University of Insubria

Department of Science and High Technology

Ph.D. in Mathematics of Computation: Models, Structures, Algorithms and Applications

Structured Matrices coming from PDE Approximation Theory: Spectral Analysis, Spectral Symbol and Design of Fast Iterative Solvers

Supervisor: Prof. Stefano Serra-Capizzano

Ph.D. Thesis of Carlo Garoni

Academic Year 2013/2014

I dedicate my Ph.D. thesis to my wonderful supervisor Stefano

Contents

Introduction

Partial Differential Equations (PDE) are extensively used in Physics, Engineering and Applied Sciences in order to model real-world problems. A closed form for the analytical solution of such PDE is normally not available and, even in the few cases in which it is available, it often reduces to a non-informative representation formula, completely useless from a practical viewpoint (think for example to the solution of the heat equation...). It is therefore of fundamental importance to approximate the solution u of a PDE by means of some numerical method.

Despite the differences that allow one to distinguish among the various numerical methods, the principle on which all of them are based is essentially the same: they first discretize the continuous PDE by introducing a mesh, related to some discretization parameter n, and then they compute the corresponding numerical solution u_n, which will converge in some topology to the solution u of the PDE when $n \to \infty$, i.e., when the mesh is progressively refined.

Now, if the considered PDE and the chosen numerical method are both linear, the actual computation of the numerical solution u_n reduces to solving a certain linear system $A_n \mathbf{u}_n = \mathbf{f}_n$ whose size d_n increases with n and tends to infinity when $n \to \infty$. Hence, what we actually have is not just a single linear system, but a whole sequence of linear systems with increasing dimensions. Furthermore, what is often verified in practice is that, when $n \to \infty$, the sequence of discretization matrices A_n enjoys an asymptotic spectral distribution, which is somehow related to the spectrum of the differential operator $\mathscr{L}$ associated with the PDE. More in detail, it often happens that, for a large set of test functions F (usually, for all continuous functions F with bounded support), the following limit relation holds:

$$\lim_{n\to\infty} \frac{1}{d_n} \sum_{j=1}^{d_n} F(\lambda_j(A_n)) = \frac{1}{m_d(D)} \int_D \frac{\sum_{i=1}^{s} F(\lambda_i(\mathbf{f}(\mathbf{x})))}{s} d\mathbf{x}, \tag{1}$$

where $\lambda_j(A_n)$, $j = 1, \ldots, d_n$, are the eigenvalues of A_n, m_d is the Lebesgue measure in $\mathbb{R}^d$, and $\lambda_i(\mathbf{f}(\mathbf{x}))$, $i = 1, \ldots, s$, are the eigenvalues of a certain matrix-valued function

$$\mathbf{f} : D \subseteq \mathbb{R}^d \to \mathbb{C}^{s\times s}. \tag{2}$$

In this situation, $\mathbf{f}$ is called the spectral symbol (or simply the symbol) of the sequence of matrices A_n, and it provides a 'compact' description of the asymptotic spectral distribution of A_n; see Remark 1.2 below.

The identification and the study of the symbol $\mathbf{f}$ are, of course, two interesting issues in themselves, because they provide a quite accurate information about the asymptotic global behavior of the eigenvalues of A_n. In particular, an often successful guesstimate for the spectral condition number $\kappa(A_n)$, at least in the case where A_n is Hermitian positive definite, can be obtained by analyzing the eigenvalue functions $\lambda_{\min}(\mathbf{f}(\mathbf{x}))$ and $\lambda_{\max}(\mathbf{f}(\mathbf{x}))$ (especially, the number and the orders of the zeros of $\lambda_{\min}(\mathbf{f}(\mathbf{x}))$, if any). Moreover, the number s, indentifying the space $\mathbb{C}^{s\times s}$ in which the symbol $\mathbf{f}$ takes values, coincides with the number of 'branches' that compose the asymptotic spectrum of A_n; refer again to Remark 1.2 for details and see [32] for recent findings, concerning the number of spectral branches that characterize the large discretization matrices associated with Galerkin-type approximations of the Laplacian eigenvalue problem $-\Delta u = \lambda u$.

At this point, we should say that the knowledge of the symbol $\mathbf{f}$ and of its properties is not only interesting in itself, but can also be used for practical purposes. In particular, the symbol can be used either to perform a convergence analysis and predict the behavior of preconditioned Krylov and multigrid methods applied to A_n, or to design effective preconditioners and multigrid solvers for the associated linear systems. The reason is clear: the convergence properties of preconditioned Krylov and multigrid methods strongly depend on the spectral features of the matrix to which they are applied. Hence, the spectral information provided by the symbol $\mathbf{f}$ can be conveniently used for designing fast solvers of this kind and/or analyzing their convergence properties. In this respect, we recall that recent estimates of the superlinear convergence of the Conjugate Gradient (CG) method are strictly related to the asymptotic spectral distribution of the matrices to which the CG method is applied; see [5].

The purpose of this thesis is to present some specific examples in which the above philosophical discussion comes to life. As our model PDE, we consider classical second-order elliptic differential equations; see (3.1), (4.1) and (5.1) below. Concerning the numerical methods that we employ for their solution, we make three choices: the classical $\mathbb{Q}_{\boldsymbol{p}}$ Lagrangian Finite Element Method (FEM), the Galerkin Isogeometric Analysis (IgA) based on B-splines, and the IgA Collocation Method based on B-splines. The first method is a classical approximation technique and, consequently, there is not much to say about it: we just refer the reader to the wide literature on the subject (see, e.g., [47, 48, 49, 18, 55]). As for the second two methods, they will be described in Chapters 4 and 5, respectively. However, we anticipate here that both of them are based on the IgA paradigm, whose goal is to improve the connection between numerical simulation of PDE and Computer Aided Design (CAD) systems, the latter being widely employed in Engineering. In its original formulation, the main idea in IgA is to use directly the geometry provided by CAD systems and to approximate the unknown solutions of differential equations by the same type of functions. Tensor-product B-splines and their rational extension, the so-called NURBS, are the dominant technology in CAD systems used in Engineering, and thus also in IgA. The reader is referred to [33, Section 1.2] for a quick overview of the IgA paradigm and to [41, 19] for a detailed introduction to this fascinating subject, which has been developed by T. J. R. Hughes and his research team since 2005 and is now emerging on the international scene.

Despite the specific features of the three mentioned numerical methods, all of them, as well as our elliptic PDE, are linear. As a consequence, the actual computation of the numerical solution u_n reduces to solving a linear system $A_n\mathbf{u}_n = \mathbf{f}_n$ whose size d_n tends to infinity when the discretization parameter $n \to \infty$. Therefore, we are precisely in the framework described at the beginning, and we may be interested in computing the symbol $\mathbf{f}$ characterizing the asymptotic spectrum of the matrices A_n in the sense (1). This will be done, for the three numerical methods under investigation, in Chapters 3–5, where we will also study the properties of the symbol. After this, in Chapters 6–7, the properties of the symbol will be used in order to design fast iterative solvers for the discretization matrices A_n associated with the two numerical methods based on IgA (the Galerkin IgA and the IgA Collocation Method). The design of fast iterative solvers for the discretization matrices A_n associated with the Lagrangian FEM approximation is an harder task, due to the 'bad' features of the related symbol, and so it will be the subject of future research.

