} & \frac{-7}{72} & \frac{-7}{576} & \frac{-17}{288} & \frac{-17}{36} & \frac{-17}{288} & \frac{5}{576} & \frac{5}{72} & \frac{5}{576} \\ \frac{7}{1152} & \frac{-7}{576} & \frac{-7}{288} & \frac{17}{576} & \frac{-17}{288} & \frac{-17}{144} & \frac{-5}{1152} & \frac{5}{576} & \frac{5}{288} \end{pmatrix};$$

$$\hat{f}_{(0,-1)} = \begin{pmatrix} \frac{5}{288} & \frac{23}{720} & \frac{-11}{1440} & \frac{5}{576} & \frac{23}{1440} & \frac{-11}{2880} & \frac{-5}{1152} & \frac{-23}{2880} & \frac{11}{5760} \\ \frac{-17}{144} & \frac{-47}{360} & \frac{23}{720} & \frac{-17}{288} & \frac{-47}{720} & \frac{23}{1440} & \frac{17}{576} & \frac{47}{1440} & \frac{-23}{2880} \\ \frac{-7}{288} & \frac{-17}{144} & \frac{5}{288} & \frac{-7}{576} & \frac{-17}{288} & \frac{5}{576} & \frac{7}{1152} & \frac{17}{576} & \frac{-5}{1152} \\ \frac{5}{576} & \frac{23}{1440} & \frac{-11}{2880} & \frac{5}{72} & \frac{23}{180} & \frac{-11}{360} & \frac{5}{576} & \frac{23}{1440} & \frac{-11}{2880} \\ \frac{-17}{288} & \frac{-47}{720} & \frac{23}{1440} & \frac{-17}{36} & \frac{-47}{90} & \frac{23}{180} & \frac{-17}{288} & \frac{-47}{720} & \frac{23}{1440} \\ \frac{-7}{576} & \frac{-17}{288} & \frac{5}{576} & \frac{-7}{72} & \frac{-17}{36} & \frac{5}{72} & \frac{-7}{576} & \frac{-17}{288} & \frac{5}{576} \\ \frac{-5}{1152} & \frac{-23}{2880} & \frac{11}{5760} & \frac{5}{576} & \frac{23}{1440} & \frac{-11}{2880} & \frac{5}{288} & \frac{23}{720} & \frac{-11}{1440} \\ \frac{17}{576} & \frac{47}{1440} & \frac{-23}{2880} & \frac{-17}{288} & \frac{-47}{720} & \frac{23}{1440} & \frac{-17}{144} & \frac{-47}{360} & \frac{23}{720} \\ \frac{7}{1152} & \frac{17}{576} & \frac{-5}{1152} & \frac{-7}{576} & \frac{-17}{288} & \frac{5}{576} & \frac{-7}{288} & \frac{-17}{144} & \frac{5}{288} \end{pmatrix};$$

$$\hat{f}_{(1,0)} = \hat{f}^T_{(-1,0)};$$

$$\hat{f}_{(0,1)} = \hat{f}^T_{(0,-1)}.$$

VI.2 Proof of the preconditioned eigenvalue expansion for $\alpha = 0$

Theorem VI.2.1. *Let f, g be real-valued cosine trigonometric polynomials (RCTP) on $[0,\pi]$ with $M_g = \max g > 0$ and $m_g = \min g \geq 0$. If $r = \frac{f}{g}$ is monotone on $[0,\pi]$ then $\exists C > 0$ such*

that

$$\left|\lambda_j(\mathcal{P}_n(f,g)) - r\left(\frac{j\pi}{n+1}\right)\right| \le Ch \quad \forall\, j, \forall\, n, \tag{VI.1}$$

where

- $\mathcal{P}_n(f,g)$ *is the "preconditioned" matrix* $\mathcal{P}_n(f,g) = T_n^{-1}(g)T_n(f)$,
- $\lambda_1(\mathcal{P}_n(f,g)), \lambda_2(\mathcal{P}_n(f,g)), \ldots, \lambda_n(\mathcal{P}_n(f,g))$ *are the eigenvalues of* $\mathcal{P}_n(f,g)$*, arranged in nondecreasing or nonincreasing order, depending on whether* r *is increasing or decreasing,*
- $h = \frac{1}{n+1}$ *and* $\theta_{j,n} = \frac{j\pi}{n+1} = j\pi h$.

Proof. For the sake of simplicity, we assume that r is nondecreasing (the other case has a similar proof).
Notice that the conditions on f and g imply that $T_n(g)$ is positive definite and we find

$$\mathcal{P}_n(f,g) \sim T_n^{-1/2}(g)T_n(f)T_n^{-1/2}(g),$$

so we can order the eigenvalues of $\mathcal{P}_n(f,g)$ as follows

$$\lambda_1(\mathcal{P}_n(f,g)) \le \lambda_2(\mathcal{P}_n(f,g)) \le \cdots \le \lambda_n(\mathcal{P}_n(f,g)).$$

We remark the decomposition (I.11) of Section I.5

$$\begin{aligned} T_n(f) &= \tau_n(f) + H_n(f), \\ T_n(g) &= \tau_n(g) + H_n(g), \end{aligned} \tag{VI.2}$$

where, for ψ RCTP of degree m and orthogonal $Q = \left(\sqrt{\frac{2}{n+1}}\sin\left(\frac{ij\pi}{n+1}\right)\right)_{i,j=1}^{n}$, $\tau_n(\psi)$ is the following τ matrix [14] of size n generated by ψ

$$\tau_n(\psi) = Q \operatorname*{diag}_{1\le j\le n}\left(\psi\left(\frac{j\pi}{n+1}\right)\right)Q, \quad Q = Q^T = Q^{-1},$$

and $H_n(\psi)$ is the Hankel matrix generated by ψ with $\operatorname{rank}(H_n(\psi)) \le 2(m-1)$.
Hence,

$$\begin{aligned} R_f &:= \operatorname{rank}(H_n(f)) \le 2(\deg(f) - 1), \\ R_g &:= \operatorname{rank}(H_n(g)) \le 2(\deg(g) - 1), \\ R_{f,g} &:= \max\{R_f, R_g\} \le 2\left(\max\{\deg(f), \deg(g)\} - 1\right). \end{aligned} \tag{VI.3}$$

Let P_n^τ be the matrix $\tau_n^{-1}(g)\tau_n(f)$,

$$\begin{aligned} P_n^\tau &= Q\left(\operatorname*{diag}_{1\le j\le n}\left(g\left(\frac{j\pi}{n+1}\right)\right)\right)^{-1} QQ \operatorname*{diag}_{1\le j\le n}\left(f\left(\frac{j\pi}{n+1}\right)\right)Q \\ &= Q \operatorname*{diag}_{1\le j\le n}\left(\frac{f}{g}\left(\frac{j\pi}{n+1}\right)\right)Q \\ &= Q \operatorname*{diag}_{1\le j\le n}\left(r\left(\frac{j\pi}{n+1}\right)\right)Q. \end{aligned}$$

Hence, for $j = 1, \ldots, n$

$$\lambda_j(P_n^\tau) = r\left(\frac{j\pi}{n+1}\right). \tag{VI.4}$$

By observing that $T_n^{-1}(g)T_n(f)$ is similar to $T_n^{-1/2}(g)T_n(f)T_n^{-1/2}(g)$, using the MinMax spectral characterization for Hermitian matrices [13], fixed $j \in \{R_{f,g}+1, \ldots, n-R_{f,g}\}$ and $T \subset \mathbb{C}^n$, $\dim(T) = n+1-j$, we obtain

$$\begin{aligned}
\lambda_j(\mathcal{P}_n(f,g)) &= \lambda_j\left(T_n^{-1}(g)T_n(f)\right) \\
&= \lambda_j\left(T_n^{-1/2}(g)T_n(f)T_n^{-1/2}(g)\right) \\
&= \max_{\dim(T)=n+1-j}\left(\min_{\substack{x\in T,\\ x\neq \underline{0}}}\left(\frac{x^*T_n^{-1/2}(g)T_n(f)T_n^{-1/2}(g)x}{x^*x}\right)\right) \\
&= \max_{\dim(T)=n+1-j}\left(\min_{\substack{x\in T,\\ x\neq 0\\ y=T_n^{-1/2}(g)x}}\left(\frac{y^*T_n(f)y}{y^*T_n(g)y}\right)\right) \\
&= \max_{\dim(\hat{T})=n+1-j}\left(\min_{\substack{y\in \hat{T},\\ y\neq \underline{0}}}\left(\frac{y^*T_n(f)y}{y^*T_n(g)y}\right)\right),
\end{aligned} \tag{VI.5}$$

because $T_n^{-1/2}(g)$ is a full rank matrix and, if $\dim(T) = n+1-j$, then $\hat{T} := \{y : y = T_n^{-1/2}(g)x,\ x \neq \underline{0}, x \in T\}$ is a new vector space having the same dimension $n+1-j$ as T.
Let F be the subspace of $\mathbb{C}^n$ generated by the union of the columns of matrices $H_n(f)$ and $H_n(g)$. Because of the particular structure of the columns of Hankel matrices $H_n(f)$ and $H_n(g)$, we deduce

$$\dim(F) = \max\{\operatorname{rank}(H_n(g)), \operatorname{rank}(H_n(f))\} = R_{f,g},$$

so that

$$\dim(F^\perp) = n - R_{f,g}.$$

Let us define $W_{f,g} = \hat{T} \cap F^\perp$,

$$n+1-j \geq \dim(W_{f,g}) \geq \max\{0, \dim(\hat{T})+\dim(F^\perp)-n\} = n+1-j+n-R_{f,g}-n = n+1-(j+R_{f,g}),$$

because $n+1-(j+R_{f,g}) \geq 1$ for $j \leq n-R_{f,g}$.The latter implies in particular that $W_{f,g} \neq \emptyset$. Thus, because of the orthogonality, $\forall\, y \neq \underline{0} \in W_{f,g}$, we find

$$H_n(f)y = \underline{0}, \quad H_n(g)y = \underline{0},$$

so that

$$y^*H_n(f)y = 0, \quad y^*H_n(g)y = 0.$$

Hence, from (VI.5)

$$\begin{aligned}
\lambda_j(\mathcal{P}_n(f,g)) &= \max_{\dim(\hat{T})=n+1-j} \left(\min_{\substack{y\in\hat{T},\\ y\neq\underline{0}}} \left(\frac{y^*(\tau_n(f)+H_n(f))y}{y^*(\tau_n(g)+H_n(g))y} \right) \right) \\
&\leq \max_{\dim(\hat{T})=n+1-j} \left(\min_{\substack{y\in W_{f,g}\\ y\neq\underline{0}}} \left(\frac{y^*(\tau_n(f)+H_n(f))y}{y^*(\tau_n(g)+H_n(g))y} \right) \right) \\
&= \max_{\dim(\hat{T})=n+1-j} \left(\min_{\substack{y\in W_{f,g}\\ y\neq\underline{0}}} \left(\frac{y^*\tau_n(f)y}{y^*\tau_n(g)y} \right) \right) \\
&= \max_{\substack{W_{f,g}=\hat{T}\cap F^\perp\\ \dim(\hat{T})=n+1-j}} \left(\min_{\substack{y\in W_{f,g},\\ y\neq\underline{0}}} \left(\frac{y^*\tau_n(f)y}{y^*\tau_n(g)y} \right) \right) \\
&\leq \max_{n+1-j\geq\dim(\hat{W}_{f,g})\geq n+1-(j+R_{f,g})} \left(\min_{\substack{y\in\hat{W}_{f,g},\\ y\neq\underline{0}}} \left(\frac{y^*\tau_n(f)y}{y^*\tau_n(g)y} \right) \right) \\
&= \max_{n+1-j\geq\dim(\hat{W})\geq n+1-(j+R_{f,g})} \left(\min_{\substack{y\in\hat{W}_{f,g},\\ y\neq\underline{0}\\ x=\tau_n^{1/2}(g)y}} \left(\frac{x^*\tau_n^{-1/2}(g)\tau_n(f)\tau_n^{-1/2}(g)x}{x^*x} \right) \right) \\
&= \max\{\lambda_j(P_n^\tau), \lambda_{j+1}(P_n^\tau), \ldots, \lambda_{j+R_{f,g}}(P_n^\tau)\} \\
&= \lambda_{j+R_{f,g}}(P_n^\tau).
\end{aligned} \tag{VI.6}$$

By fixing $j \in \{R_{f,g}+1, \ldots, n-R_{f,g}\}$ and $T \subset \mathbb{C}^n$, $\dim(T) = j$, analogously we obtain

$$\begin{aligned}
\lambda_j(\mathcal{P}_n(f,g)) &= \min_{\dim(T)=j} \left(\max_{\substack{x\in T,\\ x\neq\underline{0}}} \left(\frac{x^*T_n^{-1/2}(g)T_n(f)T_n^{-1/2}(g)x}{x^*x} \right) \right) \\
&= \min_{\dim(T)=j} \left(\max_{\substack{x\in T,\\ x\neq\underline{0}\\ y=T_n^{-1/2}(g)x}} \left(\frac{y^*T_n(f)y}{y^*T_n(g)y} \right) \right) \\
&= \min_{\dim(\hat{T})=j} \left(\max_{\substack{y\in\hat{T},\\ y\neq\underline{0}}} \left(\frac{y^*T_n(f)y}{y^*T_n(g)y} \right) \right) \\
&= \min_{\dim(\hat{T})=j} \left(\max_{\substack{y\in\hat{T},\\ y\neq\underline{0}}} \left(\frac{y^*(\tau_n(f)+H_n(f))y}{y^*(\tau_n(g)+H_n(g))y} \right) \right).
\end{aligned} \tag{VI.7}$$

Let us define $W_{f,g} = \hat{T} \cap F^\perp$,

$$j \geq \dim(W_{f,g}) \geq \max\{0, \dim(\hat{T}) + \dim(F^\perp) - n\} = j + n - R_{f,g} - n = j - R_{f,g},$$

because $j - R_{f,g} \geq 1$ for $j \geq R_{f,g} + 1$. The latter implies in particular that $W_{f,g} \neq \emptyset$, and hence, because of the orthogonality, $\forall\, y \neq \underline{0} \in W_{f,g}$, we have

$$H_n(f)y = \underline{0}, \quad H_n(g)y = \underline{0},$$

and therefore

$$y^* H_n(f)y = 0, \quad y^* H_n(g)y = 0.$$

Thus, from (VI.7)

$$\begin{aligned}
\lambda_j(\mathcal{P}_n(f,g)) &\geq \min_{\dim(\hat{T})=j} \left(\max_{\substack{y \in W_{f,g}, \\ y \neq \underline{0}}} \left(\frac{y^*(\tau_n(f) + H_n(f))y}{y^*(\tau_n(g) + H_n(g))y} \right) \right) \\
&= \min_{\dim(\hat{T})=j} \left(\max_{\substack{y \in W_{f,g}, \\ y \neq \underline{0}}} \left(\frac{y^* \tau_n(f) y}{y^* \tau_n(g) y} \right) \right) \\
&= \min_{\substack{W_{f,g} = \hat{T} \cap F^{\perp} \\ \dim(\hat{T})=j}} \left(\max_{\substack{y \in W_{f,g}, \\ x \neq \underline{0}}} \left(\frac{y^* \tau_n(f) y}{y^* \tau_n(g) y} \right) \right) \\
&\geq \min_{j \geq \dim(\hat{W}_{f,g}) \geq j - R_{f,g}} \left(\max_{\substack{y \in W_{f,g}, \\ y \neq \underline{0}}} \left(\frac{y^* \tau_n(f) y}{y^* \tau_n(g) y} \right) \right) \\
&= \min\{\lambda_j(P_n^{\tau}), \lambda_{j-1}(P_n^{\tau}), \ldots, \lambda_{j-R_{f,g}}(P_n^{\tau})\} \\
&= \lambda_{j-R_{f,g}}(P_n^{\tau}).
\end{aligned} \tag{VI.8}$$

By exploiting the previous inequality, relations (VI.4) and (VI.6), we obtain for $j = R_{f,g} + 1, \ldots, n - R_{f,g}$

$$r\left(\frac{(j-s)\pi}{n+1}\right) = \lambda_{j-s}(P_n^{\tau}) \leq \lambda_j(\mathcal{P}_n(f,g)) \leq \lambda_{j+s}(P_n^{\tau}) = r\left(\frac{(j+s)\pi}{n+1}\right), \tag{VI.9}$$

where $s = R_{f,g}$.
The function r is a RCTP on $[0, \pi]$ and a monotone increasing function so we have, $\forall\, n$ and $\forall\, j = s+1, \ldots, n-s$,

$$\lambda_j(\mathcal{P}_n(f,g)) - r\left(\frac{j\pi}{n+1}\right) \leq r\left(\frac{(j+s)\pi}{n+1}\right) - r\left(\frac{j\pi}{n+1}\right) = r'(\bar{\theta})\frac{s\pi}{n+1} \leq ||r'||_{\infty} s\pi h, \tag{VI.10}$$

with $\bar{\theta} \in (\frac{j\pi}{n+1}, \frac{(j+s)\pi}{n+1})$ and

$$\lambda_j(\mathcal{P}_n(f,g)) - r\left(\frac{j\pi}{n+1}\right) \geq r\left(\frac{(j-s)\pi}{n+1}\right) - r\left(\frac{j\pi}{n+1}\right) \geq -||r'||_{\infty} s\pi h. \tag{VI.11}$$