We now describe in more details the content of Chapters 3–7, which form the core of this thesis. Nonetheless, Chapters 1–2 are also important. Indeed, Chapter 1 provides the fundamental background that is necessary for understanding the subsequent chapters, while Chapter 2 presents new tools for computing spectral distributions, some of which are used in Chapter 3.

- In Chapter 3, we shall see that the symbol $\mathbf{f}$ of the $\mathbb{Q}_{\boldsymbol{p}}$ Lagrangian FEM stiffness matrices approximating the elliptic PDE (3.1) is a (Hermitian) matrix-valued function of the form (2) with $s = N(\boldsymbol{p}) := \prod_{i=1}^{d} p_i$, where p_i is the polynomial approximation degree in the direction x_i; see Section 3.1 for more details. In particular, this means that the (asymptotic) spectrum of the Lagrangian FEM stiffness matrices is split into $N(\boldsymbol{p})$ spectral branches (cf. Remark 1.2). We will also study the properties of the symbol $\mathbf{f}$,

and we shall see that its eigenvalues $\lambda_i(\mathbf{f}(\mathbf{x}))$, $i = 1, \ldots, N(\boldsymbol{p})$, presents an 'exponential scattering' with $\boldsymbol{p}$. This makes it difficult to design effective iterative solvers for the Lagrangian FEM stiffness matrices when the approximation parameters $\boldsymbol{p}$ are large, and, indeed, such solvers are not yet available: finding them will be the subject of future research.

- In Chapter 4, we will compute the symbol of the stiffness matrices arising from the Galerkin IgA approximation based on (tensor-product) B-splines of degree $\boldsymbol{p} = (p_1, \ldots, p_d)$ of the elliptic PDE (4.1), where again p_i is the spline approximation degree in the i-th direction (see Section 4.1). This time, the symbol is a real-valued function f, i.e., it is of the form (2) with $s = 1$ and $\mathbb{C}$ replaced by $\mathbb{R}$. Therefore, unlike the FEM matrices, the Galerkin IgA matrices have a unique spectral branch. The properties of the symbol f are deeply studied in Chapter 4 and will be used in Chapter 6 to design a fast iterative solver for the Galerkin IgA matrices.

- In Chapter 5, we will compute the symbol of the collocation matrices associated with the isogeometric collocation approximation based on (tensor-product) B-splines of degree $\boldsymbol{p} = (p_1, \ldots, p_d)$ of the full elliptic PDE (5.1). Note that such PDE is more complicated than the one considered in Chapters 3–4, and, in fact, the symbol f has a more complex structure. However, f is still a real-valued function, as in the case of the Galerkin IgA approximation considered in Chapter 4, meaning that the IgA collocation matrices have a unique spectral branch like the Galerkin IgA matrices. The properties of f will be carefully studied in Chapter 5 and will be exploited in Chapter 7 in order to design a fast iterative solver for the IgA collocation matrices.

- Chapter 6 is devoted to the design of a fast iterative solver of multigrid type for the Galerkin IgA matrices, whose symbol has been indentified and studied in Chapter 4. We point out that here the word 'fast' has a twofold meaning: first, the convergence rate of the solver must be optimal, i.e., independent of the matrix size and of the discretization parameter n; second, the convergence rate must be robust, i.e., independent of the spline approximation parameter $\boldsymbol{p} = (p_1, \ldots, p_d)$. Using the properties of the symbol provided in Chapter 4, we will succeed in designing a fast solver with these characteristics for the Galerkin IgA matrices.

- Chapter 7 is completely analogous to Chapter 6: using the properties of the symbol studied in Chapter 5, we design a fast iterative solver for the IgA collocation matrices, where the word 'fast' has again the same meaning as in Chapter 6 (see previous item).

The papers that supplied material for this thesis are [24, 25, 26, 27, 33, 34, 35, 36]. It should be emphasized, however, that Chapters 4 and 5 contain some non-trivial extensions of the results presented in the corresponding papers [33] and [26]. In order to keep the presentation concise and focused on a single subject, the results of [28, 29, 30, 31] have been eventually excluded. The only paper that has not been inserted here (because it is not finished yet), but whose content fits perfectly in the framework of this thesis, is [32]. Let us then conclude this introduction with a brief discussion about it.

In [32], we consider the Laplacian eigenvalue problem:

$$\begin{cases} -\Delta u = \lambda u & \text{in } \Omega := (0,1)^d, \\ u = 0 & \text{on } \partial\Omega. \end{cases} \tag{3}$$

For its numerical approximation, we use the Galerkin method, in which the Galerkin approximation space is chosen as the space generated by (tensor-product) B-splines of degree $p \geq 1$ and smoothness $k \in \{0, \ldots, p-1\}$ in each direction x_i, $i = 1, \ldots, d$. The choice $k = 0$ corresponds to the classical FEM with C^0 B-spline basis (instead of the Lagrangian basis used in Chapter 3), while the choice $k = p - 1$ corresponds to the Galerkin IgA approximation considered in Chapter 4, which uses C^{p-1} B-splines as basis functions. In this

context, the resulting sequence of discretization matrices A_n enjoys an asymptotic spectral distribution in the sense (1), and the associated symbol $\mathbf{f}$ is of the form (2) with $s = (p-k)^d$. It follows that the asymptotic spectrum of A_n is split into $(p-k)^d$ spectral branches. One of these branches is known in Engineering as 'acoustical branch', while the others are the so-called 'optical branches'; see, e.g., the appendix of [42] where this terminology is employed. In particular, in the case $k = 0$ the number of branches is p^d, while in the case $k = p-1$ the number of branches is 1. This is consistent with our findings in Chapters 3–4; see the discussion in the first two items above.

Chapter 1

Notation, definitions and mathematical background

1.1 Notation

- $\mathbb{R}^{m\times n}$ (resp. $\mathbb{C}^{m\times n}$) is the space of real (resp. complex) $m\times n$ matrices.
- If X is a matrix and α is a scalar, the matrix αX is sometimes denoted by $X\alpha$.
- If $\mathbf{x}$ is a vector and X is a matrix, then $\mathbf{x}^T$ and $\mathbf{x}^*$ (resp. X^T and X^*) are the transpose and the transpose conjugate of $\mathbf{x}$ (resp. X).
- O_m and I_m denote, respectively, the $m\times m$ zero matrix and the $m\times m$ identity matrix. Sometimes, when the dimension m is clear from the context, O and I are used instead of O_m and I_m.
- Given $X\in\mathbb{C}^{m\times m}$, $\Lambda(X)$ is the spectrum of X (the set of all the eigenvalues of X) and $\rho(X)$ is the spectral radius of X, i.e. $\rho(X) := \max_{\lambda\in\Lambda(X)}|\lambda|$. The eigenvalues of X are denoted by $\lambda_j(X)$, $j=1,\ldots,m$.
- Let $X\in\mathbb{C}^{m\times m}$ be a matrix with only real eigenvalues (e.g., a Hermitian matrix). We denote by $\lambda_{\min}(X)$ and $\lambda_{\max}(X)$ the minimal and the maximal eigenvalue of X, respectively. Unless otherwise stated, it is understood that the eigenvalues of X are labeled in non-increasing order: $\lambda_{\max}(X)=\lambda_1(X)\geq\ldots\geq\lambda_m(X)=\lambda_{\min}(X)$; in addition, we set $\lambda_j(X)=+\infty$ if $j<1$ and $\lambda_j(X)=-\infty$ if $j>m$ (this convention simplifies the presentation, as we shall see later).
- HPD and SPD stand for 'Hermitian Positive Definite' and 'Symmetric Positive Definite', respectively. Similarly, HPSD and SPSD stand for 'Hermitian Positive SemiDefinite' and 'Symmetric Positive SemiDefinite', respectively.
- If $X\in\mathbb{C}^{m\times n}$, we denote by $\sigma_j(X)$, $j=1,\ldots,\min(m,n)$, the singular values of X labeled, as usual, in non-increasing order: $\sigma_1(X)\geq\ldots\geq\sigma_{\min(m,n)}(X)$. $\sigma_1(X)$ and $\sigma_{\min(m,n)}(X)$ are also denoted by $\sigma_{\max}(X)$ and $\sigma_{\min}(X)$.
- If $p\in[1,\infty]$, the symbol $\|\cdot\|_p$ is used to denote both the p-norm of vectors and matrices:

$$\|\mathbf{x}\|_p := \begin{cases} \left(\sum_{i=1}^m |x_i|^p\right)^{1/p} & \text{if } 1\leq p<\infty, \\ \max_{i=1,\ldots,m}|x_i| & \text{if } p=\infty, \end{cases} \qquad \mathbf{x}\in\mathbb{C}^m,$$

$$\|X\|_p := \max_{\substack{\mathbf{x}\in\mathbb{C}^m \\ \mathbf{x}\neq\mathbf{0}}} \frac{\|X\mathbf{x}\|_p}{\|\mathbf{x}\|_p}, \qquad X\in\mathbb{C}^{m\times m}.$$

  $\|\cdot\|_2$ is often referred to as the spectral or Euclidean norm and is also denoted by $\|\cdot\|$.

- If $p \in [1, \infty]$, the Schatten p-norm of a matrix $X \in \mathbb{C}^{m\times m}$ is defined as the p-norm of the vector $\sigma(X) = (\sigma_1(X), \ldots, \sigma_m(X))$ formed by the singular values of X; see [7]. We denote this norm by $|||\cdot|||_p$:

$$|||X|||_p := \begin{cases} \left(\sum_{i=1}^m \sigma_i(X)^p\right)^{1/p} & \text{if } 1 \le p < \infty, \\ \sigma_{\max}(X) & \text{if } p = \infty, \end{cases} \qquad X \in \mathbb{C}^{m\times m}.$$

  The Schatten 1-norm is also called trace-norm.[1]

- $\Re(X)$ and $\Im(X)$ are, respectively, the real and the imaginary part of the (square) matrix X:

$$\Re(X) := \frac{X + X^*}{2}, \qquad \Im(X) := \frac{X - X^*}{2\mathrm{i}} \qquad (\mathrm{i} \text{ is the imaginary unit, } \mathrm{i}^2 = -1).$$

- $\kappa_p(X)$ is the condition number of the (invertible) matrix X, measured in the p-norm:

$$\kappa_p(X) := \|X\|_p \|X^{-1}\|_p.$$

  $\kappa_2(X)$ is often referred to as the spectral or Euclidean condition number and is also denoted by $\kappa(X)$.

- If $X, Y \in \mathbb{C}^{m\times m}$, $X \ge Y$ (resp. $X > Y$) means that X, Y are Hermitian and $X - Y$ is nonnegative definite (resp. positive definite).

- If $w_i : D_i \to \mathbb{C}$, $i = 1, \ldots, d$, are functions, then $w_1 \otimes \cdots \otimes w_d : D_1 \times \cdots \times D_d \to \mathbb{C}$ denotes the tensor-product function

$$(w_1 \otimes \cdots \otimes w_d)(\xi_1, \ldots, \xi_d) := w_1(\xi_1) \cdots w_d(\xi_d), \qquad \xi_i \in D_i,\ i = 1, \ldots, d.$$

  More generally, if $\mathbf{w}_i : D_i \to \mathbb{C}^{s_i \times s_i}$, $i = 1, \ldots, d$, are matrix-valued functions, then $\mathbf{w}_1 \otimes \cdots \otimes \mathbf{w}_d : D_1 \times \cdots \times D_d \to \mathbb{C}^{(s_1 \cdots s_d)\times(s_1 \cdots s_d)}$ is defined as

$$(\mathbf{w}_1 \otimes \cdots \otimes \mathbf{w}_d)(\xi_1, \ldots, \xi_d) := \mathbf{w}_1(\xi_1) \otimes \cdots \otimes \mathbf{w}_d(\xi_d), \qquad \xi_i \in D_i,\ i = 1, \ldots, d.$$

- m_d (a slanted lowercase m with subscript d) denotes the Lebesgue measure in $\mathbb{R}^d$. The Lebesgue measure in $\mathbb{R}$, m_1, is also denoted by m. Throughout this thesis, the words 'measure', 'measurable', 'a.e.', etc. always refer to the Lebesgue measure.

- $C_c(\mathbb{C})$ (resp. $C_c(\mathbb{R})$) is the space of complex-valued continuous functions defined over $\mathbb{C}$ (resp. $\mathbb{R}$) and with bounded support. Moreover, $C_c^1(\mathbb{R}) := C_c(\mathbb{R}) \cap C^1(\mathbb{R})$, where $C^1(\mathbb{R})$ is the space of complex-valued functions F defined on $\mathbb{R}$ and such that the real and imaginary parts, $\Re(F)$ and $\Im(F)$, are of class C^1 over $\mathbb{R}$ in the classical sense.

- For $z \in \mathbb{C}$ and $\epsilon > 0$, we denote by $D(z, \epsilon)$ the disk centered at z and with radius ϵ, i.e. $D(z, \epsilon) := \{w \in \mathbb{C} : |w - z| < \epsilon\}$. For $S \subseteq \mathbb{C}$ and $\epsilon > 0$, we denote by $D(S, \epsilon)$ the ϵ-expansion of S, defined as $D(S, \epsilon) := \bigcup_{z\in S} D(z, \epsilon)$.

- The words 'matrix-sequence', 'matrix-sequences', 'matrix-family', 'matrix-families' stand for 'sequence of matrices', 'sequences of matrices', 'family of matrices', 'families of matrices', respectively.

- A matrix-valued function $\mathbf{f} : D \to \mathbb{C}^{s\times s}$, defined on a measurable set $D \subseteq \mathbb{R}^d$, is said to be measurable (resp. continuous, in $L^p(D)$) if all its components $f_{ij} : D \to \mathbb{C}$, $i, j = 1, \ldots, s$, are measurable (resp. continuous, in $L^p(D)$). The space of functions $\mathbf{f} : D \to \mathbb{C}^{s\times s}$ belonging to $L^p(D)$ is sometimes denoted by $L^p(D, \mathbb{C}^{s\times s})$.