By setting $C = ||r'||_{\infty} s\pi$, for $s+1 \leq j \leq n-s$, we obtain

$$\left| \lambda_j(\mathcal{P}_n(f,g)) - r\left(\frac{j\pi}{n+1}\right) \right| \leq Ch. \tag{VI.12}$$

Furthermore, from [47] $\forall\, j = 1, \ldots, n$, we know that

$$m_r \leq \lambda_j(\mathcal{P}_n(f,g)) \leq M_r,$$

where

$$m_r = \min_{\theta\in[0,\pi]} r(\theta); \quad M_r = \max_{\theta\in[0,\pi]} r(\theta),$$

with strict inequalities that is $m_r < \lambda_j(\mathcal{P}_n(f,g)) < M_r$ if $m_r < M_r$, while the case $m_r = M_r$ is in fact trivial. Hence for $n - s < j \leq n$

$$\left| r\left(\frac{j\pi}{n+1}\right) - \lambda_j(\mathcal{P}_n(f,g)) \right| \leq \left| r\left(\frac{j\pi}{n+1}\right) - r\left(\frac{n\pi}{n+1}\right) \right| \leq |r'(\bar{\theta})| \left| \frac{(n-j)\pi}{n+1} \right|,$$

where $\bar{\theta} \in (\frac{j\pi}{n+1}, \frac{n\pi}{n+1})$. If $n - s < j \leq n$ then $|n - j| < s$, so that

$$\left| r\left(\frac{j\pi}{n+1}\right) - \lambda_j(\mathcal{P}_n(f,g)) \right| \leq ||r'||_\infty s\pi h = Ch.$$

For $1 \leq j < s + 1$

$$\left| r\left(\frac{j\pi}{n+1}\right) - \lambda_j(\mathcal{P}_n(f,g)) \right| \leq \left| r\left(\frac{j\pi}{n+1}\right) - r\left(\frac{\pi}{n+1}\right) \right| \leq |r'(\bar{\theta})| \left| \frac{(j-1)\pi}{n+1} \right|,$$

where $\bar{\theta} \in (\frac{\pi}{n+1}, \frac{j\pi}{n+1})$. If $1 \leq j < s + 1$ then $|j - 1| < s$, so

$$\left| r\left(\frac{j\pi}{n+1}\right) - \lambda_j(\mathcal{P}_n(f,g)) \right| \leq ||r'||_\infty s\pi h = Ch.$$

Hence

$$\left| \lambda_j(\mathcal{P}_n(f,g)) - r\left(\frac{j\pi}{n+1}\right) \right| \leq Ch \quad \forall\, j \,\forall\, n.$$

□

Remark 11. *With regard to Theorem VI.2.1, the case where r is bounded and non-monotone is almost analogous. If we consider $\hat{r}$, the monotone nondecreasing rearrangement of r on $[0,\pi]$, taking into account that the derivative of r has at most a finite number S of sign changes, we deduce that $\hat{r}$ is Lipschitz continuous and its Lipschitz constant is bounded by $\|r'\|_\infty$ (notice that $\hat{r}$ is not necessarily continuously differentiable, but the derivative of $\hat{r}$ has at most S points of discontinuity). Furthermore, the eigenvalues of $\tau_n(r)$ are exactly given*

$$r\left(\frac{j\pi}{n+1}\right)$$

so that, by ordering these values nondecreasingly, we deduce that they coincide with $\hat{r}(x_{j,h})$, with $x_{j,h}$ of the form $\frac{j\pi}{n+1}(1+o(1))$. With these premises, the proof follows exactly the same steps as in Theorem VI.2.1, using the MinMax characterization and the Interlacing theorem for Hermitian matrices.

VI.3 Proofs of the theorems stated in Section IV.2 of Chapter IV

We first recall from [72, Section 3] that, for every $p \geq 0$ and $\theta \in [0,\pi]$,

$$g_p(\theta) = \sum_{k\in\mathbb{Z}} \left| \widehat{\phi_p}(\theta + 2k\pi) \right|^2, \tag{VI.13}$$

where $\widehat{\phi_p}$ is the Fourier transform of the cardinal B-spline ϕ_p, whose modulus is given by

$$\left|\widehat{\phi_p}(\theta)\right|^2 = \left(\frac{2-2\cos(\theta)}{\theta^2}\right)^{p+1}; \tag{VI.14}$$

see [38]. The next lemma is fundamental to our purposes.

Lemma VI.3.1. *For $p \geq 1$ and $\theta \in [0,\pi]$ we have*

$$\frac{9}{5}\pi(\pi-\theta)\left(\frac{\theta}{2\pi-\theta}\right)^{2p+2} \leq e_p(\theta)-\theta^2 \leq 4\pi(\pi-\theta)\left(\frac{\theta}{2\pi-\theta}\right)^{2p+2} + 5\theta^2\left(\frac{\theta}{2\pi+\theta}\right)^{2p}. \tag{VI.15}$$

Proof. From (IV.16) and (VI.13)–(VI.14) we obtain

$$f_p(\theta) = (2-2\cos(\theta))^{p+1}\sum_{k\in\mathbb{Z}}\frac{1}{(\theta+2k\pi)^{2p}} = (2-2\cos(\theta))^{p+1}\left[\frac{1}{\theta^{2p}}+\sum_{k\neq 0}\frac{1}{(\theta+2k\pi)^{2p}}\right],$$

$$g_p(\theta) = (2-2\cos(\theta))^{p+1}\sum_{k\in\mathbb{Z}}\frac{1}{(\theta+2k\pi)^{2p+2}} = (2-2\cos(\theta))^{p+1}\left[\frac{1}{\theta^{2p+2}}+\sum_{k\neq 0}\frac{1}{(\theta+2k\pi)^{2p+2}}\right].$$

By setting

$$r_p(\theta) = \theta^{2p}\sum_{k\neq 0}\frac{1}{(\theta+2k\pi)^{2p}} \geq 0,$$

we see that

$$e_p(\theta)-\theta^2 = \frac{f_p(\theta)}{g_p(\theta)}-\theta^2 = \theta^2\frac{1+r_p(\theta)}{1+r_{p+1}(\theta)}-\theta^2 = \theta^2\frac{r_p(\theta)-r_{p+1}(\theta)}{1+r_{p+1}(\theta)}. \tag{VI.16}$$

Furthermore,

$$r_p(\theta)-r_{p+1}(\theta) = \theta^{2p}(A_{p,+}(\theta)+A_{p,-}(\theta)) \tag{VI.17}$$

where

$$A_{p,+}(\theta) = \sum_{k\geq 1}\frac{1}{(2k\pi+\theta)^{2p}}\left(1-\frac{\theta^2}{(2k\pi+\theta)^2}\right), \tag{VI.18}$$

$$A_{p,-}(\theta) = \sum_{k\geq 1}\frac{1}{(2k\pi-\theta)^{2p}}\left(1-\frac{\theta^2}{(2k\pi-\theta)^2}\right). \tag{VI.19}$$

Assume $\theta \in [0,\pi]$. We observe that

$$0 \leq 1-\frac{\theta^2}{(2k\pi+\theta)^2} \leq 1, \qquad k \geq 1,$$

which implies

$$\begin{aligned} A_{p,+}(\theta) &\leq \frac{1}{(2\pi+\theta)^{2p}}+\sum_{k\geq 2}\frac{1}{(2k\pi+\theta)^{2p}} \leq \frac{1}{(2\pi+\theta)^{2p}}+\int_1^{+\infty}\frac{\mathrm{d}\kappa}{(2\pi\kappa+\theta)^{2p}} \\ &= \frac{1}{(2\pi+\theta)^{2p}}+\frac{1}{2\pi(2p-1)(2\pi+\theta)^{2p-1}} \leq \frac{5}{2}\frac{1}{(2\pi+\theta)^{2p}}. \end{aligned}$$

Similarly,

$$\begin{aligned} A_{p,-}(\theta) &\le \frac{4\pi(\pi-\theta)}{(2\pi-\theta)^{2p+2}} + \frac{8\pi(2\pi-\theta)}{(4\pi-\theta)^{2p+2}} + \sum_{k\ge 3} \frac{1}{(2k\pi-\theta)^{2p}} \\ &\le \frac{4\pi(\pi-\theta)}{(2\pi-\theta)^{2p+2}} + \frac{8\pi(2\pi-\theta)}{(4\pi-\theta)^{2p+2}} + \int_2^{+\infty} \frac{\mathrm{d}\kappa}{(2\pi\kappa-\theta)^{2p}} \\ &= \frac{4\pi(\pi-\theta)}{(2\pi-\theta)^{2p+2}} + \frac{8\pi(2\pi-\theta)}{(4\pi-\theta)^{2p+2}} + \frac{1}{2\pi(2p-1)(4\pi-\theta)^{2p-1}} \\ &\le \frac{4\pi(\pi-\theta)}{(2\pi-\theta)^{2p+2}} + \frac{5}{2}\frac{1}{(2\pi+\theta)^{2p}}, \end{aligned}$$

where we have exploited the fact that

$$4\pi-\theta \ge 2\pi+\theta, \qquad \frac{8\pi(2\pi-\theta)}{(4\pi-\theta)^2} \le 1.$$

By combining (VI.16) and (VI.17) with the obtained upper bounds for $A_{p,+}$ and $A_{p,-}$, we get the upper bound in (VI.15).

To prove the lower bound in (VI.15), we use the inequality

$$\begin{aligned} r_{p+1}(\theta) &\le \theta^{2p+2}\left(\frac{1}{(2\pi+\theta)^{2p+2}} + \frac{1}{(2\pi-\theta)^{2p+2}} + \int_1^{+\infty}\left[\frac{1}{(2\pi\kappa+\theta)^{2p+2}} + \frac{1}{(2\pi\kappa-\theta)^{2p+2}}\right]\mathrm{d}\kappa\right) \\ &= \theta^{2p+2}\left(\frac{1}{(2\pi+\theta)^{2p+2}} + \frac{1}{(2\pi-\theta)^{2p+2}} + \frac{1}{2\pi(2p+1)}\left[\frac{1}{(2\pi+\theta)^{2p+1}} + \frac{1}{(2\pi-\theta)^{2p+1}}\right]\right). \end{aligned}$$

Note that

$$\frac{1}{(2\pi+\theta)^q} + \frac{1}{(2\pi-\theta)^q} \le \frac{1}{(3\pi)^q} + \frac{1}{\pi^q}, \qquad q \ge 1,$$

since the function on the left-hand side is monotone increasing for $\theta \in [0,\pi]$. Therefore, for $p \ge 1$,

$$r_{p+1}(\theta) \le \left(\frac{\theta}{\pi}\right)^{2p+2}\left(\frac{1}{3^{2p+2}} + 1 + \frac{1}{2(2p+1)}\left[\frac{1}{3^{2p+1}} + 1\right]\right) \le \frac{1}{81} + 1 + \frac{1}{6}\left[\frac{1}{27} + 1\right] = \frac{32}{27}.$$

Moreover, from (VI.18) and (VI.19) we deduce that

$$A_{p,+}(\theta) + A_{p,-}(\theta) = \sum_{k\ge 1} \frac{4k\pi(k\pi+\theta)}{(2k\pi+\theta)^{2p+2}} + \frac{4k\pi(k\pi-\theta)}{(2k\pi-\theta)^{2p+2}} \ge \frac{4\pi(\pi-\theta)}{(2\pi-\theta)^{2p+2}}.$$

Taking into account (VI.17), we arrive at

$$\frac{r_p(\theta) - r_{p+1}(\theta)}{1 + r_{p+1}(\theta)} \ge \frac{4\pi(\pi-\theta)}{(2\pi-\theta)^2}\left(\frac{\theta}{2\pi-\theta}\right)^{2p}\frac{27}{59}.$$

In view of (VI.16), this immediately gives the lower bound in (VI.15). □

We are now ready to prove Theorems IV.2.1 and IV.2.3.

Proof of Theorem IV.2.1. From the upper bound in (VI.15) we have

$$\max_{\theta\in[0,\pi]} \left|e_p(\theta) - \theta^2\right| \le \max_{\theta\in[0,\pi]} \left[4\pi(\pi-\theta)\left(\frac{\theta}{2\pi-\theta}\right)^{2p+2} + 5\theta^2\left(\frac{\theta}{2\pi+\theta}\right)^{2p}\right].$$

By setting $z = \frac{\theta}{\pi} \in [0,1]$ we obtain

$$\max_{\theta\in[0,\pi]} \left|e_p(\theta) - \theta^2\right| \leq \max_{z\in[0,1]} \left[4\pi^2(1-z)\left(\frac{z}{2-z}\right)^{2p+2} + 5\pi^2 z^2 \left(\frac{z}{2+z}\right)^{2p}\right]$$
$$\leq \max_{z\in[0,1]} 5\pi^2 \left[\left(\frac{z}{2-z}\right)^{2p+2}\left(1 - \frac{z}{2-z}\right) + \frac{1}{3^{2p}}\right].$$

Finally, by setting $y = \frac{z}{2-z} \in [0,1]$ and observing that

$$\max_{y\in[0,1]} y^{2p+2}(1-y) = \left(1 - \frac{1}{2p+3}\right)^{2p+2} \frac{1}{2p+3} \leq \frac{1}{2p+3},$$

we get

$$\max_{\theta\in[0,\pi]} \left|e_p(\theta) - \theta^2\right| \leq 5\pi^2 \left(\frac{1}{2p+3} + \frac{1}{3^{2p}}\right).$$

This concludes the proof. □

Proof of Theorem IV.2.3. For $p = 1$ the bounds $1/3 \leq w_p(\theta) \leq 1$ stated in the theorem hold because from (IV.14) we know that

$$g_0(\theta) = 1, \qquad g_1(\theta) = \frac{2}{3} + \frac{1}{3}\cos(\theta).$$

In the following we focus on the case $p \geq 2$. From (IV.17) it is clear that the bounds hold for $\theta = 0$. From (IV.16) and (VI.15) we deduce that, for $\theta \in (0,\pi]$,

$$1 \leq \frac{1}{\theta^2}\frac{f_p(\theta)}{g_p(\theta)} = \frac{2-2\cos(\theta)}{\theta^2}\frac{g_{p-1}(\theta)}{g_p(\theta)} \leq 1 + \frac{4\pi(\pi-\theta)}{(2\pi-\theta)^2}\left(\frac{\theta}{2\pi-\theta}\right)^{2p} + 5\left(\frac{\theta}{2\pi+\theta}\right)^{2p}$$
$$\leq 1 + \frac{4\pi(\pi-\theta)}{(2\pi-\theta)^2}\left(\frac{\theta}{2\pi-\theta}\right)^{4} + 5\left(\frac{1}{3}\right)^{4} \leq 1 + \frac{3}{20} + \frac{5}{81} < \frac{12}{\pi^2}.$$

Since

$$1 \leq \frac{\theta^2}{2-2\cos(\theta)} \leq \frac{\pi^2}{4}, \qquad \theta \in (0,\pi],$$

we obtain

$$1 \leq \frac{g_{p-1}(\theta)}{g_p(\theta)} < 3,$$

which is equivalent to $1/3 < w_p(\theta) \leq 1$. □

In order to prove Theorem IV.2.2, further work is needed. In particular, we shall need to analyze the auxiliary functions

$$R_{k,p}(\omega) = \left(\frac{\omega}{k\pi+\omega}\right)^{2p+1} - \left(\frac{\omega}{k\pi-\omega}\right)^{2p+1}, \qquad k,p \geq 1, \qquad \omega \in \left[0, \frac{\pi}{2}\right]. \tag{VI.20}$$

The next three technical lemmas are devoted to this purpose.

Lemma VI.3.2. *For $p \geq 1$ and $k \geq 2$ the function*

$$R_{k,p+1}(\omega) - R_{k,p}(\omega) \tag{VI.21}$$

is nonnegative, monotone increasing and convex for $\omega \in [0, \frac{\pi}{2}]$.