[1] My choice of using the symbol $|||\cdot|||_p$ to denote the Schatten p-norm was inspired by the fact that Bhatia, in his book [7], uses the symbol $|||\cdot|||$ for the unitarily invariant norms. Note that the Schatten p-norms are unitarily invariant, being defined in terms of singular values.

- $\mathbb{P}_p$ is the space of polynomials of degree less than or equal to p.
- $\mathbb{Q}_+^d := \{\mathbf{q} \in \mathbb{Q}^d : \ q_i > 0 \ \text{ for all } i = 1, \dots, d\}$.
- Given two sequences $\{a_n\}$ and $\{b_n\}$ with $a_n,\ b_n \geq 0$ for all n, the notation

  $$a_n = O(b_n)$$

  means that there exists a constant C, independent of n, such that $a_n \leq C b_n$ for all n.
- Given two sequences $\{a_n\}$ and $\{b_n\}$ with $a_n,\ b_n \neq 0$ for all sufficiently large n, the notation

  $$a_n \overset{n\to\infty}{\sim} b_n$$

  means that $a_n/b_n \to 1$ as $n \to \infty$.

1.1.1 Multi-index notation

Throughout the thesis, we will systematically use the multi-index notation, expounded by Tyrtyshnikov in [70, Section 6]. When discretizing a linear PDE defined over a d-dimensional domain $\Omega \subset \mathbb{R}^d$ by means of a linear numerical method, the actual computation of the numerical solution reduces to solving a certain linear system, whose coefficient matrix usually shows a d-level structure; see [70, Section 6] for the corresponding definition. As we shall see in Chapters 3–5, the multi-index notation is a powerful tool that allows us to give a compact expression of this matrix, treating the dimensionality parameter d as any other parameter involved in the considered numerical method. In this way, the dependency of the matrix structure from d is highlighted and a compact presentation is made possible.

A multi-index $\boldsymbol{i}$ is simply a vector in $\mathbb{Z}^d$; its components are denoted by $i_1, \dots, i_d$. A multi-index $\boldsymbol{i} \in \mathbb{Z}^d$ is also called a d-index.

- $\mathbf{0}$, $\mathbf{1}$, $\mathbf{2}$, ... are the vectors of all zeros, all ones, all twos, ... (their size will be clear from the context).
- If $\boldsymbol{i}, \boldsymbol{j}$ are d-indices, $\boldsymbol{i} \leq \boldsymbol{j}$ means that $i_\ell \leq j_\ell$ for all $\ell = 1, \dots, d$.
- If $\boldsymbol{h}, \boldsymbol{k}$ are d-indices such that $\boldsymbol{h} \leq \boldsymbol{k}$, the multi-index range $\boldsymbol{h}, \dots, \boldsymbol{k}$ is the set $\{\boldsymbol{j} \in \mathbb{Z}^d : \boldsymbol{h} \leq \boldsymbol{j} \leq \boldsymbol{k}\}$. We assume for the multi-index range $\boldsymbol{h}, \dots, \boldsymbol{k}$ the standard lexicographic ordering:

  $$\Big[\ \dots \ \Big[\ [\ (j_1, \dots, j_d) \ ]_{j_d = h_d, \dots, k_d} \ \Big]_{j_{d-1} = h_{d-1}, \dots, k_{d-1}} \ \dots \ \Big]_{j_1 = h_1, \dots, k_1}. \tag{1.1}$$

  For instance, in the case $d = 2$ the ordering is

  $$(h_1, h_2),\ (h_1, h_2{+}1),\ \dots,\ (h_1, k_2),\ (h_1{+}1, h_2),\ (h_1{+}1, h_2{+}1),\ \dots,\ (h_1{+}1, k_2),\ \dots\ \dots,\ (k_1, h_2),\ (k_1, h_2{+}1),\ \dots,\ (k_1, k_2).$$

- When a d-index $\boldsymbol{j}$ varies over a multi-index range $\boldsymbol{h}, \dots, \boldsymbol{k}$ (this is sometimes written as $\boldsymbol{j} = \boldsymbol{h}, \dots, \boldsymbol{k}$ or $(j_1, \dots, j_d) = (h_1, \dots, h_d), \dots, (k_1, \dots, k_d)$), it is always understood that $\boldsymbol{j}$ varies from $\boldsymbol{h}$ to $\boldsymbol{k}$ following the specific ordering (1.1). For instance, if $\boldsymbol{m} \in \mathbb{N}^d$ and if we write $X = [x_{\boldsymbol{i}\boldsymbol{j}}]_{\boldsymbol{i},\boldsymbol{j}=\boldsymbol{1}}^{\boldsymbol{m}}$, then X is a matrix in $\mathbb{C}^{(m_1 \cdots m_d) \times (m_1 \cdots m_d)}$ whose components are indexed by two d-indices $\boldsymbol{i}, \boldsymbol{j}$, both varying over the multi-index range $\mathbf{1}, \dots, \boldsymbol{m}$ according to (1.1). Similarly, if $\mathbf{x} = [x_{\boldsymbol{i}}]_{\boldsymbol{i}=\boldsymbol{1}}^{\boldsymbol{m}}$ then $\mathbf{x}$ is a vector in $\mathbb{C}^{m_1 \cdots m_d}$ whose components $x_{\boldsymbol{i}}$, $\boldsymbol{i} = \mathbf{1}, \dots, \boldsymbol{m}$, are ordered in accordance with (1.1): the first component is $x_{\mathbf{1}} = x_{(1,\dots,1,1)}$, the second component is $x_{(1,\dots,1,2)}$, and so on until the last component, which is $x_{\boldsymbol{m}} = x_{(m_1,\dots,m_d)}$.

- If $\boldsymbol{i}, \boldsymbol{j} \in \mathbb{Z}^d$ are multi-indices, $\boldsymbol{i} \preceq \boldsymbol{j}$ means that $\boldsymbol{i}$ precedes (or equals) $\boldsymbol{j}$ in the lexicographic ordering (which is a total ordering on $\mathbb{Z}^d$). Moreover, we define

$$\boldsymbol{i} \wedge \boldsymbol{j} := \begin{cases} \boldsymbol{i} & \text{if } \boldsymbol{i} \preceq \boldsymbol{j}, \\ \boldsymbol{j} & \text{if } \boldsymbol{i} \succ \boldsymbol{j}. \end{cases} \tag{1.2}$$

  Note that $\boldsymbol{i} \wedge \boldsymbol{j}$ is the minimum among $\boldsymbol{i}$ and $\boldsymbol{j}$ with respect to the lexicographic ordering.

- Given $\boldsymbol{h}, \boldsymbol{k} \in \mathbb{Z}^d$ with $\boldsymbol{h} \le \boldsymbol{k}$, the notation $\sum_{\boldsymbol{j}=\boldsymbol{h}}^{\boldsymbol{k}}$ indicates the summation over all $\boldsymbol{j}$ in the multi-index range $\boldsymbol{h}, \dots, \boldsymbol{k}$.

- For a multi-index $\boldsymbol{m} \in \mathbb{N}^d$, $N(\boldsymbol{m}) := \prod_{j=1}^d m_j$ and $\boldsymbol{m} \to \infty$ means that $\min(m_1, \dots, m_d) \to \infty$.