Proof. Assume $\omega \in [0, \frac{\pi}{2}]$. We have

$$R_{k,p+1}(\omega) - R_{k,p}(\omega) = z_k^{2p+3} - z_k^{2p+1} + y_k^{2p+1} - y_k^{2p+3}, \qquad k \geq 1,$$

where

$$y_k = \frac{\omega}{k\pi - \omega}, \qquad z_k = \frac{\omega}{k\pi + \omega}. \tag{VI.22}$$

It is easy to check that y_k is a monotone increasing and convex function of ω. Similarly, z_k is a monotone increasing and concave function of ω. Moreover,

$$\frac{z_k}{y_k} = \frac{1}{1+2y_k}, \qquad \frac{z_k'}{y_k'} = \left(\frac{z_k}{y_k}\right)^2, \qquad \frac{z_k''}{y_k''} = -\left(\frac{z_k}{y_k}\right)^3, \qquad k \geq 1, \tag{VI.23}$$

and

$$0 \leq z_k \leq y_k \leq \frac{1}{2k-1} \leq \frac{1}{3}, \qquad k \geq 2. \tag{VI.24}$$

Proving the nonnegativity of the function in (VI.21) is equivalent to showing that

$$y_k^{2p+1}(1 - y_k^2) \geq z_k^{2p+1}(1 - z_k^2).$$

In view of (VI.23), this is equivalent to

$$\frac{1 - y_k^2}{1 - z_k^2} \geq \frac{1}{(1+2y_k)^{2p+1}}.$$

Since

$$\frac{1 - y_k^2}{1 - z_k^2} \geq 1 - y_k^2,$$

it suffices to prove that

$$1 - y_k^2 \geq \frac{1}{(1+2y_k)^{2p+1}}.$$

A direct computation shows that the above inequality holds for $y_k \in [0, 1/3]$ (it is enough to verify it for $p = 1$ as the right-hand side decreases with p). Taking into account (VI.24), this proves the nonnegativity of (VI.21) for $k \geq 2$.

We now show that the function (VI.21) is convex. With some elementary manipulations we obtain

$$R''_{k,p+1}(\omega) - R''_{k,p}(\omega) = A_k + B_k - C_k - D_k, \tag{VI.25}$$

where

$$A_k = 2y_k^{2p-1}(y_k')^2 \left[p(2p+1) - (p+1)(2p+3)y_k^2\right], \qquad B_k = y_k^{2p} y_k'' \left[2p+1 - (2p+3)y_k^2\right],$$
$$C_k = 2z_k^{2p-1}(z_k')^2 \left[p(2p+1) - (p+1)(2p+3)z_k^2\right], \qquad D_k = z_k^{2p} z_k'' \left[2p+1 - (2p+3)z_k^2\right].$$

From (VI.24) it follows that, for $p \geq 1$ and $k \geq 2$,

$$p(2p+1) - (p+1)(2p+3)x_k^2 > 0, \qquad 2p+1 - (2p+3)x_k^2 > 0, \qquad x_k = y_k, z_k.$$

As a consequence, we have $B_k \geq 0$ and $D_k \leq 0$ because $y_k'' \geq 0$ and $z_k'' \leq 0$. In the following we show that $A_k \geq C_k$. Taking into account (VI.23), this is equivalent to proving that

$$\frac{p(2p+1) - (p+1)(2p+3)y_k^2}{p(2p+1) - (p+1)(2p+3)z_k^2} \geq \left(\frac{z_k}{y_k}\right)^{2p-1} \left(\frac{z_k'}{y_k'}\right)^2 = \frac{1}{(1+2y_k)^{2p+3}}.$$

Since

$$\frac{p(2p+1)-(p+1)(2p+3)y_k^2}{p(2p+1)-(p+1)(2p+3)z_k^2} \geq \frac{p(2p+1)-(p+1)(2p+3)y_k^2}{p(2p+1)},$$

it suffices to prove that

$$1-\frac{(p+1)(2p+3)}{p(2p+1)}y_k^2 \geq \frac{1}{(1+2y_k)^{2p+3}}.$$

The above inequality holds for $p \geq 1$ and $y_k \in [0,1/3]$ (it is enough to verify it for $p=1$). Recalling (VI.24), this shows the convexity of (VI.21).

Finally, the monotonicity of the function (VI.21) follows from the convexity by observing that the first derivative vanishes at $\omega = 0$. □

Lemma VI.3.3. *For $p \geq 1$ the function*

$$R_{1,p+1}(\omega) - R_{1,p}(\omega) \tag{VI.26}$$

is nonnegative for $\omega \in [0,\omega_p^*]$ and concave for $\omega \in [\omega_p^*, \frac{\pi}{2}]$, where

$$\omega_p^* = \frac{\pi}{2}\left(1-\frac{1}{48p-1}\right). \tag{VI.27}$$

Proof. Along the proof we use the same notation as in the proof of Lemma VI.3.2. We first address the nonnegativity. With the same line of arguments as in the proof of Lemma VI.3.2 we deduce that the function in (VI.26) is nonnegative if

$$1-y_1^2 \geq \frac{1}{(1+2y_1)^{2p+1}}.$$

The above inequality holds for $p \geq 1$ whenever

$$0 \leq y_1 \leq 1-\frac{1}{24p} = y_{1,p}^*. \tag{VI.28}$$

In view of (VI.22) and (VI.27), this is equivalent to $0 \leq \omega \leq \omega_p^*$.

We now prove the concavity. Similarly to (VI.25), we have

$$R''_{1,p+1}(\omega) - R''_{1,p}(\omega) = A_1 + B_1 - C_1 - D_1,$$

where

$$\begin{aligned} A_1 &= 2y_1^{2p-1}(y_1')^2\left[p(2p+1)-(p+1)(2p+3)y_1^2\right], & B_1 &= y_1^{2p}y_1''\left[2p+1-(2p+3)y_1^2\right], \\ C_1 &= 2z_1^{2p-1}(z_1')^2\left[p(2p+1)-(p+1)(2p+3)z_1^2\right], & D_1 &= z_1^{2p}z_1''\left[2p+1-(2p+3)z_1^2\right]. \end{aligned}$$

Since $0 \leq z_1 \leq \frac{1}{3}$ we have

$$p(2p+1)-(p+1)(2p+3)z_1^2 > 0, \qquad 2p+1-(2p+3)z_1^2 > 0.$$

Moreover, for $y_1 \geq y_{1,p}^*$,

$$p(2p+1)-(p+1)(2p+3)y_1^2 < 0, \qquad 2p+1-(2p+3)y_1^2 < 0.$$

Hence, $A_1 < 0$ and $C_1 > 0$ for $\omega \in [\omega_p^*, \frac{\pi}{2}]$. In the following, for $\omega \in [\omega_p^*, \frac{\pi}{2}]$, we prove that $B_1 \leq D_1$, or equivalently

$$\frac{2p+1-(2p+3)y_1^2}{2p+1-(2p+3)z_1^2} \leq \left(\frac{z_1}{y_1}\right)^{2p} \frac{z_1''}{y_1''}.$$

By (VI.23) this is equivalent to

$$\frac{(2p+3)y_1^2-(2p+1)}{2p+1-(2p+3)z_1^2} \geq \frac{1}{(1+2y_1)^{2p+3}}.$$

Since

$$\frac{(2p+3)y_1^2-(2p+1)}{2p+1-(2p+3)z_1^2} \geq \frac{(2p+3)y_1^2-(2p+1)}{2p+1-(2p+1)z_1^2} = \left(\frac{2p+3}{2p+1}y_1^2-1\right)\frac{1}{1-z_1^2} \geq \frac{2p+3}{2p+1}y_1^2-1,$$

it suffices to prove that, for $\omega \in [\omega_p^*, \frac{\pi}{2}]$,

$$\frac{2p+3}{2p+1}y_1^2-1 \geq \frac{1}{(1+2y_1)^{2p+3}}. \tag{VI.29}$$

Note that the left-hand side in (VI.29) is monotone increasing while the right-hand side is monotone decreasing. Thus, the observation that the inequality (VI.29) holds for $y_1 = y_{1,p}^*$ and $p \geq 1$ concludes the proof. □

Lemma VI.3.4. *For $p \geq 1$ and $\omega \in [0, \frac{\pi}{2}]$ we have*

$$1+(p+1)\sum_{k\geq 1} R_{k,p+1}(\omega) - p\sum_{k\geq 1} R_{k,p}(\omega) \geq 0. \tag{VI.30}$$

Proof. Assume $\omega \in [0, \frac{\pi}{2}]$. When taking the derivative of $R_{k,p}$,

$$R_{k,p}'(\omega) = (2p+1)\left[\left(\frac{\omega}{k\pi+\omega}\right)^{2p}\frac{k\pi}{(k\pi+\omega)^2} - \left(\frac{\omega}{k\pi-\omega}\right)^{2p}\frac{k\pi}{(k\pi-\omega)^2}\right] \leq 0, \tag{VI.31}$$

we see that $R_{k,p}(\omega)$ is a monotone decreasing function with $R_{k,p}(0) = 0$. In addition,

$$\sum_{k\geq 1} R_{k,p}\left(\frac{\pi}{2}\right) = \sum_{k\geq 1}\left[\frac{1}{(2k+1)^{2p+1}} - \frac{1}{(2k-1)^{2p+1}}\right] = -1, \tag{VI.32}$$

so

$$1+\sum_{k\geq 1} R_{k,p+1}(\omega) \geq 1+\sum_{k\geq 1} R_{k,p}\left(\frac{\pi}{2}\right) = 0. \tag{VI.33}$$

In the following we prove that the sum of the remaining terms in (VI.30) is nonnegative as well, i.e.,

$$p\sum_{k\geq 1}\left[R_{k,p+1}(\omega) - R_{k,p}(\omega)\right] \geq 0.$$

From Lemmas VI.3.2 and VI.3.3 it follows that this is true for $\omega \in [0, \omega_p^*]$. Therefore, it remains to show that

$$S_p(\omega) = \sum_{k\geq 2}\left[R_{k,p+1}(\omega) - R_{k,p}(\omega)\right] \geq R_{1,p}(\omega) - R_{1,p+1}(\omega), \qquad \omega \in \left[\omega_p^*, \frac{\pi}{2}\right]. \tag{VI.34}$$

To this end, we first deduce from (VI.32) that

$$\sum_{k\geq 1}\left[R_{k,p+1}\left(\frac{\pi}{2}\right)-R_{k,p}\left(\frac{\pi}{2}\right)\right]=0,$$

implying

$$S_p\left(\frac{\pi}{2}\right)=R_{1,p}\left(\frac{\pi}{2}\right)-R_{1,p+1}\left(\frac{\pi}{2}\right)=\frac{1}{3^{2p+1}}-\frac{1}{3^{2p+3}}\geq 0.$$

Moreover, from (VI.31) we get

$$R'_{k,p}\left(\frac{\pi}{2}\right)=(2p+1)\frac{4k}{\pi}\left[\frac{1}{(2k+1)^{2p+2}}-\frac{1}{(2k-1)^{2p+2}}\right],$$

which gives [1]

$$\begin{aligned}
S'_p\left(\frac{\pi}{2}\right)&=R'_{2,p+1}\left(\frac{\pi}{2}\right)-R'_{2,p}\left(\frac{\pi}{2}\right)+\sum_{k\geq 3}\left[R'_{k,p+1}\left(\frac{\pi}{2}\right)-R'_{k,p}\left(\frac{\pi}{2}\right)\right]\\
&\leq\frac{16(p+1)}{3^{2p+2}\pi}+\frac{4}{\pi}\sum_{k\geq 3}\left[\frac{(2p+3)k}{(2k+1)^{2p+4}}+\frac{(2p+1)k}{(2k-1)^{2p+2}}\right]\\
&\leq\frac{16(p+1)}{3^{2p+2}\pi}+\frac{12}{5\pi}\sum_{k\geq 3}\left[\frac{2p+3}{(2k+1)^{2p+3}}+\frac{2p+1}{(2k-1)^{2p+1}}\right]\\
&\leq\frac{16(p+1)}{3^{2p+2}\pi}+\frac{12}{5\pi}\int_2^{+\infty}\left[\frac{2p+3}{(2\kappa+1)^{2p+3}}+\frac{2p+1}{(2\kappa-1)^{2p+1}}\right]\mathrm{d}\kappa\\
&=\frac{16(p+1)}{3^{2p+2}\pi}+\frac{6}{5\pi}\left[\frac{2p+3}{2(p+1)\,5^{2p+2}}+\frac{2p+1}{2p\,3^{2p}}\right]\\
&\leq\frac{p+1}{3^{2p-1}\pi}=m_p.
\end{aligned}$$

From Lemma VI.3.2 it follows that $S_p(\omega)$ is convex on $[0,\frac{\pi}{2}]$, so

$$S_p(\omega)\geq\left(\omega-\frac{\pi}{2}\right)m_p+S_p\left(\frac{\pi}{2}\right)=T_p(\omega). \tag{VI.35}$$

The straight line $T_p(\omega)$ vanishes at

$$\hat{\omega}_p=\frac{\pi}{2}-S_p\left(\frac{\pi}{2}\right)\frac{1}{m_p}=\frac{\pi}{2}-\left(\frac{1}{3^{2p+1}}-\frac{1}{3^{2p+3}}\right)\frac{3^{2p-1}\pi}{p+1}=\frac{\pi}{2}-\frac{8\pi}{81(p+1)},$$

and

$$\hat{\omega}_p=\frac{\pi}{2}\left(1-\frac{16}{81(p+1)}\right)<\omega_p^*.$$

From Lemma VI.3.3 we know that $R_{1,p}(\omega)-R_{1,p+1}(\omega)$ is convex on $[\omega_p^*,\frac{\pi}{2}]$, and so

$$R_{1,p}(\omega)-R_{1,p+1}(\omega)\leq T_p(\omega),\qquad\omega\in\left[\omega_p^*,\frac{\pi}{2}\right]. \tag{VI.36}$$

These functions are illustrated in Figure VI.1. By combining (VI.35) and (VI.36), we get (VI.34). □

We are now ready to prove Theorem IV.2.2.

[1] The equality holds due to the uniform convergence of the series.

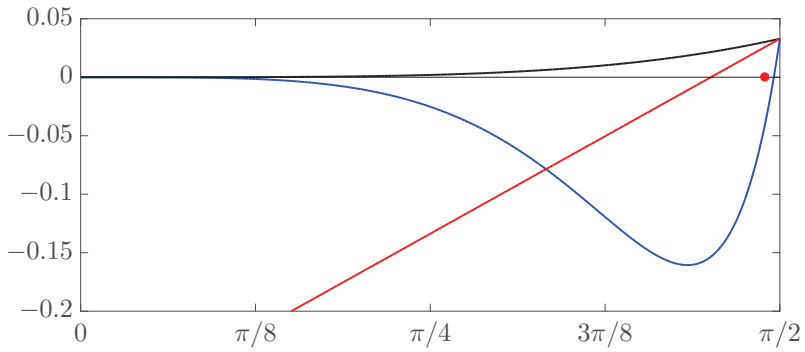


Figure VI.1: Graphs of $S_1(\omega)$ (black), $R_{1,1}(\omega) - R_{1,2}(\omega)$ (blue), and $T_1(\omega)$ (red). The value ω_1^* is marked with a red dot.

Proof of Theorem IV.2.2. Assume $\theta \in [0, \pi]$. From [52, Lemma A.2] we know that $g_p'(\theta) \leq 0$.[2] Moreover, by (IV.16)–(IV.17) we have $g_p(\theta), f_p(\theta) \geq 0$ and $f_p(\theta) = (2-2\cos(\theta))g_{p-1}(\theta)$. Finally, from the lower bound in (VI.15) we deduce that $f_p(\theta) \geq \theta^2 g_p(\theta)$. Therefore,

$$\begin{aligned} e_p'(\theta) &= \frac{f_p'(\theta)g_p(\theta) - f_p(\theta)g_p'(\theta)}{(g_p(\theta))^2} \geq \frac{f_p'(\theta) - \theta^2 g_p'(\theta)}{g_p(\theta)} \\ &= \frac{2\sin(\theta)g_{p-1}(\theta) + (2-2\cos(\theta))g_{p-1}'(\theta) - \theta^2 g_p'(\theta)}{g_p(\theta)}. \end{aligned}$$

This means that, in order to prove the monotonicity of e_p, it suffices to show that

$$G_p(\theta) = 2\sin(\theta)g_{p-1}(\theta) + (2-2\cos(\theta))g_{p-1}'(\theta) - \theta^2 g_p'(\theta) \geq 0, \qquad \theta \in [0, \pi]. \tag{VI.37}$$

From (IV.14) it follows that $g_p'(0) = g_p'(\pi) = 0$ for $p \geq 0$, so that $G_p(0) = G_p(\pi) = 0$ for $p \geq 1$. It remains to prove the inequality in (VI.37) for $\theta \in (0, \pi)$.