- Operations involving multi-indices that do not have a meaning when considering multi-indices as normal vectors must be always understood in the componentwise sense. For instance, $\boldsymbol{np} = (n_1 p_1, \dots, n_d p_d)$, $\alpha \boldsymbol{i}/\boldsymbol{j} = (\alpha i_1/j_1, \dots, \alpha i_d/j_d)$ for all $\alpha \in \mathbb{C}$ (of course, the division is defined when $j_1, \dots, j_d \neq 0$), $\boldsymbol{i}^2 = (i_1^2, \dots, i_d^2)$, $\boldsymbol{i} \bmod \boldsymbol{m} = (i_1 \bmod m_1, \dots, i_d \bmod m_d)$, and so on.

- When a multi-index appears as subscript or superscript, we often suppress the parentheses to simplify the notation. For instance, the component of the vector $\mathbf{x} = [x_{\boldsymbol{i}}]_{\boldsymbol{i}=\boldsymbol{1}}^{\boldsymbol{m}}$ corresponding to the multi-index $\boldsymbol{i}$ is denoted by $x_{\boldsymbol{i}}$ or by $x_{i_1,\dots,i_d}$, and we preferably avoid the heavy notation $x_{(i_1,\dots,i_d)}$.

1.2 Preliminaries on Linear Algebra and Matrix Analysis

We recall in this section some results from Linear Algebra and Matrix Analysis that will be used later on. Most of the results that we are going to see can be found in [7] or [8].

For every $X \in \mathbb{C}^{m\times m}$, $\|X\|_1$ is the maximum among the 1-norms of the column vectors of X, while $\|X\|_\infty$ is the maximum among the 1-norms of the row vectors of X. As a consequence, $\|X\|_1 = \|X^T\|_\infty$. An important relation between the p-norms with $p = 1, 2, \infty$ is the following:

$$\|X\| = \|X\|_2 \le \sqrt{\|X\|_1 \|X\|_\infty} = \sqrt{\|X\|_\infty \|X^T\|_\infty}; \tag{1.3}$$

see [8, p. 121].

Given $X \in \mathbb{C}^{m\times m}$, we know from the Singular Value Decomposition (SVD) that $\operatorname{rank}(X)$ is the number of nonzero singular values of X and

$$\|X\| = \sigma_{\max}(X) = \max_{\|\mathbf{u}\|=\|\mathbf{v}\|=1} \mathbf{u}^* X \mathbf{v}. \tag{1.4}$$

As a consequence, $|||X|||_\infty = \|X\|$ and

$$|||X|||_1 = \sum_{i=1}^m \sigma_i(X) \le \operatorname{rank}(X)\|X\| \le m\|X\|, \qquad \forall X \in \mathbb{C}^{m\times m}. \tag{1.5}$$

From the SVD we also know that the formula $\|X^{-1}\|_2 = \frac{1}{\sigma_{\min}(X)}$ holds whenever X is invertible, hence

$$\kappa(X) = \frac{\sigma_{\max}(X)}{\sigma_{\min}(X)}, \qquad \text{for all invertible matrices } X. \tag{1.6}$$

If $X \in \mathbb{C}^{m\times m}$ is a normal matrix, i.e. $XX^* = X^*X$, then X is unitarily diagonalizable, meaning that there exist a unitary matrix U and a diagonal matrix D such that $X = UDU^*$. Using this, it can be shown that the singular values of X coincide with the moduli of the eigenvalues, $|\lambda_j(X)|$, $j = 1, \dots, m$. Consequently,

$\|X\| = \rho(X)$ and $|||X|||_1 = \sum_{j=1}^m |\lambda_j(X)|$. Note that, if X is Hermitian ($X^* = X$) or skew-Hermitian ($X^* = -X$), then X is normal.

For any square matrix X, $\Re(X)$ and $\Im(X)$ are Hermitian matrices and $X = \Re(X) + \mathrm{i}\,\Im(X)$. If λ is an eigenvalue of X and $\mathbf{x}$ is a corresponding eigenvector, then, by the minimax principle [7, 8], we have

$$\lambda = \frac{\mathbf{x}^* X \mathbf{x}}{\mathbf{x}^*\mathbf{x}} = \frac{\mathbf{x}^* \Re(X) \mathbf{x}}{\mathbf{x}^*\mathbf{x}} + \mathrm{i}\,\frac{\mathbf{x}^* \Im(X) \mathbf{x}}{\mathbf{x}^*\mathbf{x}} \in [\lambda_{\min}(\Re(X)), \lambda_{\max}(\Re(X))] \times [\lambda_{\min}(\Im(X)), \lambda_{\max}(\Im(X))] \subset \mathbb{C}.$$

This implies that

$$\Lambda(X) \subseteq [\lambda_{\min}(\Re(X)), \lambda_{\max}(\Re(X))] \times [\lambda_{\min}(\Im(X)), \lambda_{\max}(\Im(X))]\,, \qquad \text{for all square matrices } X. \tag{1.7}$$

Other consequences of the minimax principle are the following:

$$\lambda_{\min}(X + Y) \geq \lambda_{\min}(X) + \lambda_{\min}(Y), \quad \text{for all Hermitian matrices } X, Y, \tag{1.8}$$

$$\lambda_{\max}(X + Y) \leq \lambda_{\max}(X) + \lambda_{\max}(Y), \quad \text{for all Hermitian matrices } X, Y, \tag{1.9}$$

$$\lambda_j(X) \geq \lambda_j(Y), \quad \forall\, j = 1, \ldots, m, \quad \text{for all Hermitian matrices } X, Y \in \mathbb{C}^{m\times m} \text{ such that } X \geq Y. \tag{1.10}$$

An important relation between the singular values of X and the eigenvalues of $\Re(X)$ is provided by the Fan-Hoffman theorem [7, Proposition III.5.1]. We report below the corresponding statement, together with the statement of the Ky-Fan theorem [7, Proposition III.5.3]. The latter provides a relation between the real parts of the eigenvalues of X and the eigenvalues of $\Re(X)$. Recall that the eigenvalues of a Hermitian matrix, such as $\Re(X)$, are labeled in non-increasing order (see Section 1.1).

Theorem 1.1 (**Fan-Hoffman**). *Let $X \in \mathbb{C}^{m\times m}$, then*

$$\sigma_j(X) \geq \lambda_j(\Re(X)), \quad \forall\, j = 1, \ldots, m.$$

We shall see in Chapters 3–4 that the Fan-Hoffman theorem is very useful for estimating the spectral condition number (1.6) of a non-singular matrix X coming from the numerical approximation of a PDE.

Theorem 1.2 (**Ky-Fan**). *Let $X \in \mathbb{C}^{m\times m}$ and label the eigenvalues of X so that $\Re(\lambda_1(X)) \geq \ldots \geq \Re(\lambda_m(X))$. Then*

$$\sum_{j=1}^{k} \Re(\lambda_j(X)) \leq \sum_{j=1}^{k} \lambda_j(\Re(X)), \tag{1.11}$$

for all $k = 1, \ldots, m$. Moreover, for $k = m$, the equality holds in (1.11).