Let $\omega = \frac{\theta}{2} \in (0, \frac{\pi}{2})$. From [52, Proof of Lemma A.2] we know that

$$g_p(\theta) = \sum_{k\in\mathbb{Z}} \left(\frac{\sin(\omega)}{\omega + k\pi}\right)^{2p+2},$$

and

$$g_p'(\theta) = (p+1)(\sin(\omega))^{2p+1}\cos(\omega) \sum_{k\in\mathbb{Z}} \left[\frac{1}{(\omega+k\pi)^{2p+2}} - \frac{\tan(\omega)}{(\omega+k\pi)^{2p+3}}\right].$$

Therefore, recalling that $2 - 2\cos(\theta) = 4(\sin(\omega))^2$ and $\sin(\theta) = 2\sin(\omega)\cos(\omega)$, with some manipulations we obtain

$$\begin{aligned} G_p(\theta) = 4(\sin(\omega))^{2p+1} \Bigg(&\cos(\omega)(p+1) \sum_{k\in\mathbb{Z}} \frac{1}{(\omega+k\pi)^{2p}} \left[1 - \left(\frac{\omega}{\omega+k\pi}\right)^2\right] \\ &+ \sin(\omega) \sum_{k\in\mathbb{Z}} \frac{1}{(\omega+k\pi)^{2p+1}} \left[(p+1)\left(\frac{\omega}{\omega+k\pi}\right)^2 - p\right]\Bigg). \end{aligned} \tag{VI.38}$$

[2] Note that in [52] the function $g_p(\theta)$ is denoted by $h_p(\theta)$.

Considering the positivity of the first sum in (VI.38), it suffices to show that

$$\frac{4(\sin(\omega))^{2p+2}}{\omega^{2p+1}}\left(1+\sum_{k\geq 1}\left(\frac{\omega}{k\pi+\omega}\right)^{2p+1}\left[(p+1)\left(\frac{\omega}{k\pi+\omega}\right)^{2}-p\right]\right.$$
$$\left.-\sum_{k\geq 1}\left(\frac{\omega}{k\pi-\omega}\right)^{2p+1}\left[(p+1)\left(\frac{\omega}{k\pi-\omega}\right)^{2}-p\right]\right)\geq 0.$$

This inequality follows from (VI.30). □

VI.4 Proof of the IgA eigenvalue expansion for $\alpha = 0$

This section is devoted to the proof of the following theorem, that is, the expansion (IV.18) for $\alpha = 0$ and $j = 1, \ldots, N^{n,p} - (4p-2)$.

Theorem VI.4.1. *For every $p \geq 3$, every n, and every $j = 1, \ldots, N^{n,p} - (4p-2) = n - 3p$, we have*

$$\lambda_j(n^{-2}L_n^{[p]}) = e_p(\theta_{j,n}) + E_{j,n,0}^{[p]}, \tag{VI.39}$$

where:

- *the eigenvalues of $n^{-2}L_n^{[p]}$ are arranged in non decreasing order, $\lambda_1(n^{-2}L_n^{[p]}) \leq \ldots \leq \lambda_{n+p-2}(n^{-2}L_n^{[p]})$;*
- *e_p is the function defined in* (IV.15)*;*
- *$h = \frac{1}{n}$ and $\theta_{j,n} = \frac{j\pi}{n} = j\pi h$ for $j = 1, \ldots, n$;*
- *$|E_{j,n,0}^{[p]}| \leq C^{[p]}h$ for some constant $C^{[p]}$ depending only on p.*

Proof. Throughout this proof, we will use the simplified notations $N = N(p,n)$ and $\rho = 4p-2$. Moreover, we will write $V \subseteq_{\text{sp.}} \mathbb{C}^N$ to indicate that V is a vector subspace of $\mathbb{C}^N$. If A is an $N \times N$ matrix and $V \subseteq_{\text{sp.}} \mathbb{C}^N$, the symbol $A(V)$ will denote the subspace of $\mathbb{C}^N$ defined as $\{A\mathbf{x} : \mathbf{x} \in V\}$. Note that $A(V)$ has the same dimension as V whenever A is invertible.

We know from [115, Section 3] that

$$T_N(f_p) = \tau_N(f_p) + H_N(f_p), \tag{VI.40}$$
$$T_N(g_p) = \tau_N(g_p) + H_N(g_p), \tag{VI.41}$$

where, for any cosine trigonometric polynomial $\psi(\theta) = \psi_0 + 2\sum_{k=1}^{p}\psi_k\cos(k\theta)$,

- $\tau_N(\psi)$ is the tau matrix of order N generated by ψ, that is, the matrix in $\tau_N(0,0)$ defined as

$$\tau_N(\psi) = Q_N(0,0)\left(\operatorname*{diag}_{j=1,\ldots,N}\psi\left(\frac{j\pi}{N+1}\right)\right)Q_N(0,0);$$

- $H_N(\psi)$ is the Hankel matrix defined as

$$H_N(\psi) = \begin{bmatrix} \psi_2 & \psi_3 & \cdots & \psi_p & & & & \\ \psi_3 & & \iddots & & & & & \\ \vdots & \iddots & & & & & & \\ \psi_p & & & & & & & \\ & & & & & & & \psi_p \\ & & & & & & \iddots & \vdots \\ & & & & & \iddots & & \psi_3 \\ & & & & \psi_p & \cdots & \psi_3 & \psi_2 \end{bmatrix}.$$

Considering that $(H_N(f_p))_{ij} = (H_N(g_p))_{ij} = 0$ for $2p \le i \le N - 2p + 1 = n - p - 1$, in view of (IV.19)–(IV.22) we have

$$n^{-1}K_n^{[p]} = \tau_N(f_p) + \hat{R}_N^{[p]}, \tag{VI.42}$$

$$nM_n^{[p]} = \tau_N(g_p) + \hat{S}_N^{[p]}, \tag{VI.43}$$

where the rank corrections $\hat{R}_N^{[p]} = H_N(f_p) + R_N^{[p]}$ and $\hat{S}_N^{[p]} = H_N(g_p) + S_N^{[p]}$ satisfy

$$(\hat{R}_n^{[p]})_{ij} = 0, \quad 2p \le i \le n - p - 1 \quad \Longrightarrow \quad \operatorname{rank}(\hat{R}_N^{[p]}) \le \rho, \tag{VI.44}$$

$$(\hat{S}_n^{[p]})_{ij} = 0, \quad 2p \le i \le n - p - 1 \quad \Longrightarrow \quad \operatorname{rank}(\hat{S}_N^{[p]}) \le \rho. \tag{VI.45}$$

Since $M_n^{[p]}$ is symmetric positive definite and $L_n^{[p]} = (M_n^{[p]})^{-1}K_n^{[p]}$ is similar to

$$(M_n^{[p]})^{-1/2}K_n^{[p]}(M_n^{[p]})^{-1/2},$$

by the minimax principle for the eigenvalues of Hermitian matrices [13] we have, for every $j = 1, \dots, N$,

$$\begin{aligned} \lambda_j(n^{-2}L_n^{[p]}) &= \lambda_j(n^{-2}(M_n^{[p]})^{-1/2}K_n^{[p]}(M_n^{[p]})^{-1/2}) \\ &= \max_{\substack{V \subseteq_{\mathrm{sp.}} \mathbb{C}^N \\ \dim V = N-j+1}} \min_{\substack{\mathbf{x} \in V \\ \mathbf{x} \neq \mathbf{0}}} \frac{n^{-2}\mathbf{x}^*(M_n^{[p]})^{-1/2}K_n^{[p]}(M_n^{[p]})^{-1/2}\mathbf{x}}{\mathbf{x}^*\mathbf{x}} \\ &= \max_{\substack{V \subseteq_{\mathrm{sp.}} \mathbb{C}^N \\ \dim V = N-j+1}} \min_{\substack{\mathbf{y} \in (M_n^{[p]})^{-1/2}(V) \\ \mathbf{y} \neq \mathbf{0}}} \frac{n^{-2}\mathbf{y}^*K_n^{[p]}\mathbf{y}}{\mathbf{y}^*M_n^{[p]}\mathbf{y}} \\ &= \max_{\substack{U \subseteq_{\mathrm{sp.}} \mathbb{C}^N \\ \dim U = N-j+1}} \min_{\substack{\mathbf{y} \in U \\ \mathbf{y} \neq \mathbf{0}}} \frac{\mathbf{y}^*(n^{-1}K_n^{[p]})\mathbf{y}}{\mathbf{y}^*(nM_n^{[p]})\mathbf{y}}. \end{aligned} \tag{VI.46}$$

Let F be the subspace of $\mathbb{C}^N$ generated by the union of the nonzero columns of $\hat{R}_n^{[p]}$ and $\hat{S}_n^{[p]}$. By (VI.44)–(VI.45), we have $\dim F \le \rho$ and, consequently, $\dim F^\perp \ge N - \rho$. Moreover, if U is any subspace of $\mathbb{C}^N$ such that $\dim U = u$, we have $\dim(U \cap F^\perp) = \dim U + \dim F^\perp - \dim(U + F^\perp) \ge$

$u + (N-\rho) - N = u - \rho$. Thus, for $j = 1, \dots, N-\rho$, from (VI.42)–(VI.43) and (VI.46) we obtain

$$\begin{aligned}
\lambda_j(n^{-2}L_n^{[p]}) &\le \max_{\substack{U \subseteq_{\mathrm{sp.}} \mathbb{C}^N \\ \dim U = N-j+1}} \min_{\substack{\mathbf{y} \in U \cap F^\perp \\ \mathbf{y} \neq \mathbf{0}}} \frac{\mathbf{y}^*(\tau_N(f_p) + \hat{R}_n^{[p]})\mathbf{y}}{\mathbf{y}^*(\tau_N(g_p) + \hat{S}_n^{[p]})\mathbf{y}} \\
&= \max_{\substack{U \subseteq_{\mathrm{sp.}} \mathbb{C}^N \\ \dim U = N-j+1}} \min_{\substack{\mathbf{y} \in U \cap F^\perp \\ \mathbf{y} \neq \mathbf{0}}} \frac{\mathbf{y}^*\tau_N(f_p)\mathbf{y}}{\mathbf{y}^*\tau_N(g_p)\mathbf{y}} \\
&\le \max_{\substack{W \subseteq_{\mathrm{sp.}} \mathbb{C}^N \\ \dim W \ge N-(j+\rho)+1}} \min_{\substack{\mathbf{y} \in W \\ \mathbf{y} \neq \mathbf{0}}} \frac{\mathbf{y}^*\tau_N(f_p)\mathbf{y}}{\mathbf{y}^*\tau_N(g_p)\mathbf{y}} \\
&= \max_{\substack{W \subseteq_{\mathrm{sp.}} \mathbb{C}^N \\ \dim W \ge N-(j+\rho)+1}} \min_{\substack{\mathbf{x} \in (\tau_N(g_p))^{1/2}(W) \\ \mathbf{x} \neq \mathbf{0}}} \frac{\mathbf{x}^*(\tau_N(g_p))^{-1/2}\tau_N(f_p)(\tau_N(g_p))^{-1/2}\mathbf{x}}{\mathbf{x}^*\mathbf{x}} \\
&= \max_{\substack{V \subseteq_{\mathrm{sp.}} \mathbb{C}^N \\ \dim V \ge N-(j+\rho)+1}} \min_{\substack{\mathbf{x} \in V \\ \mathbf{x} \neq \mathbf{0}}} \frac{\mathbf{x}^*\tau_N(e_p)\mathbf{x}}{\mathbf{x}^*\mathbf{x}} \\
&= \max_{\substack{V \subseteq_{\mathrm{sp.}} \mathbb{C}^N \\ \dim V = N-(j+\rho)+1}} \min_{\substack{\mathbf{x} \in V \\ \mathbf{x} \neq \mathbf{0}}} \frac{\mathbf{x}^*\tau_N(e_p)\mathbf{x}}{\mathbf{x}^*\mathbf{x}} \\
&= \lambda_{j+\rho}(\tau_N(e_p)) = e_p\Big(\frac{(j+\rho)\pi}{N+1}\Big), \qquad \text{(VI.47)}
\end{aligned}$$

where the last equality is because of the monotonicity of e_p (Theorem IV.2.2). Similarly, using again the minimax principle for Hermitian matrices, for $j = \rho+1, \dots, N$ we obtain

$$\begin{aligned}
\lambda_j(n^{-2}L_n^{[p]}) &= \lambda_j(n^{-2}(M_n^{[p]})^{-1/2}K_n^{[p]}(M_n^{[p]})^{-1/2}) \\
&= \min_{\substack{V \subseteq_{\mathrm{sp.}} \mathbb{C}^N \\ \dim V = j}} \max_{\substack{\mathbf{x} \in V \\ \mathbf{x} \neq \mathbf{0}}} \frac{n^{-2}\mathbf{x}^*(M_n^{[p]})^{-1/2}K_n^{[p]}(M_n^{[p]})^{-1/2}\mathbf{x}}{\mathbf{x}^*\mathbf{x}} \\
&= \min_{\substack{V \subseteq_{\mathrm{sp.}} \mathbb{C}^N \\ \dim V = j}} \max_{\substack{\mathbf{y} \in (M_n^{[p]})^{-1/2}(V) \\ \mathbf{y} \neq \mathbf{0}}} \frac{n^{-2}\mathbf{y}^*K_n^{[p]}\mathbf{y}}{\mathbf{y}^*M_n^{[p]}\mathbf{y}} \\
&= \min_{\substack{U \subseteq_{\mathrm{sp.}} \mathbb{C}^N \\ \dim U = j}} \max_{\substack{\mathbf{y} \in U \\ \mathbf{y} \neq \mathbf{0}}} \frac{\mathbf{y}^*(n^{-1}K_n^{[p]})\mathbf{y}}{\mathbf{y}^*(nM_n^{[p]})\mathbf{y}} \\
&\ge \min_{\substack{U \subseteq_{\mathrm{sp.}} \mathbb{C}^N \\ \dim U = j}} \max_{\substack{\mathbf{y} \in U \cap F^\perp \\ \mathbf{y} \neq \mathbf{0}}} \frac{\mathbf{y}^*(\tau_N(f_p) + \hat{R}_n^{[p]})\mathbf{y}}{\mathbf{y}^*(\tau_N(g_p) + \hat{S}_n^{[p]})\mathbf{y}} \\
&= \min_{\substack{U \subseteq_{\mathrm{sp.}} \mathbb{C}^N \\ \dim U = j}} \max_{\substack{\mathbf{y} \in U \cap F^\perp \\ \mathbf{y} \neq \mathbf{0}}} \frac{\mathbf{y}^*\tau_N(f_p)\mathbf{y}}{\mathbf{y}^*\tau_N(g_p)\mathbf{y}} \\
&\ge \min_{\substack{W \subseteq_{\mathrm{sp.}} \mathbb{C}^N \\ \dim W \ge j-\rho}} \max_{\substack{\mathbf{y} \in W \\ \mathbf{y} \neq \mathbf{0}}} \frac{\mathbf{y}^*\tau_N(f_p)\mathbf{y}}{\mathbf{y}^*\tau_N(g_p)\mathbf{y}} \\
&= \min_{\substack{W \subseteq_{\mathrm{sp.}} \mathbb{C}^N \\ \dim W \ge j-\rho}} \max_{\substack{\mathbf{x} \in (\tau_N(g_p))^{1/2}(W) \\ \mathbf{x} \neq \mathbf{0}}} \frac{\mathbf{x}^*(\tau_N(g_p))^{-1/2}\tau_N(f_p)(\tau_N(g_p))^{-1/2}\mathbf{x}}{\mathbf{x}^*\mathbf{x}}
\end{aligned}$$

$$
\begin{aligned}
&= \min_{\substack{V\subseteq_{\mathrm{sp.}}\mathbb{C}^N \\ \dim V\geq j-\rho}} \max_{\substack{\mathbf{x}\in V \\ \mathbf{x}\neq\mathbf{0}}} \frac{\mathbf{x}^*\tau_N(e_p)\mathbf{x}}{\mathbf{x}^*\mathbf{x}} \\
&= \min_{\substack{V\subseteq_{\mathrm{sp.}}\mathbb{C}^N \\ \dim V=j-\rho}} \max_{\substack{\mathbf{x}\in V \\ \mathbf{x}\neq\mathbf{0}}} \frac{\mathbf{x}^*\tau_N(e_p)\mathbf{x}}{\mathbf{x}^*\mathbf{x}} \\
&= \lambda_{j-\rho}(\tau_N(e_p)) = e_p\Big(\frac{(j-\rho)\pi}{N+1}\Big). \qquad \text{(VI.48)}
\end{aligned}
$$

Putting together (VI.47) and (VI.48), we get

$$
e_p\Big(\frac{(j-\rho)\pi}{N+1}\Big) \leq \lambda_j(n^{-2}L_n^{[p]}) \leq e_p\Big(\frac{(j+\rho)\pi}{N+1}\Big), \qquad j=\rho+1,\dots,N-\rho. \qquad \text{(VI.49)}
$$

From (VI.49) we immediately obtain

$$
\begin{aligned}
&\Big|\lambda_j(n^{-2}L_n^{[p]}) - e_p\Big(\frac{j\pi}{N+1}\Big)\Big| \qquad \text{(VI.50)} \\
&\leq \max\Big(\Big|e_p\Big(\frac{(j-\rho)\pi}{N+1}\Big) - e_p\Big(\frac{j\pi}{N+1}\Big)\Big|, \Big|e_p\Big(\frac{(j+\rho)\pi}{N+1}\Big) - e_p\Big(\frac{j\pi}{N+1}\Big)\Big|\Big) \\
&\leq \|e_p'\|_\infty \frac{\rho\pi}{N+1} \leq \|e_p'\|_\infty \rho\pi h, \qquad j=\rho+1,\dots,N-\rho. \qquad \text{(VI.51)}
\end{aligned}
$$

Moreover, since the eigenvalues of $n^{-2}L_n^{[p]}$ are positive (because of the similarity between $L_n^{[p]}$ and the symmetric positive definite matrix $(M_n^{[p]})^{-1/2}K_n^{[p]}(M_n^{[p]})^{-1/2}$) and $e_p(0) = 0 = \min_{\theta\in[0,\pi]} e_p(\theta)$ (by (IV.16)–(IV.17)), for $j = 1,\dots,\rho$ we have

$$
\begin{aligned}
&\Big|\lambda_j(n^{-2}L_n^{[p]}) - e_p\Big(\frac{j\pi}{N+1}\Big)\Big| = \qquad \text{(VI.52)} \\
&= \begin{cases} \lambda_j(n^{-2}L_n^{[p]}) - e_p\Big(\frac{j\pi}{N+1}\Big), & \text{if } \lambda_j(n^{-2}L_n^{[p]}) - e_p\Big(\frac{j\pi}{N+1}\Big) \geq 0, \\ e_p\Big(\frac{j\pi}{N+1}\Big) - \lambda_j(n^{-2}L_n^{[p]}), & \text{otherwise,} \end{cases} \\
&\leq \begin{cases} \lambda_{\rho+1}(n^{-2}L_n^{[p]}) - e_p\Big(\frac{j\pi}{N+1}\Big), & \text{if } \lambda_j(n^{-2}L_n^{[p]}) - e_p\Big(\frac{j\pi}{N+1}\Big) \geq 0, \\ e_p\Big(\frac{j\pi}{N+1}\Big), & \text{otherwise,} \end{cases} \\
&\leq \begin{cases} \Big|\lambda_{\rho+1}(n^{-2}L_n^{[p]}) - e_p\Big(\frac{(\rho+1)\pi}{N+1}\Big)\Big| + e_p\Big(\frac{(\rho+1)\pi}{N+1}\Big) - e_p\Big(\frac{j\pi}{N+1}\Big), & \\ \qquad\qquad\qquad\qquad \text{if } \lambda_j(n^{-2}L_n^{[p]}) - e_p\Big(\frac{j\pi}{N+1}\Big) \geq 0, & \\ e_p\Big(\frac{\rho\pi}{N+1}\Big) - e_p(0), & \text{otherwise,} \end{cases} \\
&\leq \begin{cases} \|e_p'\|_\infty\rho\pi h + \|e_p'\|_\infty\rho\pi h, & \text{if } \lambda_j(n^{-2}L_n^{[p]}) - e_p\Big(\frac{j\pi}{N+1}\Big) \geq 0, \\ \|e_p'\|_\infty\rho\pi h, & \text{otherwise,} \end{cases} \\
&\leq 2\|e_p'\|_\infty\rho\pi h. \qquad \text{(VI.53)}
\end{aligned}
$$

Combining (VI.50) and (VI.53), we obtain

$$
\Big|\lambda_j(n^{-2}L_n^{[p]}) - e_p\Big(\frac{j\pi}{N+1}\Big)\Big| \leq 2\|e_p'\|_\infty\rho\pi h, \qquad j=1,\dots,N-\rho. \qquad \text{(VI.54)}
$$

To conclude the proof, we note that the stepsizes $h = \frac{1}{n}$ and $H = \frac{1}{N+1}$ are such that

$$0 < h - H = \frac{N+1-n}{n(N+1)} = \frac{p-1}{n(n+p-1)} < \frac{p}{n^2}$$

and, consequently, the grid points $\theta_{j,n} = j\pi h$ and $\Theta_{j,n} = j\pi H$ satisfy

$$0 < \theta_{j,n} - \Theta_{j,n} < \frac{p\pi}{n}, \qquad j = 1, \dots, n.$$

Thus, the inequality (VI.54) yields the thesis (VI.39) with

$$\begin{aligned} |E^{[p]}_{j,n,0}| &= |\lambda_j(n^{-2}L^{[p]}_n) - e_p(\theta_{j,n})| \le |\lambda_j(n^{-2}L^{[p]}_n) - e_p(\Theta_{j,n})| + |e_p(\Theta_{j,n}) - e_p(\theta_{j,n})| \\ &\le 2\|e'_p\|_\infty \rho\pi h + \|e'_p\|_\infty p\pi h = C^{[p]}h, \qquad j = 1, \dots, N-\rho, \end{aligned}$$

where $C^{[p]} = (2\rho + p)\pi\|e'_p\|_\infty$. □

VI.5 $\mathbb{Q}_p$ Lagrangian FEM matrix symbol for $p = 2, 3, 4$

Recall that the $p \times p$ matrix-valued symbols of $K^{(p)}_n$ and $M^{(p)}_n$ are

$$\mathbf{f}(\theta) = \hat{\mathbf{f}}_0 + \hat{\mathbf{f}}_1 e^{\iota\theta} + \hat{\mathbf{f}}_1^{\mathrm{T}} e^{-\iota\theta}$$

and

$$\mathbf{g}(\theta) = \hat{\mathbf{g}}_0 + \hat{\mathbf{g}}_1 e^{\iota\theta} + \hat{\mathbf{g}}_1^{\mathrm{T}} e^{-\iota\theta}$$

respectively. The detailed expressions of $\hat{\mathbf{f}}_0$, $\hat{\mathbf{f}}_1$ and $\hat{\mathbf{g}}_0$, $\hat{\mathbf{g}}_1$ for the particular degrees $p = 2, 3, 4$ are given below.
For $p = 2$,

$$\hat{\mathbf{f}}_0 = \frac{1}{3}\begin{bmatrix} 16 & -8 \\ -8 & 14 \end{bmatrix}, \quad \hat{\mathbf{f}}_1 = \frac{1}{3}\begin{bmatrix} 0 & -8 \\ 0 & 1 \end{bmatrix},$$

$$\hat{\mathbf{g}}_0 = \frac{1}{30}\begin{bmatrix} 16 & 2 \\ 2 & 8 \end{bmatrix}, \quad \hat{\mathbf{g}}_1 = \frac{1}{30}\begin{bmatrix} 0 & 2 \\ 0 & -1 \end{bmatrix}.$$

For $p = 3$,

$$\hat{\mathbf{f}}_0 = \frac{1}{40}\begin{bmatrix} 432 & -297 & 54 \\ -297 & 432 & -189 \\ 54 & -189 & 296 \end{bmatrix}, \quad \hat{\mathbf{f}}_1 = \frac{1}{40}\begin{bmatrix} 0 & 0 & -189 \\ 0 & 0 & 54 \\ 0 & 0 & -13 \end{bmatrix},$$

$$\hat{\mathbf{g}}_0 = \frac{1}{1680}\begin{bmatrix} 648 & -81 & -36 \\ -81 & 648 & 99 \\ -36 & 99 & 256 \end{bmatrix}, \quad \hat{\mathbf{g}}_1 = \frac{1}{1680}\begin{bmatrix} 0 & 0 & 99 \\ 0 & 0 & -36 \\ 0 & 0 & 19 \end{bmatrix}.$$

For $p = 4$,

$$\hat{\mathbf{f}}_0 = \frac{1}{945}\begin{bmatrix} 16640 & -14208 & 5888 & -1472 \\ -14208 & 22320 & -14208 & 3048 \\ 5888 & -14208 & 16640 & -6848 \\ -1472 & 3048 & -6848 & 9850 \end{bmatrix}, \quad \hat{\mathbf{f}}_1 = \frac{1}{945}\begin{bmatrix} 0 & 0 & 0 & -6848 \\ 0 & 0 & 0 & 3048 \\ 0 & 0 & 0 & -1472 \\ 0 & 0 & 0 & 347 \end{bmatrix},$$

$$\hat{\mathbf{g}}_0 = \frac{1}{5670}\begin{bmatrix} 1792 & -384 & 256 & 56 \\ -384 & 1872 & -384 & -174 \\ 256 & -384 & 1792 & 296 \\ 56 & -174 & 296 & 584 \end{bmatrix}, \quad \hat{\mathbf{g}}_1 = \frac{1}{5670}\begin{bmatrix} 0 & 0 & 0 & 296 \\ 0 & 0 & 0 & -174 \\ 0 & 0 & 0 & 56 \\ 0 & 0 & 0 & -29 \end{bmatrix}.$$

VI.6 Proof of the block eigenvalue expansion for $\alpha = 0$

Theorem VI.6.1. *Let $s > 1$, $N = N(n, s) = sn$ and $\mathbf{f}$ be an Hermitian matrix-valued trigonometric polynomial (HTP) with Fourier coefficients $\hat{\mathbf{f}}_0, \hat{\mathbf{f}}_1, \ldots, \hat{\mathbf{f}}_m \in \mathbb{R}^{s\times s}$. Suppose that $\mathbf{f}$ is of the form*

$$\mathbf{f}(\theta) = \sum_{k=-m}^{m} \hat{\mathbf{f}}_k \mathrm{e}^{\iota k\theta} = \hat{\mathbf{f}}_0 + \sum_{k=1}^{m} \left(\hat{\mathbf{f}}_k \mathrm{e}^{\iota k\theta} + \hat{\mathbf{f}}_k^T \mathrm{e}^{-\mathbf{i}k\theta}\right), \qquad m = \deg\left(\mathbf{f}(\theta)\right) \in \mathbb{N},$$

such that

$$\hat{\mathbf{f}}_{-k} = \hat{\mathbf{f}}_k^T \quad k = 0, \ldots, m. \tag{VI.55}$$

Suppose that the eigenvalue functions of $\mathbf{f}$, $\lambda^{(q)}(\mathbf{f}) : [0, \pi] \to \mathbb{R}^{s\times s}$, $q = 1, \ldots, s$, are monotone on $[0, \pi]$ and such that

$$\max_{\theta\in[0,\pi]} \lambda^{(q)}(\mathbf{f}) < \min_{\theta\in[0,\pi]} \lambda^{(q+1)}(\mathbf{f}) \tag{VI.56}$$

$q = 1, \ldots, s-1$, then, fixed $q \in \{1, \ldots, s\}$,

$$\left|\lambda_\gamma(T_n(\mathbf{f})) - \lambda^{(q)}\left(\mathbf{f}\left(\theta_{j,n}\right)\right)\right| \le Ch \tag{VI.57}$$

$\forall\, n$, for $j = 1, \ldots, n$, and $\gamma = \gamma(q, j) = (q-1)n + j$, where

- *$\lambda_\gamma(T_n(\mathbf{f}))$, $\gamma \in \{1, \ldots, N\}$, are the eigenvalues of $T_n(\mathbf{f})$, such that, for a fixed $\bar{q} \in \{1, \ldots, s\}$, $\lambda_{(\bar{q}-1)n+j}(T_n(\mathbf{f}))$ are arranged in non decreasing or non increasing order, depending on whether $\lambda^{(\bar{q})}(\mathbf{f})$ is increasing or decreasing.*
- *$h = \frac{1}{n+1}$ and $\theta_{j,n} = \frac{j\pi}{n+1} = j\pi h$, $j = 1, \ldots, n$;*

Proof. For the sake of simplicity, we assume that for $q = 1, \ldots, s$, $\lambda^{(q)}(\mathbf{f})$ is monotone non decreasing (the other cases have a similar proof).
Notice that the conditions on $\mathbf{f}$ imply that the $N \times N$ block Toeplitz matrix generated by $\mathbf{f}$, $T_n(\mathbf{f})$, is Hermitian positive definite so we can order its eigenvalues in non decreasing order of as follows

$$\left\{\left\{\lambda_{(q-1)n+j}(T_n(\mathbf{f}))\right\}_{j=1}^{n}\right\}_{q=1}^{s} \tag{VI.58}$$

We remark from the relation I.13 of Section I.5 that

$$T_n(\mathbf{f}) = \tau_N(\mathbf{f}) + H_N(\mathbf{f}), \tag{VI.59}$$

where, for ψ (HTP) of degree m and orthogonal $Q = \left(\sqrt{\frac{2}{n+1}}\sin\left(\frac{ij\pi}{n+1}\right)\right)_{i,j=1}^{n}$, $\tau_N(\psi)$ is the following τ matrix [14] of size N generated by ψ

$$\tau_N(\psi) = (Q \otimes I_s)\operatorname*{diag}_{1\le j\le n}\left(\psi\left(\frac{j\pi}{n+1}\right)\right)(Q \otimes I_s), \quad Q = Q^T = Q^{-1},$$

where $H_N(\psi)$ is the Hankel matrix associated to ψ with $\nu := \nu(s,m) = \operatorname{rank}(H_N(\psi)) \le 2s(m-1)$.
For $q = 1,\dots,s$, $j = 1,\dots,n$, setting $\gamma = (q-1)n + j$, we find

$$\lambda_\gamma(\tau_N(\mathbf{f})) = \lambda^{(q)}\left(\mathbf{f}\left(\frac{j\pi}{n+1}\right)\right). \tag{VI.60}$$

Note that $T_n(\mathbf{f})$ is similar to the matrix

$$\begin{aligned}\tilde{T}_n(\mathbf{f}) &= (Q \otimes I_s)T_n(\mathbf{f})(Q \otimes I_s)\\ &= \operatorname*{diag}_{1\le j\le n}\left(\mathbf{f}\left(\frac{j\pi}{n+1}\right)\right) + (Q \otimes I_s)H_N(\mathbf{f})(Q \otimes I_s)\\ &= \operatorname*{diag}_{1\le j\le n}\left(\mathbf{f}\left(\frac{j\pi}{n+1}\right)\right) + \tilde{H}_\nu,\end{aligned}$$

with $\operatorname{rank}(\tilde{H}_\nu) = \nu$, so $T_n(\mathbf{f})$ and $\tilde{T}_n(\mathbf{f})$ have the same eigenvalues.

Using the MinMax spectral characterization for Hermitian matrices [13], we obtain, for $\gamma = (q-1)n + j \in \{\nu+1,\dots,N-\nu\}$,

$$\lambda_{\gamma-\nu}(\tau_N(\mathbf{f})) = \lambda^{(q)}\left(\mathbf{f}\left(\frac{(j-\nu)\pi}{n+1}\right)\right) \le \lambda_\gamma(T_n(\mathbf{f})) \le \lambda_{\gamma+\nu}(\tau_N(\mathbf{f})) = \lambda^{(q)}\left(\mathbf{f}\left(\frac{(j+\nu)\pi}{n+1}\right)\right). \tag{VI.61}$$

The eigenvalue functions $\lambda^{(q)}(\mathbf{f})$ are monotone non decreasing function so we have, $\forall\, n$ and for $\gamma = (q-1)n + j \in \{\nu+1,\dots,N-\nu\}$,

$$\begin{aligned}&\lambda_\gamma(T_n(\mathbf{f})) - \lambda^{(q)}\left(\mathbf{f}\left(\frac{j\pi}{n+1}\right)\right) \le \lambda^{(q)}\left(\mathbf{f}\left(\frac{(j+\nu)\pi}{n+1}\right)\right) - \lambda^{(q)}\left(\mathbf{f}\left(\frac{j\pi}{n+1}\right)\right) =\\ &= \left(\lambda^{(q)}(\mathbf{f}(\bar{\theta}))\right)'\frac{\nu\pi}{n+1} \le \left\|\left(\lambda^{(q)}(\mathbf{f})\right)'\right\|_\infty \frac{\nu\pi}{n+1},\end{aligned} \tag{VI.62}$$

with $\bar{\theta} \in \left(\frac{j\pi}{n+1}, \frac{(j+\nu)\pi}{n+1}\right)$ and

$$\begin{aligned}\lambda_\gamma(T_n(\mathbf{f})) - \lambda^{(q)}\left(\mathbf{f}\left(\frac{j\pi}{n+1}\right)\right) &\ge \lambda^{(q)}\left(\mathbf{f}\left(\frac{(j-\nu)\pi}{n+1}\right)\right) - \lambda^{(q)}\left(\mathbf{f}\left(\frac{j\pi}{n+1}\right)\right) \ge\\ &\ge -\left\|\left(\lambda^{(q)}(\mathbf{f})\right)'\right\|_\infty \frac{\nu\pi}{n+1}.\end{aligned} \tag{VI.63}$$

By setting $C = \left\| \left(\lambda^{(q)}(\mathbf{f})\right)' \right\|_\infty \nu\pi$, for $\gamma = (q-1)n + j \in \{\nu+1, \dots, N-\nu\}$, we obtain

$$\left| \lambda_\gamma(T_n(\mathbf{f})) - \lambda^{(q)}\left(\mathbf{f}\left(\frac{j\pi}{n+1}\right)\right) \right| \leq Ch. \tag{VI.64}$$