We now provide the statement of two classical interlacing theorems; see [7, Corollary III.1.5] for the first one and [7, p. 63] for the second one. We recall that Y is a principal submatrix of $X \in \mathbb{C}^{m\times m}$ if there exists $E \subset \{1, \ldots, m\}$ such that Y is obtained from X by removing the rows and columns corresponding to indices $i \in E$. In this case, Y is called the principal submatrix of X corresponding to the set of indices $F = \{1, \ldots, m\} \backslash E$.

Theorem 1.3 (**Cauchy's interlacing theorem**). *Let $X \in \mathbb{C}^{m\times m}$ be Hermitian and let Y be a principal submatrix of X of order ℓ. Then*

$$\lambda_j(X) \geq \lambda_j(Y) \geq \lambda_{j+m-\ell}(X), \qquad \forall\, j = 1, \ldots, \ell.$$

In the statement of Theorem 1.4 we use the convention introduced in Section 1.1 for a matrix $X \in \mathbb{C}^{m\times m}$ with only real eigenvalues, namely $\lambda_i(X) = -\infty$ if $i < 1$ and $\lambda_i(X) = +\infty$ if $i > m$.

Theorem 1.4. *Let $Y = X + E$, where $X, E \in \mathbb{C}^{m\times m}$ are Hermitian. Let $k^+, k^- \geq 0$ be respectively the number of positive and the number of negative eigenvalues of E, i.e.*

$$k^+ := \#\{j \in \{1, \ldots, m\} : \lambda_j(E) > 0\}, \qquad k^- := \#\{j \in \{1, \ldots, m\} : \lambda_j(E) < 0\}.$$

Then

$$\lambda_{j-k^+}(X) \geq \lambda_j(Y) \geq \lambda_{j+k^-}(X), \qquad \forall\, j = 1, \ldots, m.$$

1.2.1 Tensor products and direct sums

If X, Y are matrices of any dimension, say $X \in \mathbb{C}^{m_1 \times m_2}$ and $Y \in \mathbb{C}^{\ell_1 \times \ell_2}$, then

- $X \otimes Y$ is the tensor (or Kronecker) product of X and Y, that is the $m_1\ell_1 \times m_2\ell_2$ matrix

$$X \otimes Y := [x_{ij}Y]_{\substack{i=1,\dots,m_1 \\ j=1,\dots,m_2}} = \begin{bmatrix} x_{11}Y & \cdots & x_{1m_2}Y \\ \vdots & & \vdots \\ x_{m_11}Y & \cdots & x_{m_1m_2}Y \end{bmatrix};$$

- $X \oplus Y$ is the direct sum of X and Y, that is the $(m_1 + \ell_1) \times (m_2 + \ell_2)$ matrix

$$X \oplus Y := \left[\begin{array}{c|c} X & O \\ \hline O & Y \end{array}\right].$$

Tensor products and direct sums possess a lot of nice algebraic properties.

(i) Associativity: for all matrices X, Y, Z, $(X \otimes Y) \otimes Z = X \otimes (Y \otimes Z)$ and $(X \oplus Y) \oplus Z = X \oplus (Y \oplus Z)$. This means that we can omit parentheses in expressions like $X_1 \otimes X_2 \otimes \cdots \otimes X_d$ or $X_1 \oplus X_2 \oplus \cdots \oplus X_d$.

(ii) Multi-index notation (for tensor products): if we have d matrices $X_k \in \mathbb{C}^{m_k \times m_k}$, $k = 1, \dots, d$, then

$$(X_1 \otimes X_2 \otimes \cdots \otimes X_d)_{\boldsymbol{ij}} = (X_1)_{i_1 j_1}(X_2)_{i_2 j_2} \cdots (X_d)_{i_d j_d}, \qquad \forall \boldsymbol{i}, \boldsymbol{j} = \mathbf{1}, \dots, \boldsymbol{m}, \tag{1.12}$$

where $\boldsymbol{m} := (m_1, m_2, \dots, m_d)$. This means that, for all $\boldsymbol{i}, \boldsymbol{j}$ in the multi-index range $\mathbf{1}, \dots, \boldsymbol{m}$, the $(\boldsymbol{i}, \boldsymbol{j})$-th entry of $X_1 \otimes X_2 \otimes \cdots \otimes X_d$ is given by (1.12). Note that it makes sense to talk about the $(\boldsymbol{i}, \boldsymbol{j})$-th entry of $X_1 \otimes X_2 \otimes \cdots \otimes X_d$, because we have fixed for the set $\mathbf{1}, \dots, \boldsymbol{m}$ the lexicographic ordering (1.1). Note also that (1.12) can be rewritten as

$$X_1 \otimes \cdots \otimes X_d = \left[(X_1)_{i_1 j_1}(X_2)_{i_2 j_2} \cdots (X_d)_{i_d j_d}\right]_{\boldsymbol{i},\boldsymbol{j}=\mathbf{1}}^{\boldsymbol{m}}.$$

The equality (1.12) is of fundamental importance and, indeed, it motivates the introduction of multi-indices for indexing the entries of a matrix formed by a sum of one or more tensor products. To understand better the importance of (1.12), try to write the (i, j)-th entry of $X_1 \otimes X_2 \otimes \cdots \otimes X_d$ as a function of two linear indices $i, j = 1, \dots, N(\boldsymbol{m})$.

(iii) The relations $(X_1 \otimes Y_1)(X_2 \otimes Y_2) = (X_1X_2) \otimes (Y_1Y_2)$ and $(X_1 \oplus Y_1)(X_2 \oplus Y_2) = (X_1X_2) \oplus (Y_1Y_2)$ hold whenever X_1, X_2 can be multiplied and Y_1, Y_2 can be multiplied.

(iv) For all matrices X, Y, $(X \otimes Y)^* = X^* \otimes Y^*$, $(X \oplus Y)^* = X^* \oplus Y^*$ and $(X \otimes Y)^T = X^T \otimes Y^T$, $(X \oplus Y)^T = X^T \oplus Y^T$.

(v) Bilinearity (of tensor products): $(\alpha_1 X_1 + \alpha_2 X_2) \otimes (\beta_1 Y_1 + \beta_2 Y_2) = \alpha_1\beta_1(X_1 \otimes Y_1) + \alpha_1\beta_2(X_1 \otimes Y_2) + \alpha_2\beta_1(X_2 \otimes Y_1) + \alpha_2\beta_2(X_2 \otimes Y_2)$ for all $\alpha_1, \alpha_2, \beta_1, \beta_2 \in \mathbb{C}$ and for all matrices X_1, X_2, Y_1, Y_2 such that X_1, X_2 are summable and Y_1, Y_2 are summable.

From (i)–(v), a lot of other interesting properties follow. We recall some of them. If X, Y are invertible, then $X \otimes Y$ is invertible, its inverse being $X^{-1} \otimes Y^{-1}$. If X, Y are normal (resp. Hermitian, symmetric, unitary) then $X \otimes Y$ is also normal (resp. Hermitian, symmetric, unitary). If $X \in \mathbb{C}^{m \times m}$ and $Y \in \mathbb{C}^{\ell \times \ell}$, then the eigenvalues and the singular values of $X \otimes Y$ (resp. $X \oplus Y$) are $\{\lambda_i(X)\lambda_j(Y) : i = 1, \dots, m,\ j = 1, \dots, \ell\}$ and

$\{\sigma_i(X)\sigma_j(Y) : i = 1,\dots,m,\ \ j = 1,\dots,\ell\}$ (resp. $\{\lambda_i(X) : i = 1,\dots,m\} \bigcup \{\lambda_j(Y) : j = 1,\dots,\ell\}$, and $\{\sigma_i(X) : i = 1,\dots,m\} \bigcup \{\sigma_j(Y) : j = 1,\dots,\ell\}$). As a consequence, for all $X \in \mathbb{C}^{m\times m}$ and $Y \in \mathbb{C}^{\ell\times\ell}$,

$$\|X \otimes Y\| = \|X\|\,\|Y\|, \qquad \|X \oplus Y\| = \max(\|X\|, \|Y\|), \tag{1.13}$$

$$|||X \otimes Y|||_1 = |||X|||_1\,|||Y|||_1, \qquad |||X \oplus Y|||_1 = |||X|||_1 + |||Y|||_1, \tag{1.14}$$

$$\mathrm{rank}(X \otimes Y) = \mathrm{rank}(X)\mathrm{rank}(Y), \qquad \mathrm{rank}(X \oplus Y) = \mathrm{rank}(X) + \mathrm{rank}(Y), \tag{1.15}$$

and if X, Y are HPD (resp. HPSD), then $X \otimes Y$ is HPD (resp. HPSD), with

$$\lambda_{\min}(X \otimes Y) = \lambda_{\min}(X)\lambda_{\min}(Y), \qquad \lambda_{\max}(X \otimes Y) = \lambda_{\max}(X)\lambda_{\max}(Y). \tag{1.16}$$

In particular,

$$X \otimes Y \geq X' \otimes Y', \quad \text{for all HPSD matrices } X, Y, X', Y' \text{ such that } X \geq X' \text{ and } Y \geq Y', \tag{1.17}$$

because $X \otimes Y - X' \otimes Y' = (X - X') \otimes Y + X' \otimes (Y - Y')$ is a sum of two HPSD matrices. We also point out the following property: suppose we are given $2d$ matrices $X_1, \dots, X_d, Y_1, \dots, Y_d$ with $X_i, Y_i \in \mathbb{C}^{m_i\times m_i}$ for all $i = 1,\dots,d$, then

$$\mathrm{rank}(X_1 \otimes \cdots \otimes X_d - Y_1 \otimes \cdots \otimes Y_d) \leq \sum_{i=1}^{d} \mathrm{rank}(X_i - Y_i) m_1 \cdots m_{i-1} m_{i+1} \cdots m_d = N(\boldsymbol{m}) \sum_{i=1}^{d} \frac{\mathrm{rank}(X_i - Y_i)}{m_i}, \tag{1.18}$$

where, of course, $\boldsymbol{m} = (m_1, \dots, m_d)$. This is true because

$$\begin{aligned}
\mathrm{rank}(X_1 \otimes \cdots \otimes X_d - Y_1 \otimes \cdots \otimes Y_d) &= \mathrm{rank}\left(\sum_{i=1}^{d} \ Y_1 \otimes \cdots \otimes Y_{i-1} \otimes (X_i - Y_i) \otimes X_{i+1} \otimes \cdots \otimes X_d \right) \\
&\leq \sum_{i=1}^{d} \mathrm{rank}\left(Y_1 \otimes \cdots \otimes Y_{i-1} \otimes (X_i - Y_i) \otimes X_{i+1} \otimes \cdots \otimes X_d\right) \\
&= \sum_{i=1}^{d} \mathrm{rank}(Y_1 \otimes \cdots \otimes Y_{i-1})\mathrm{rank}(X_i - Y_i)\mathrm{rank}(X_{i+1} \otimes \cdots \otimes X_d) \\
&\leq \sum_{i=1}^{d} m_1 \cdots m_{i-1}\mathrm{rank}(X_i - Y_i) m_{i+1} \cdots m_d.
\end{aligned}$$

A property of tensor products, which can be deduced from the definition but is not as popular as the previous ones, is given in Lemma 1.1; see also [38].

Lemma 1.1. *For all $\boldsymbol{m} \in \mathbb{N}^2$ there exists a permutation matrix $\Pi_{\boldsymbol{m}} \in \mathbb{C}^{N(\boldsymbol{m})\times N(\boldsymbol{m})}$ such that*

$$X_2 \otimes X_1 = \Pi_{\boldsymbol{m}}(X_1 \otimes X_2)\Pi_{\boldsymbol{m}}^T, \quad \forall X_1 \in \mathbb{C}^{m_1\times m_1},\ \forall X_2 \in \mathbb{C}^{m_2\times m_2}. \tag{1.19}$$

Proof. Let $\Pi_{\boldsymbol{m}}$ be the permutation matrix associated with the permutation σ of $\{1,\dots,m_1m_2\}$ given by

$$\sigma := [1, m_2+1, 2m_2+1, \dots, (m_1-1)m_2+1, 2, m_2+2, 2m_2+2, \dots, (m_1-1)m_2+2, \dots\dots, m_2, 2m_2, 3m_2 \dots, m_1m_2],$$

i.e., by

$$\sigma(i) := ((i-1) \bmod m_1)m_2 + \left\lfloor \frac{i-1}{m_1} \right\rfloor + 1, \qquad i = 1,\dots,m_1m_2.$$

In other words, $\Pi_{\boldsymbol{m}}$ is the matrix whose rows are (in this order) $\mathbf{e}_{\sigma(i)},\ i = 1,\dots,m_1m_2$, where $\mathbf{e}_i,\ i = 1,\dots,m_1m_2$, are the vectors of the canonical basis of $\mathbb{C}^{m_1m_2}$. It can be verified that $\Pi_{\boldsymbol{m}}$ defined in this way satisfies (1.19) for all $X_1 \in \mathbb{C}^{m_1\times m_1}$ and $X_2 \in \mathbb{C}^{m_2\times m_2}$. The verification can be done componentwise, by showing that the (i,j)-th entry of the first matrix in (1.19) is equal to the (i,j)-th entry of the second matrix, for all $i, j = 1,\dots,m_1m_2$. □

Lemma 1.1 says that the tensor product of two matrices is 'almost' commutative. It is important to notice that the permutation matrix $\Pi_{\boldsymbol{m}}$ depend only on $\boldsymbol{m}$ and not on the specific matrices X_1, X_2. By induction, we now extend the result of Lemma 1.1 to the case of tensor products with more than two factors.