Furthermore, from [47] $\forall\, \gamma = 1, \dots, N$, we know that

$$m_{\mathbf{f}} \leq \lambda_\gamma(T_n(\mathbf{f})) \leq M_{\mathbf{f}},$$

where

$$m_{\mathbf{f}} = \min_{\theta\in[0,\pi]} \left(\lambda^{(1)}(\mathbf{f}(\theta))\right); \quad M_{\mathbf{f}} = \max_{\theta\in[0,\pi]} \left(\lambda^{(s)}(\mathbf{f}(\theta))\right),$$

with strict inequalities that is $m_{\mathbf{f}} < \lambda_\gamma(T_n(\mathbf{f})) < M_{\mathbf{f}}$ since, by the assumptions, the extreme eigenvalue functions are not constant. Hence for $N - \nu < \gamma \leq N$

$$\begin{aligned} \left| \lambda^{(s)}\left(\mathbf{f}\left(\frac{j\pi}{n+1}\right)\right) - \lambda_\gamma(T_n(\mathbf{f})) \right| &\leq \left| \lambda^{(s)}\left(\mathbf{f}\left(\frac{j\pi}{n+1}\right)\right) - \lambda^{(s)}\left(\mathbf{f}\left(\frac{n\pi}{n+1}\right)\right) \right| \\ &\leq \left| \left(\lambda^{(s)}\left(\mathbf{f}\left(\bar{\theta}\right)\right)\right)' \right| \left| \frac{(n-j)\pi}{n+1} \right|, \end{aligned}$$

where $\bar{\theta} \in \left(\frac{j\pi}{n+1}, \frac{n\pi}{n+1}\right)$. If $N - \nu < \gamma \leq N$ then $|N-\nu| < |(s-1)n + j| \to |n-j| < \nu$, so that

$$\left| \lambda^{(s)}\left(\mathbf{f}\left(\frac{j\pi}{n+1}\right)\right) - \lambda_\gamma(T_n(\mathbf{f})) \right| \leq \left\| \left(\lambda^{(s)}\left(\mathbf{f}\right)\right)' \right\|_\infty \frac{\nu\pi}{n+1} = Ch.$$

For $1 \leq \gamma < \nu + 1$

$$\begin{aligned} \left| \lambda^{(1)}\left(\mathbf{f}\left(\frac{j\pi}{n+1}\right)\right) - \lambda_\gamma(T_n(\mathbf{f})) \right| &\leq \left| \lambda^{(1)}\left(\mathbf{f}\left(\frac{j\pi}{n+1}\right)\right) - \lambda^{(1)}\left(\mathbf{f}\left(\frac{\pi}{n+1}\right)\right) \right| \\ &\leq \left| \left(\lambda^{(1)}\left(\mathbf{f}\left(\bar{\theta}\right)\right)\right)' \right| \left| \frac{(j-1)\pi}{n+1} \right|, \end{aligned}$$

where $\bar{\theta} \in \left(\frac{\pi}{n+1}, \frac{j\pi}{n+1}\right)$. If $1 \leq \gamma < \nu + 1$ then $|j| > |\nu + 1| \Rightarrow |j-1| < \nu$, so

$$\left| \lambda^{(1)}\left(\mathbf{f}\left(\frac{j\pi}{n+1}\right)\right) - \lambda_\gamma(T_n(\mathbf{f})) \right| \leq \left\| \left(\lambda^{(1)}\left(\mathbf{f}\right)\right)' \right\|_\infty \frac{\nu\pi}{n+1} = Ch.$$

Hence for $q = 1, \dots, s$, $j = 1, \dots, n$, $\gamma = (q-1)n + j \in \{1, \dots, N\}$,

$$\left| \lambda_\gamma(T_n(\mathbf{f})) - \lambda^{(q)}\left(\mathbf{f}\left(\frac{j\pi}{n+1}\right)\right) \right| \leq Ch.$$

□

Remark 12. *With regard to Theorem VI.6.1, for $q = 1, \dots, s$, the case where $\lambda^{(q)}(\mathbf{f})$ are bounded and non-monotone is almost analogous. If we consider $\hat{\lambda}^{(q)}(\mathbf{f})$, the monotone non decreasing rearrangement of $\lambda^{(q)}(\mathbf{f})$ on $[0, \pi]$, taking into account that the derivative of $\lambda^{(q)}(\mathbf{f})$ has at most a finite number S of sign changes, we deduce that $\hat{\lambda}^{(q)}(\mathbf{f})$ is Lipschitz continuous and its Lipschitz constant is bounded by $\| \left(\lambda^{(q)}(\mathbf{f})\right)' \|_\infty$ (notice that $\hat{\lambda}^{(q)}(\mathbf{f})$ is not necessarily*

continuously differentiable but the derivative of $\hat{\lambda}^{(q)}(\mathbf{f})$ *has at most* S *points of discontinuity). Furthermore the eigenvalues* $\lambda_\gamma(\tau_N(\mathbf{f}))$*, are exactly given by*

$$\lambda^{(q)}\left(\mathbf{f}\left(\frac{j\pi}{n+1}\right)\right), \quad q=1,\ldots,s \quad j=1,\ldots,n,$$

so that, by ordering these values non decreasingly, we deduce that they coincide with $\hat{\lambda}^{(q)}\left(\mathbf{f}\left(x_{j,n}\right)\right)$*, with* $x_{j,n}$ *of the form* $\frac{j\pi}{n+1}(1+o(1))$*. With these premises, the proof follows exactly the same steps as in Theorem VI.6.1, using the MinMax characterization and the Interlacing theorem for Hermitian matrices.*

Conclusions

In most of the applications the interest in studying the spectral properties of structured matrix sequences is two-fold. In fact, on one hand there are problems in which the information on the eigenvalues are indirectly useful in finding efficiently the numerical solution, on the other hand there are situations (for example, this is the case of eigenvalue problems [42, 92]) where the eigenvalues have a physical meaning or represent the approximation of a quantity of interest.

These reasons, and many others, make the research of more and more efficient eigensolvers relevant and topical.

This thesis faces up to the mentioned double requests with a double strategy. It presents both several standard issues treated with a new class of techniques, and few novel computational problems never solved with classical tools.

In particular this is the case of the **Chapter II** where for the first time the spectral analysis with GLT techniques is applied to the recent discretization by the novel family of high order accurate Discontinuous Galerkin (DG) methods on *staggered* meshes.

On the other direction, **Chapter III, IV, V** are devoted to present new fast extrapolation–interpolation methods for computing the approximation of the spectrum of large Toeplitz and Toeplitz-like sequences in various settings.

The future purpose will be to combine the two strategy and provide new useful tools to deal with new computational problems and those arising from some recent discretization techniques.

A first achievement can be obtained from the possible future developments of the topics treated in **Chapter II**.

We have have studied in detail the resulting (structured) matrices coming from the discretization by staggered DG methods of the incompressible Navier-Stokes equations. The classical theory of Toeplitz matrices generated by a function (in the most general block, multilevel form) and the more recent theory of GLT matrix-sequences have been the key tools for analyzing the spectral properties of the considered large matrices. We have obtained a quite complete picture of the spectral properties of the underlying linear systems that result after the discretization of the PDE. This information has been employed for giving a forecast of the convergence history of the CG method and for proposing a basic, still effective, Strang-type block circulant preconditioner and for designing the essentials of the Two grid technique.

Starting from the preliminary findings in Subsection II.3.3 and Subsection II.3.4, the use of these results will be the ground for further research in the direction of new more advanced techniques (involving preconditioning, multigrid, multi-iterative solvers [113]), by taking into account variable coefficients, compressibility, graded meshes in geometrically complex domains, and various boundary conditions.

Furthermore there are possible other situations where the multilevel block matrix sequences are involved and the proposed analysis could be similarly applied. This is the case of the structured matrix sequences arising from some different PDEs discretization, e.g., the Virtual Element Methods, or the optimal control problems that will be subjects of future investigation.

On the other hand the second goal of this thesis have been to provide new tools for computing the spectrum of:

1. preconditioned banded symmetric Toeplitz matrices [1];
2. Toeplitz-like matrices, $n^{-1}K_n^{[p]}$, $nM_n^{[p]}$, $n^{-2}L_n^{[p]}$, coming from the B-spline IgA approximation of $-u'' = \lambda u$, plus its multivariate counterpart for $-\Delta u = \lambda u$ [58];
3. block and preconditioned block banded symmetric Toeplitz matrices [60].

The proposed algorithms are based on the classical concept of symbol, but with an innovative view on the errors of the approximation of eigenvalues by the uniform sampling of the symbol. This new approach was used in the independent works [16, 17, 19] and [62] where the authors conjectured the existence of an asymptotic spectral expansion for banded symmetric Toeplitz matrices. From a theoretical viewpoint in the **Chapter VI** we have proved, for all the Items, the first order asymptotic term of the expansion, using purely linear algebra tools. The theoretical proof of the asymptotic expansion for higher-order $\alpha \geq 1$ will be a future research line. Considering that the asymptotic eigenvalue expansion in IgA context is strongly connected with the eigenvalue expansion for preconditioned Toeplitz matrices of Section III.1, a proof of the former may suggest the way to prove the latter, and vice versa.

We also complement the results of [51, 71, 72, 73, 74, 76, 77], proving several important analytic properties of $e_p(\theta)$, spectral symbol of $\{n^{-2}L_n^{[p]}\}_n$.

We have extended for all contexts above the extrapolation algorithm based on the asymptotic expansion and we have demonstrated that the simple-loop requirement treated in [7, 16, 17], in standard double precision computations, is not a problem when using our proposed algorithms.

In **Chapter III** we have considered the problem of computing the spectrum of the sequence of preconditioned Toeplitz matrices $\{\mathcal{P}_n(f,g) = T_n^{-1}(g)T_n(f)\}$, for f trigonometric polynomial, g nonnegative and not identically zero trigonometric polynomial. Moreover we have shown numerical evidences showing that some of the assumptions proposed by Bogoya et al. [16, 17, 19] can be relaxed. We have extended the extrapolation algorithm for computing the eigenvalues in this setting, here the key has been the sampling of the function $r = f/g$ that plays the same role as f in the non-preconditioned case.

This generalization have potential application to the computation of the spectrum of differential operators. In fact, up to low rank corrections, matrices of the form $\mathcal{P}_n(f,g)$ appear in the context of the spectral approximation of differential operators in which a low rank correction of $T_n(g)$ is the mass matrix and a low rank correction of $T_n(f)$ is the stiffness matrix.

Therefore a plan for the future has to include: the analysis of the non-monotone case and its relations with the study in [63] for the special case where $f(\theta) = 2 - 2\cos(\omega\theta)$, $\omega \geq 2$ integer, and $g(\theta) = 1$; the extension of the results by [8] to the preconditioned Toeplitz case and the study of its connection with the treated general expansion; the extension of the numerical and theoretical study to possible other contexts.

The positive result of the preconditioned case have suggested that same kind of asymptotic expansion holds, at least in the context of the IgA approximation of second order differential operators.

In **Chapter IV** we have further explored the B-spline IgA approximation of the Laplacian eigenvalue problem $-\Delta u = \lambda u$ over the k-dimensional hypercube $(0,1)^k$. We have provided the exact eigenvalue–eigenvector structure of the resulting discretization matrices $K_n^{[p]}$, $M_n^{[p]}$, and $L_n^{[p]}$, for $p = 1, 2$. For the cases $p \geq 3$ we have proposed a parallel interpolation–extrapolation algorithm based on the asymptotic spectral expansion for computing the eigenvalues of $L_n^{[p]}$, excluding the largest $n_p^{\text{out}} = p - 2 + \text{mod}(p, 2)$ outliers. The performance of the algorithm has been illustrated through several numerical experiments. By using tensor-product arguments, it is plain to extend the whole analysis to the general k-dimensional setting.

The matrices arising from the discretization of a linear PDE by a linear Numerical Method (NM) usually have a Toeplitz or Toeplitz-like structure. For example, in the case of a constant-coefficient PDE, the matrix structure is often a small perturbation of a pure Toeplitz structure, whereas in the case of a variable-coefficient PDE, the matrix structure is often the so-called Generalized Locally Toeplitz structure [76, 77, 125, 126]; see in particular [77, Section 7.1]. Hence the natural question is:

"*Do we have an asymptotic expansions for the eigenvalues of generic PDE discretization matrices?*"

The chapter has provided a positive answer in the case where the PDE a the Laplacian eigenproblem and the discretization is the B-spline IgA. It is clear, however, that the previous question opens the doors to a series of possible future researches. Hence the purpose will be ascertain the existence of an asymptotic eigenvalue expansion for PDE discretization matrices and exploit this expansion (if any) for computing the eigenvalues themselves through fast interpolation–extrapolation procedures.

A big step forward in this direction has been the generalization of the proposed theory to the block and preconditioned block context, presented in **Chapter V**. Special attention has been dedicate to the generalization of the results of **Chapters III-IV** under the assumptions that $\mathbf{f}$ of is an $s \times s$ matrix-valued trigonometric polynomial with $s \geq 1$, and $T_n(\mathbf{f})$ is the associated block Toeplitz matrix, whose size is $N(n, s) = sn$.

First we numerically have derived the conditions which ensure the existence of an asymptotic expansion for the eigenvalues, generalizing those for the scalar-valued setting $s = 1$. Furthermore, following the proposal for $s = 1$ in the previous chapters, we have devised an interpolation–extrapolation algorithm for computing the eigenvalues of banded symmetric block Toeplitz matrices with a high level of accuracy and a low computational cost, and we have presented several examples of practical interest. Furthermore we have provided the exact formulae for the eigenvalues of the matrices coming from the $\mathbb{Q}_p$ Lagrangian Finite Element approximation of a second order elliptic differential problem and the preconditioned block matrices coming from the classical Lagrangian Finite Element approximation of the classical eigenvalue problem for the Laplacian operator in one dimension.

The natural step in order to investigate the existence of an asymptotic eigenvalue expansion for many other PDE discretization matrices will be a feasible extension of the proposed approach to the multilevel contexts, in cases where the tensor product argument of **Chapter IV** cannot

be exploited. This is a real challenge and we still do not know how to race the problem: here the principal open question concerns the formalization, in both scalar or block case, of the asymptotic spectral expansion for k-level matrices, that in turn depends on the lack of the monotonicity concept for a k variate symbol.

The matrices produced by most types of discretizations possess a structure, namely they are often banded. As seen, the latter is strongly related with the concept of trigonometric polynomial generating function, and this case has been the main object of the present thesis. However, we stress that there are many other contexts, such as discretized fractional, differential or integral operators where the involved symbols are not of polynomial type.

Hence, moved by preliminary positive results, we plan also to extend the theory and the algorithms to the eigenvalues of possibly dense matrix sequences, possibly with some regularity conditions on the symbol. The aim is that of continuing to provide and analyze methods in order to deal with the most general classes of structured matrix sequences and discretizations of partial differential equations using spectral methods or of fractional differential equations by means of standard local methods.

Bibliography

[1] F. Ahmad, E.S. Al-Aidarous, D. Abdullah Alrehaili, S.-E. Ekström, I. Furci, and S. Serra-Capizzano. Are the eigenvalues of preconditioned banded symmetric Toeplitz matrices known in almost closed form? *Numer. Algorithms*, 78(3):867–893, 2018. p. viii, xii, 25, 82, 109, 162

[2] A. Aricò and M. Donatelli. A V-cycle multigrid for multilevel matrix algebras: proof of optimality. *Numer. Math.*, 105(4):511–547, 2007. p. viii, 20, 22

[3] A. Aricò, M. Donatelli, and S. Serra-Capizzano. V-cycle optimal convergence for certain (multilevel) structured matrices. *SIAM J. Matrix Anal. Appl.*, 26(1):186–214, 2004. p. viii, 20, 22, 56

[4] B. F. Armaly, F. Durst, J. C. F. Pereira, and B. Schonung. Experimental and theoretical investigation on backward-facing step flow. *J. Fluid Mech.*, 127:473–496, 1983. p. 27

[5] O. Axelsson and G. Lindskog. On the rate of convergence of the preconditioned conjugate gradient method. *Numer. Math.*, 48(5):499–523, 1986. p. vii, 18, 32

[6] O. Axelsson and M. Neytcheva. The algebraic multilevel iteration methods–theory and applications. In *Proceedings of the 2nd International Colloquium on Numerical Analysis, Plovdiv, Bulgaria (August) by D. D. Bainov, and V. Covachev*, pages 13–23, 1993. p. 18

[7] M. Barrera, Böttcher A., S. M. Grudsky, and E. A. Maximenko. Eigenvalues of even very nice Toeplitz matrices can be unexpectedly erratic. *Operator Theory Adv. Appl.*, 268:51–77, 2018. p. viii, 26, 109, 130, 162

[8] M. Barrera and S. M. Grudsky. Asymptotics of eigenvalues for pentadiagonal symmetric Toeplitz matrices. *Oper. Theory Adv. Appl.*, 259:51–77, 2017. p. 66, 67, 162

[9] F. Bassi, A. Crivellini, D. A. Di Pietro, and S. Rebay. On a robust discontinuous Galerkin technique for the solution of compressible flow. *J. Comput. Phys.*, 218:208–221, 2006. p. 27

[10] F. Bassi, A. Crivellini, D. A. Di Pietro, and S. Rebay. An implicit high-order discontinuous Galerkin method for steady and unsteady incompressible flows. *Comput. Fluids*, 36:1529–1546, 2007. p. 27

[11] B. Beckermann and A. B. J. Kuijlaars. Superlinear convergence of conjugate gradients. *SIAM J. Numer. Anal.*, 39(1):300–329, 2001. p. vii, 32

[12] J. B. Bell, P. Coletta, and H. M. Glaz. A second-order projection method for the incompressible Navier-Stokes equations. *J. Comput. Phys.*, 85:257–283, 1989. p. 48, 50

[13] R. Bhatia. *Matrix Analysis.* Springer-Verlag, New York, 1997. p. 26, 45, 140, 153, 158

[14] D. Bini and M. Capovani. Spectral and computational properties of band symmetric Toeplitz matrices . *Linear Algebra Appl*, 52–53:99–126, 1983. p. 12, 13, 84, 139, 158

[15] D. A. Bini, G. Latouche, and B. Meini. *Numerical methods for structured Markov chains. (Numerical Mathematics and Scientific Computation).* Oxford University Press, New York, 2005. p. vi, 7

[16] J. M. Bogoya, S. M. Grudsky, and E. A. Maximenko. Eigenvalues of Hermitian Toeplitz matrices generated by simple-loop symbols with relaxed smoothness. *Oper. Theory Adv. Appl.*, 259:179–212, 2017. p. viii, xi, 23, 26, 57, 60, 109, 162

[17] J.M. Bogoya, A. Böttcher, S. M. Grudsky, and E. A. Maximenko. Eigenvalues of Hermitian Toeplitz matrices with smooth simple-loop symbols . *J. Math. Anal. Appl.*, 422:1308–1334, 2015. p. viii, xi, 23, 26, 57, 60, 109, 110, 162

[18] A. Böttcher and S. M. Grudsky. On the condition numbers of large semi-definite Toeplitz matrices. *Linear Algebra Appl.*, 279:285–301, 1998. p. 10

[19] A. Böttcher, S. M. Grudsky, and E. A. Maximenko. Inside the eigenvalues of certain Hermitian Toeplitz band matrices. *J. Comput. Appl. Math.*, 233:2245–2264, 2010. p. viii, xi, 23, 57, 60, 109, 162

[20] A. Böttcher and B. Silbermann. *Introduction to Large Truncated Toeplitz Matrices.* Springer, 1999. p. 26, 58

[21] E. Bozzo and C. Di Fiore. On the use of certain matrix algebras associated with discrete trigonometric transforms in matrix displacement decomposition. *SIAM. J. Matrix Anal. Appl.*, 16:312–326, 1995. p. 80, 84, 85, 132

[22] M. E. Brachet, D. I. Meiron, and S. A. Orszag. Small-scale structure of the Taylor-Green vortex. *J. Fluid Mech.*, 130:411–452, 1983. p. 48

[23] C. Brezinski and Zaglia M. Redivo. *Extrapolation Methods: Theory and Practice.* Elsevier Science Publishers B. V, Amsterdam, 1991. p. 60, 89, 112

[24] A. N. Brooks and T. J. R. Hughes. Stream-line upwind/Petrov Galerkin formulstion for convection dominated flows with particular emphasis on the incompressible Navier-Stokes equations. *Comput. Methods Appl. Mech. Engrg.*, 32:199–259, 1982. p. 27

[25] V. Casulli. A semi-implicit finite difference method for non-hydrostatic free-surface flows. *Internat. J. Numer. Methods Fluids*, 30:425–440, 1999. p. 28

[26] V. Casulli. A high-resolution wetting and drying algorithm for free-surface hydrodynamics. *Internat. J. Numer. Methods Fluids*, 60:391–408, 2009. p. 28

[27] V. Casulli. A semi–implicit numerical method for the free–surface Navier–Stokes equations. *Internat. J. Numer. Methods Fluids*, 74:605–622, 2014. p. 28

[28] V. Casulli and R. T. Cheng. Semi-implicit finite difference methods for three–dimensional shallow water flow. *Internat. J. Numer. Methods Fluids*, 15:629–648, 1992. p. 28

[29] V. Casulli and G. S. Stelling. Semi-implicit subgrid modelling of three-dimensional free-surface flows. *Internat. J. Numer. Methods Fluids*, 67:441–449, 2011. p. 28

[30] V. Casulli and R. A. Walters. An unstructured grid, three–dimensional model based on the shallow water equations. *Internat. J. Numer. Methods Fluids*, 32:331–348, 2000. p. 28

[31] R. H. Chan and X. Jin. An introduction to iterative Toeplitz solvers. *SIAM, Philadelphia*, 5, 2007. p. 8

[32] R. H. Chan and M. Ng. Conjugate gradient methods for Toeplitz systems. *SIAM Review*, 38(3):427–482, 1996. p. viii, 19, 59

[33] R. H. Chan and G. Strang. Toeplitz equations by Conjugate Gradients with circulant preconditioner. *SIAM J. Sci. Statist. Comput.*, 10(1):104–119, 1989. p. 19

[34] R. H. Chan and P. Tang. Fast band–Toeplitz preconditioners for Hermitian Toeplitz systems. *SIAM J. Sci. Comput*, 15:164–171, 1994. p. 59

[35] R. H. Chan and M. C. Yeung. Circulant preconditioners for Toeplitz matrices with positive continuous generating functions. *Math. Comp.*, 58:233–240, 1992. p. 20

[36] T. Chan. An optimal circulant preconditioner for Toeplitz systems. *SIAM J. Sci. Comput.*, 9(4):766–771, 1988. p. 19

[37] H. Chen, S. Jia, and H. Xie. Postprocessing and higher order convergence for the mixed finite element approximations of the eigenvalue problem. *Appl. Numer. Math.*, 61:615–629, 2011. p. 82

[38] C. K. Chui. *An Introduction to Wavelets.* Academic Press, 1992. p. 144

[39] B. Cockburn and C. W. Shu. The Runge-Kutta discontinuous Galerkin method for conservation laws V: multidimensional systems. *J. Comput. Phys.*, 141:199–224, 1998. p. 28

[40] B. Cockburn and C. W. Shu. Runge-Kutta discontinuous Galerkin methods for convection-dominated problems. *J. Sci. Comput.*, 16:173–261, 2001. p. 28

[41] J. A. Cottrell, T. J. R. Hughes, and Y. Bazilevs. *Isogeometric Analysis: Toward Integration of CAD and FEA.* John Wiley & Sons, 2009. p. 77, 79

[42] J. A. Cottrell, A. Reali, Y. Bazilevs, and T. J. R. Hughes. Isogeometric analysis of structural vibrations. *Comput. Methods Appl. Mech. Eng.*, 195(41):5257–5296, 2006. p. vi, 77, 161

[43] A. Crivellini, V. D'Alessandro, and F. Bassi. High-order discontinuous Galerkin solutions of three-dimensional incompressible RANS equations. *Comput. Fluids*, 81:122–133, 2013. p. 27

[44] P. Davis. *Circulant Matrices*. J. Wiley and Sons, New York, 1979. p. 15

[45] C. De Boor. *A Practical Guide to Splines*. Revised Edition, Springer, 2001. p. 79

[46] V. Del Prete, F. Di Benedetto, M. Donatelli, and S. Serra-Capizzano. Symbol approach in a signal-restoration problem involving block Toeplitz matrices. *J. Comput. Appl. Math.*, 272:399–416, 2014. p. vi, 7

[47] F. Di Benedetto, G. Fiorentino, and S. Serra-Capizzano. C.G. Preconditioning for Toeplitz Matrices. *Comput. Math. Appl.*, 25(6):33–45, 1993. p. 59, 142, 159

[48] M. Donatelli, A. Dorostkar, M. Mazza, M. Neytcheva, and S. Serra-Capizzano. Function-based block multigrid strategy for a two-dimensional linear elasticity-type problem. *Comput. Math. Appl.*, 74(5):1015–1028, 2017. p. viii

[49] M. Donatelli, C. Garoni, C. Manni, S. Serra-Capizzano, and H. Speleers. Robust and optimal multi-iterative techniques for IgA collocation linear systems. *Comput. Methods Appl. Mech. Engrg.*, 284:1120–1146, 2015. p. 77

[50] M. Donatelli, C. Garoni, C. Manni, S. Serra-Capizzano, and H. Speleers. Robust and optimal multi-iterative techniques for IgA Galerkin linear systems. *Comput. Methods Appl. Mech. Engrg.*, 284:230–264, 2015. p. 77, 80, 84, 88

[51] M. Donatelli, C. Garoni, C. Manni, S. Serra-Capizzano, and H. Speleers. Spectral analysis and spectral symbol of matrices in Isogeometric collocation methods. *Math. Comp.*, 85:1639–1680, 2016. p. viii, 77, 162

[52] M. Donatelli, C. Garoni, C. Manni, S. Serra-Capizzano, and H. Speleers. Symbol-based multigrid methods for Galerkin B-spline Isogeometric analysis. *SIAM J. Numer. Anal.*, 55-1:31–62, 2017. p. 77, 80, 84, 88, 151

[53] M. Donatelli, M. Neytcheva, and S Serra-Capizzano. Canonical eigenvalue distribution of multilevel block Toeplitz sequences with non-Hermitian symbols. *Oper. Theory Adv. Appl.*, 221:269–291, 2012. p. 106

[54] M. Dumbser and V. Casulli. A staggered semi-implicit spectral discontinuous Galerkin scheme for the shallow water equations. *Appl. Math. Comput.*, 219(15):8057–8077, 2013. p. 30

[55] M. Dumbser, F. Fambri, I. Furci, M. Mazza, S. Serra-Capizzano, and M. Tavelli. Staggered discontinuous Galerkin methods for the incompressible Navier–Stokes equations: spectral analysis and computational results. *Numer. Linear Algebra Appl.*, 25(5), 2018. p. xii

[56] M. Dumbser, O. Zanotti, R. Loubère, and S. Diot. A posteriori subcell limiting of the discontinuous Galerkin finite element method for hyperbolic conservation laws. *J. Comput. Phys.*, 278:47–75, 2014. p. 28

[57] S.-E. Ekström. *Matrix-less Methods for Computing Eigenvalues of Large Structured Matrices. Ph.D. Thesis*. PhD thesis, Uppsala University, 2018. p. x, 109, 130

[58] S.-E. Ekström, I. Furci, C. Garoni, C. Manni, S. Serra-Capizzano, and H. Speleers. Are the eigenvalues of the B-spline IgA approximation of $-\Delta u = \lambda u$ known in almost closed form? (Early version by S.-E. Ekström, I. Furci, S. Serra–Capizzano with the same title in Technical report, 2017-016, Department of Information Technology, Uppsala University). *Numer. Linear Algebra Appl.*, 25(5), 2018. p. viii, xii, 26, 83, 88, 90, 91, 109, 110, 112, 113, 162

[59] S.-E. Ekström, I. Furci, and S. Serra-Capizzano. Exact formulae and matrix-less eigensolvers for preconditioned block banded symmetric Toeplitz-like matrices. In preparation. p. xii

[60] S.-E. Ekström, I. Furci, and S. Serra-Capizzano. Exact formulae and matrix-less eigensolvers for block banded symmetric Toeplitz matrices. *BIT*, 2018. (In press). p. viii, xii, 26, 162

[61] S.-E. Ekström and C. Garoni. A matrix-less and parallel interpolation–extrapolation algorithm for computing the eigenvalues of preconditioned banded symmetric Toeplitz matrices. *Numer. Algorithms*, 2018. (In press). p. 58, 90, 117

[62] S-E. Ekström, C. Garoni, and S. Serra-Capizzano. Are the eigenvalues of banded symmetric Toeplitz matrices known in almost closed form? *Exp. Math.*, pages 1–10, 2017. p. viii, 23, 24, 58, 60, 82, 109, 162

[63] S.-E. Ekström and S. Serra-Capizzano. Eigenvalues and eigenvectors of banded Toeplitz matrices and the related symbols. *Numer. Linear Algebra Appl.*, 2018. (In press). p. 69, 109, 129, 130, 162

[64] C. Estatico and S. Serra-Capizzano. Superoptimal approximation for unbounded symbols. *Linear Algebra Appl.*, 428(2-3):564–585, 2008. p. 17

[65] F. Fambri and M. Dumbser. Spectral semi-implicit and space-time discontinuous Galerkin methods for the incompressible Navier-Stokes equations on staggered Cartesian grids. *Appl. Numer. Math.*, 110:41–74, 2016. p. vii, 27, 28, 30, 31, 32, 46, 47, 48, 137

[66] F. Fambri and M. Dumbser. Semi-implicit discontinuous Galerkin methods for the incompressible Navier-Stokes equations on adaptive staggered Cartesian grids. *Comput. Methods Appl. Mech. Engrg.*, 324:170–203, 2017. p. vii, 27, 28

[67] P. Ferrari, I. Furci, S. Hon, M. A. Mursaleen, and S. Serra-Capizzano. The eigenvalue distribution of special 2-by-2 block matrix sequences, with applications to the case of symmetrized Toeplitz structures. Submitted. p. xii

[68] E. Ferrer and R. H. J. Willden. A high order discontinuous Galerkin finite element solver for the incompressible Navier-Stokes equations. *Comput. & Fluids*, 46:224–230, 2011. p. 27

[69] G. Fiorentino and S. Serra-Capizzano. Multigrid methods for symmetric positive definite block Toeplitz matrices with nonnegative generating functions. *SIAM J. Sci. Comput.*, 17(5):1068–1081, 1996. p. viii, 22

[70] M. Fortin. Old and new finite elements for incompressible flows. *Internat. J. Numer. Methods Fluids*, 1:347–364, 1981. p. 27

[71] C. Garoni. Spectral distribution of PDE discretization matrices from isogeometric analysis: the case of L^1 coefficients and non-regular geometry. *J. Spectr. Theory*, 8:297–313, 2018. p. viii, 77, 162

[72] C. Garoni, C. Manni, F. Pelosi, S. Serra-Capizzano, and H. Speleers. On the spectrum of stiffness matrices arising from Isogeometric analysis. *Numer. Math.*, 127:751–799, 2014. p. viii, 77, 80, 82, 143, 162

[73] C. Garoni, C. Manni, S. Serra-Capizzano, D. Sesana, and H. Speleers. Lusin theorem, glt sequences and matrix computations: an application to the spectral analysis of PDE discretization matrices. *J. Math. Anal. Appl.*, 446:365–382, 2017. p. viii, 77, 162

[74] C. Garoni, C. Manni, S. Serra-Capizzano, D. Sesana, and H. Speleers. Spectral analysis and spectral symbol of matrices in Isogeometric Galerkin methods. *Math. Comp.*, 86:1343–1373, 2017. p. viii, 77, 162

[75] C. Garoni and S. Serra Capizzano. Generalized Locally Toeplitz sequences: a spectral analysis tool for discretized differential equations. To appear in a volume of the Springer book series "Lecture Notes in Mathematics, C.I.M.E. Foundation Subseries". p. 83

[76] C. Garoni and S. Serra-Capizzano. Generalized Locally Toeplitz sequences: Theory and applications. Technical Report 2017-002, Department of Information Technology, Uppsala University, 2017. p. viii, 3, 77, 99, 162, 163

[77] C. Garoni and S. Serra-Capizzano. *The Theory of Generalized Locally Toeplitz Sequences: Theory and Applications, Vol. I.* Springer Monographs in Mathematics, Berlin, 2017. p. viii, ix, 9, 26, 58, 77, 80, 106, 162, 163

[78] C. Garoni, S. Serra-Capizzano, and D. Sesana. Spectral analysis and spectral symbol of d-variate $\mathbb{Q}_{\mathbf{p}}$ lagrangian FEM stiffness matrices. *SIAM J. Matrix Anal. Appl.*, 36(3):1100–1128, 2015. p. 16, 115, 131, 132, 135

[79] C. Garoni, S. Serra Capizzano, and P. Vassalos. A general tool for determining the asymptotic spectral distribution of Hermitian matrix-sequences. *Oper. Matrices*, 9(3):549–561, 2015. p. 9

[80] C. Garoni, H. Speleers, S.-E. Ekström, A. Reali, S. Serra-Capizzano, and T. J. R. Hughes. Symbol-based analysis of finite element and Isogeometric B-spline discretizations of eigenvalue problems: Exposition and review. *Arch. Comput. Methods Eng.*, 2018. (In press). p. 77, 83

[81] U. Grenander and G. Szegő. *Toeplitz forms and their applications.* Chelsea Publishing Co., New York, second edition, 1984. p. 9

[82] P. C. Hansen, J. G. Nagy, and D. P. O'Leary. *Deblurring Images: Matrices, Spectra, and Filtering (Fundamentals of Algorithms 3).* SIAM, Philadelphia, 2006. p. vi, 7

[83] F. H. Harlow and J. E. Welch. Numerical calculation of time-dependent viscous incompressible flow of fluid with a free surface. *Phys. Fluids*, 8:2182–2189, 1965. p. 27, 28

[84] M. R. Hestenes and E. Stiefel. Methods of conjugate gradients for solving linear systems. *J. Res. Natl. Bur. Stand.*, 49:409–436, 1952. p. 28

[85] J. G. Heywood and R. Rannacher. Finite element approximation of the nonstationary Navier-Stokes Problem. I. Regularity of solutions and second order error estimates for spatial discretization. *SIAM J. Numer. Anal.*, 19:275–311, 1982. p. 27

[86] J. G. Heywood and R. Rannacher. Finite element approximation of the nonstationary Navier-Stokes problem. iii. smoothing property and higher order error estimates for spatial discretization. *SIAM J. Numer. Anal.*, 25:489–512, 1988. p. 27

[87] C. W. Hirt and B. D. Nichols. Volume of fluid (VOF) method for dynamics of free boundaries. *J. Comput. Phys.*, 39:201–225, 1981. p. 28

[88] T. Huckle, S. Serra-Capizzano, and C. Tablino-Possio. Preconditioning strategies for Hermitian indefinite Toeplitz linear systems. *SIAM J. Sci. Comput.*, 25(5):1633–1654, 2004. p. 59

[89] T. Huckle, S. Serra-Capizzano, and C. Tablino-Possio. Preconditioning strategies for non-Hermitian Toeplitz linear systems. *Numer. Linear Algebra Appl.*, 12(2–3):211–220, 2005. p. 59

[90] T. J. R. Hughes, J. A. Evans, and A. Reali. Finite element and nurbs approximations of eigenvalue, boundary-value, and initial-value problems. *Comput. Methods Appl. Mech. Engrg.*, 272:290–320, 2014. p. 77

[91] T. J. R. Hughes, M. Mallet, and M. Mizukami. A new finite element formulation for computational fluid dynamics: II. Beyond SUPG. *Comput. Methods Appl. Mech. Engrg.*, 54:341–355, 1986. p. 27

[92] T. J. R. Hughes, A. Reali, and G. Sangalli. Duality and unified analysis of discrete approximations in structural dynamics and wave propagation: Comparison of p-method finite elements with k-method NURBS. *Comput. Methods Appl. Mech. Eng.*, 197(49):4104–4124, 2008. p. vi, 77, 161

[93] T. Kailath and A. H. Sayed. Displacement structures: theory and applications. *SIAM Review*, 37:297–386, 1995. p. viii

[94] T. Kailath, A. Vieira, and M. Morf. Inverses of Topelitz operators, innovations and orthogonal polynomials. *SIAM Review*, 20:106–119, 1978. p. viii

[95] B. Klein, F. Kummer, and M. Oberlack. A SIMPLE based discontinuous Galerkin solver for steady incompressible flows. *J. Comput. Phys.*, 237:235–250, 2013. p. 27

[96] M. Ng. *Iterative methods for Toeplitz systems (Numerical Mathematics and Scientific Computation)*. Oxford University Press, New York., 2004. p. 8, 19

[97] N. C. Nguyen, J. Peraire, and B. Cockburn. An implicit high-order hybridizable discontinuous Galerkin method for the incompressible Navier-Stokes equations. *J. Comput. Phys.*, 230:1147–1170, 2011. p. 27

[98] D. Noutsos, S. Serra-Capizzano, and P. Vassalos. Matrix algebra preconditioners for multilevel Toeplitz systems do not insure optimal convergence rate. *Theoret. Comput. Sci.*, 315(2):557–579, 2004. p. vii, xi, 20

[99] V. S. Patankar. *Numerical Heat Transfer and Fluid Flow.* Hemisphere Publishing Corporation, New York, 1980. p. 27, 28

[100] V. S. Patankar and B. Spalding. A calculation procedure for heat, mass and momentum transfer in three-dimensional parabolic flows. *Int. J. Heat Mass Transfer.*, 15:1787–1806, 1972. p. 27, 28

[101] G. Plonka and M. Tasche. Fast and numerically stable algorithms for discrete cosine transforms. *Linear Algebra Appl.*, 394:309–345, 2005. p. 80

[102] D. Potts, G. Steidl, and M. Tasche. Numerical stability of fast trigonometric transforms–A worst case study. *J. Concr. Appl. Math.*, 1(1):1–35, 2003. p. 80

[103] A. Reali. An Isogeometric analysis approach for the study of structural vibrations. *J. Earthquake Engrg.*, 10:1–30, 2006. p. 77

[104] S. Rhebergen and B. Cockburn. A space-time hybridizable discontinuous Galerkin method for incompressible flows on deforming domains. *J. Comput. Phys.*, 231:4185–4204, 2012. p. 27

[105] S. Rhebergen, B. Cockburn, and Jaap J. W. Van der Vegt. A space-time discontinuous Galerkin method for the incompressible Navier-Stokes equations. *J. Comput. Phys.*, 233:339–358, 2013. p. 27

[106] J. W. Ruge and K. Stüben. *Algebraic multigrid. In Multigrid Methods.* S. McCormick, ed., Frontiers Appl. Math. 3, SIAM, Philadelphia, 1987. p. 21

[107] V. V. Rusanov. Calculation of Interaction of Non–Steady Shock Waves with Obstacles. *Comput. Math. Math. Phys. USSR*, 1:267–279, 1961. p. 28

[108] Y. Saad and M. H. Schultz. GMRES: a generalized minimal residual algorithm for solving nonsymmetric linear systems. *SIAM J. Sci. Statist. Comput.*, 7:856–869, 1986. p. 28

[109] H. Sakamoto and H. Haniu. A study on vortex shedding from spheres in a uniform flow. *J. Fluids Eng.*, 112:386–392, 1990. p. 27

[110] E. Salinelli, S. Serra-Capizzano, and D. Sesana. Eigenvalue-eigenvector structure of Schoenmakers-Coffey matrices via Toeplitz technology and applications. *Linear Algebra Appl.*, 491:138–160, 2016. p. vi, 7

[111] G. Sangalli and M. Tani. Isogeometric preconditioners based on fast solvers for the Sylvester equation. *SIAM J. Sci. Comput.*, 38:A3644–A3671, 2016. p. 88

[112] L. L. Schumaker. *Spline Functions: Basic Theory.* Third Edition, Cambridge University Press, 2007. p. 79

[113] S. Serra-Capizzano. Multi-iterative Methods. *Comput. Math. Appl.*, 26(4):65–87, 1993. p. 161

[114] S. Serra-Capizzano. New PCG based algorithms for the solution of Hermitian Toeplitz systems. *Calcolo*, 32:53–176, 1995. p. 66

[115] S. Serra-Capizzano. On the extreme spectral properties of Toeplitz matrices generated by L^1 functions with several minima/maxima. *Linear Algebra Appl.*, 36:135–142, 1996. p. 10, 152

[116] S. Serra-Capizzano. Optimal, quasi-optimal and superlinear band-Toeplitz preconditioners for asymptotically ill-conditioned positive definite Toeplitz systems. *Math. Comp.*, 66(218):651–665, 1997. p. 59

[117] S. Serra-Capizzano. Asymptotic results on the spectra of block Toeplitz preconditioned matrices. *SIAM J. Matrix Anal. Appl.*, 20–1:31–44, 1998. p. 6, 10, 107, 115

[118] S. Serra-Capizzano. An ergodic theorem for classes of preconditioned matrices. *Linear Algebra Appl.*, 282(1–3):161–183, 1998. p. 59

[119] S. Serra-Capizzano. On the extreme eigenvalues of Hermitian (block) Toeplitz matrices. *Linear Algebra Appl.*, 270:109–128, 1998. p. 10

[120] S. Serra-Capizzano. A Korovkin-type theory for finite Toeplitz operators via matrix algebras. *Numer. Math.*, 82:117–142, 1999. p. 19, 20

[121] S. Serra-Capizzano. Spectral and computational analysis of block Toeplitz matrices having nonnegative definite matrix-valued generating functions. *BIT*, 39–1:152–175, 1999. p. 10, 107, 115

[122] S. Serra-Capizzano. Superlinear PCG methods for symmetric Toeplitz systems. *Math. Comp.*, 88:793–803, 1999. p. 19, 20

[123] S. Serra-Capizzano. Spectral behavior of matrix sequences and discretized boundary value problems. *Linear Algebra Appl.*, 337(1):37 – 78, 2001. p. 18

[124] S. Serra-Capizzano. Matrix algebra preconditioners for multilevel Toeplitz matrices are not superlinear. *Linear Algebra Appl.*, 343:303–319, 2002. p. vii, xi, 20

[125] S. Serra-Capizzano. Generalized Locally Toeplitz sequences: spectral analysis and applications to discretized partial differential equations. *Linear Algebra Appl.*, 366:371–402, 2003. p. 17, 24, 163

[126] S. Serra-Capizzano. The GLT class as a Generalized Fourier Analysis and applications. *Linear Algebra Appl.*, 419(1):180–233, 2006. p. 17, 163

[127] S. Serra-Capizzano and C. Tablino-Possio. Multigrid methods for multilevel circulant matrices. *SIAM J. Sci. Comput.*, 26(1):55–85, 2004. p. 20, 55, 56

[128] S. Serra-Capizzano and C. Tablino-Possio. Two-grid methods for Hermitian positive definite linear systems connected with an order relation. *Calcolo*, 51(2):261–285, 2014. p. 22, 23, 54, 56

[129] S. Serra-Capizzano and E. Tyrtyshnikov. Any circulant-like preconditioner for multilevel matrices is not superlinear. *SIAM J. Matrix Anal. Appl.*, 21(2):431–439, 2000. p. vii, xi, 20

[130] D. Sesana. Multigrid methods for block-circulant/tau linear systems. In preparation. p. viii, 22

[131] K. Shahbazi, P. F. Fischer, and C. R. Ethier. A high-order discontinuous Galerkin method for the unsteady incompressible Navier-Stokes equations. *J. Comput. Phys.*, 222:391–407, 2007. p. 27

[132] J. Stoer and R. Bulirsch. *Introduction to Numerical Analysis, 3rd edn.* Springer, New York, 2002. p. 60, 89, 110, 112

[133] G. Strang. A proposal for toeplitz matrix calculations. *Stud. Appl. Math.*, 74(2):171–176, 1986. p. 19

[134] M. Tani. *FFT-based fast diagonalization methods for Galerkin IgA.* Private Communication, 2017. p. 88

[135] M. Tavelli and M. Dumbser. A staggered semi-implicit discontinuous Galerkin method for the two dimensional incompressible Navier-Stokes equations. *Appl. Math. Comput.*, 248:70–92, 2014. p. vii, 27, 28

[136] M. Tavelli and M. Dumbser. A staggered arbitrary high order semi-implicit discontinuous Galerkin method for the two dimensional incompressible Navier-Stokes equations. *Comput. & Fluids*, 119:235–249, 2015. p. 28, 48

[137] M. Tavelli and M. Dumbser. A staggered space-time discontinuous Galerkin method for the three-dimensional incompressible Navier-Stokes equations on unstructured tetrahedral meshes. *J. Comput. Phys.*, 319:294–323, 2016. p. vii, 27, 28, 48

[138] M. Tavelli and M. Dumbser. A pressure-based semi-implicit space-time discontinuous Galerkin method on staggered unstructured meshes for the solution of the compressible Navier-Stokes equations at all Mach numbers. *J. Comput. Phys.*, 341:341–376, 2017. p. 28

[139] C. Taylor and P. Hood. A numerical solution of the Navier-Stokes equations using the finite element technique. *Comput. & Fluids*, 1:73–100, 1973. p. 27

[140] P. Tilli. A note on the spectral distribution of Toeplitz matrices. *Linear Multilinear Algebra*, 45(2–3):147–159, 1998. p. 9, 106

[141] P. Tilli. Locally Toeplitz sequences: spectral properties and applications. *Linear Algebra Appl.*, 278(1):91–120, 1998. p. 9, 17

[142] P. Tilli. Some results on complex Toeplitz eigenvalues. *J. Math. Anal. Appl.*, 239:390–401, 1999. p. 106

[143] O. Toeplitz. Zur transformation der scharen bilinear formen von unendlichvielen veränderlichen. *Nachrichten von der Gesellschaft der Wissenschaften zu Göttingen, Mathematisch-Physikalische Klasse*, pages 110–115, 1910. p. 7

[144] O. Toeplitz. Theorie der L-formen. *Mathematische Annalen*, 70:351–376, 1911. p. 7

[145] U. Trottenberg, C. W. Oosterlee, and A. Schüller. *Multigrid*. Academic Press, Inc., San Diego, CA, 2001. With contributions by A. Brandt, P. Oswald and K. Stüben. p. 21, 54

[146] E. Tyrtyshnikov and N. Zamarashkin. Spectra of multilevel Toeplitz matrices: advanced theory via simple matrix relationships. *Linear Algebra Appl.*, 270:15–27, 1998. p. 9

[147] J. Van Kan. A second-order accurate pressure correction method for viscous incompressible flow. *SIAM J. Sci. Statist. Comput.*, 7:870–891, 1986. p. 27, 28

[148] C. Van Loan. *Computational Frameworks for the Fast Fourier Transform*. SIAM, Philadelphia, 1992. p. 17, 19

[149] R. Verfürth. Finite element approximation of incompressible Navier-Stokes equations with slip boundary condition II. *Numer. Math.*, 59:615–636, 1991. p. 27

[150] C. H. K. Williamson. The existence of two stages in the transition to three-dimensionality of a cylinder wake. *Phys. Fluids*, 24:855–882, 1988. p. 27

[151] X. Yin, H. Xie, S. Jia, and S. Gao. Asymptotic expansions and extrapolations of eigenvalues for the stokes problem by mixed finite element methods. *J. Comput. Appl. Math.*, 215:127–141, 2008. p. 82

[152] N. Zamarashkin and E. Tyrtyshnikov. Distribution of the eigenvalues and singular numbers of Toeplitz matrices under weakened requirements on the generating function. *Mat. Sb.*, 188(3):83–92, 1997. p. 9

[153] O. Zanotti, F. Fambri, M. Dumbser, and A. Hidalgo. Space-time adaptive ADER discontinuous Galerkin finite element schemes with a posteriori subcell finite volume limiting. *Comput. Fluids*, 118:204–224, 2015. p. 28

[154] A. Zygmund. *Trigonometric Series*. Cambridge University Press, Cambridge, 1959. p. 17

Acknowledgments

If you only do what you can do,
you'll never be more than you are now.

Master Shifu, Kung Fu Panda III

I would like to thank all the people that allowed me to be more (and more) now than I could ever have expected.

I thank my advisor Stefano Serra-Capizzano and my co-advisor Sven-Erik Ekström for the privilege of being accompanied by their scientific and human knowledge through the challenges of research. I am grateful to our PhD coordinator Marco Donatelli for his advice and suggestions, and for listening to me whenever I needed it. I would like to thank the referees for their helpful and constructive comments.

I express my sincere gratitude to Maria Italia Gualtieri, the first to ever believe in me and to "plant" in my mind the seed of mathematical curiosity, making the beginning of this experience possible.

Although working together with many different minds has sometimes been a delicate and tricky task, the exceptional collaborators and friends I've met have enriched my way of doing research and have allowed me to grow as a person. Therefore I thank Carlo Garoni, Mariarosa Mazza, and Debora Sesana, for being such exemplary older academic siblings and for welcoming me into the "Insubria family".

I acknowledge the collaborators from Trento University: Michael Dumbser, Francesco Fambri, and Maurizio Tavelli. Thanks for your kind hospitality, friendship, and for the interesting discussions at some random conferences.

I extend a big thank to my other co-authors: Fayyaz Ahmad, Eman Salem Al-Aidarous, Dina Alrehaili, Carla Manni, and Hendrik Speleers, for their precious collaboration and friendship.

I acknowledge all the PhD colleagues and friends (from north to the furthest south), with whom I've shared moments of deep frustration but mostly hope, the joy and laughter.

I thank my family, Mom, Dad, and Serafino, whom, without understanding exactly what I'm doing, give me their unconditional support and love in the inevitable hard days.

Finally I dedicate every single page of this Thesis to my sister Rosamaria, who constantly helps me find the heart and the grit I need to persevere in this wonderful and pure folly.