Lemma 1.2. *For all $\boldsymbol{m} \in \mathbb{N}^d$ and all permutations σ of the set $\{1, \dots, d\}$, there exists a permutation matrix $\Pi_{\boldsymbol{m};\sigma} \in \mathbb{C}^{N(\boldsymbol{m}) \times N(\boldsymbol{m})}$ such that*

$$X_{\sigma(1)} \otimes \cdots \otimes X_{\sigma(d)} = \Pi_{\boldsymbol{m};\sigma}(X_1 \otimes \cdots \otimes X_d)\Pi_{\boldsymbol{m};\sigma}^T, \quad \forall X_1 \in \mathbb{C}^{m_1 \times m_1}, \dots, \forall X_d \in \mathbb{C}^{m_d \times m_d}.$$

Proof. The case $d = 1$ is trivial. For $d = 2$, the result is clear when σ is the identity, and it has been proved in Lemma 1.1 when $\sigma = [2, 1]$. Now we fix $d \geq 3$, we assume the result is true for $d - 1$, and we prove that it is true also for d. Let $\boldsymbol{m} \in \mathbb{N}^d$ and let σ be a permutation of $\{1, \dots, d\}$. Denote by i the index such that $\sigma(i) = d$, and let τ be the permutation of $\{1, \dots, d-1\}$ defined as $\tau(j) := \sigma(j)$ for $j = 1, \dots, i-1$ and $\tau(j) := \sigma(j+1)$ for $j = i, \dots, d-1$. Then, keeping in mind the properties of tensor products, for all $X_1, \dots, X_d$ with $X_j \in \mathbb{C}^{m_j \times m_j}$, $j = 1, \dots, d$, we have

$$\begin{aligned}
&X_{\sigma(1)} \otimes \cdots \otimes X_{\sigma(d)} = X_{\sigma(1)} \otimes \cdots \otimes X_{\sigma(i-1)} \otimes X_d \otimes X_{\sigma(i+1)} \otimes \cdots \otimes X_{\sigma(d)} \\
&= X_{\sigma(1)} \otimes \cdots \otimes X_{\sigma(i-1)} \otimes \left[\Pi_{(m_{\sigma(i+1)} \cdots m_{\sigma(d)}, m_d)}(X_{\sigma(i+1)} \otimes \cdots \otimes X_{\sigma(d)} \otimes X_d)\Pi_{(m_{\sigma(i+1)} \cdots m_{\sigma(d)}, m_d)}^T\right] \quad \text{(Lemma 1.1)} \\
&= \left(I_{m_{\sigma(1)} \cdots m_{\sigma(i-1)}} \otimes \Pi_{(m_{\sigma(i+1)} \cdots m_{\sigma(d)}, m_d)}\right)(X_{\sigma(1)} \otimes \cdots \otimes X_{\sigma(i-1)} \otimes X_{\sigma(i+1)} \otimes \cdots \otimes X_{\sigma(d)} \otimes X_d)\left(I_{m_{\sigma(1)} \cdots m_{\sigma(i-1)}} \otimes \Pi_{(m_{\sigma(i+1)} \cdots m_{\sigma(d)}, m_d)}^T\right) \\
&= \left(I_{m_{\sigma(1)} \cdots m_{\sigma(i-1)}} \otimes \Pi_{(m_{\sigma(i+1)} \cdots m_{\sigma(d)}, m_d)}\right)(X_{\tau(1)} \otimes \cdots \otimes X_{\tau(d-1)} \otimes X_d)\left(I_{m_{\sigma(1)} \cdots m_{\sigma(i-1)}} \otimes \Pi_{(m_{\sigma(i+1)} \cdots m_{\sigma(d)}, m_d)}\right)^T \\
&= \left(I_{m_{\sigma(1)} \cdots m_{\sigma(i-1)}} \otimes \Pi_{(m_{\sigma(i+1)} \cdots m_{\sigma(d)}, m_d)}\right)\left\{\left[\Pi_{(m_1, \dots, m_{d-1});\tau}(X_1 \otimes \cdots \otimes X_{d-1})\Pi_{(m_1, \dots, m_{d-1});\tau}^T\right] \otimes X_d\right\} \cdot \\
&\quad \cdot \left(I_{m_{\sigma(1)} \cdots m_{\sigma(i-1)}} \otimes \Pi_{(m_{\sigma(i+1)} \cdots m_{\sigma(d)}, m_d)}\right)^T \quad \text{(induction hypothesis)} \\
&= \left(I_{m_{\sigma(1)} \cdots m_{\sigma(i-1)}} \otimes \Pi_{(m_{\sigma(i+1)} \cdots m_{\sigma(d)}, m_d)}\right)(\Pi_{(m_1, \dots, m_{d-1});\tau} \otimes I_{m_d})(X_1 \otimes \cdots \otimes X_{d-1} \otimes X_d)(\Pi_{(m_1, \dots, m_{d-1});\tau} \otimes I_{m_d})^T \cdot \\
&\quad \cdot \left(I_{m_{\sigma(1)} \cdots m_{\sigma(i-1)}} \otimes \Pi_{(m_{\sigma(i+1)} \cdots m_{\sigma(d)}, m_d)}\right)^T = \Pi_{\boldsymbol{m};\sigma}(X_1 \otimes \cdots \otimes X_d)\Pi_{\boldsymbol{m};\sigma}^T,
\end{aligned}$$

where $\Pi_{\boldsymbol{m};\sigma} := \left(I_{m_{\sigma(1)} \cdots m_{\sigma(i-1)}} \otimes \Pi_{(m_{\sigma(i+1)} \cdots m_{\sigma(d)}, m_d)}\right)(\Pi_{(m_1, \dots, m_{d-1});\tau} \otimes I_{m_d})$ is a permutation matrix, being a product of two permutation matrices. □

Now we turn to the 'distributive properties' of tensor products with respect to direct sums. Again, it turns out that these properties hold modulo permutation transformations which depend only on the dimensions of the involved matrices.

Remark 1.1. From the definition of tensor products and direct sums, for all matrices $X_1, \dots, X_d, Y$ we have

$$(X_1 \oplus X_2 \oplus \cdots \oplus X_d) \otimes Y = (X_1 \otimes Y) \oplus (X_2 \otimes Y) \oplus \cdots \oplus (X_d \otimes Y).$$

Lemma 1.3. *For all $\ell \in \mathbb{N}$ and $\boldsymbol{m} \in \mathbb{N}^2$ there exists a permutation matrix $Q_{\ell,\boldsymbol{m}} \in \mathbb{C}^{\ell(m_1+m_2) \times \ell(m_1+m_2)}$ such that*

$$X \otimes (Y_1 \oplus Y_2) = Q_{\ell,\boldsymbol{m}}\left[(X \otimes Y_1) \oplus (X \otimes Y_2)\right]Q_{\ell,\boldsymbol{m}}^T, \quad \forall X \in \mathbb{C}^{\ell \times \ell}, \ \forall Y_1 \in \mathbb{C}^{m_1 \times m_1}, \ \forall Y_2 \in \mathbb{C}^{m_2 \times m_2}.$$

Proof. Let $X \in \mathbb{C}^{\ell \times \ell}$, $Y_1 \in \mathbb{C}^{m_1 \times m_1}$, $Y_2 \in \mathbb{C}^{m_2 \times m_2}$. Then, keeping in mind the properties of tensor products and direct sums,

$$\begin{aligned}
X \otimes (Y_1 \oplus Y_2) &= \Pi_{(m_1+m_2,\ell)}\left[(Y_1 \oplus Y_2) \otimes X\right]\Pi_{(m_1+m_2,\ell)}^T \quad \text{(Lemma 1.1)} \\
&= \Pi_{(m_1+m_2,\ell)}\left[(Y_1 \otimes X) \oplus (Y_2 \otimes X)\right]\Pi_{(m_1+m_2,\ell)}^T \quad \text{(Remark 1.1)} \\
&= \Pi_{(m_1+m_2,\ell)}\left\{\left[\Pi_{(\ell,m_1)}(X \otimes Y_1)\Pi_{(\ell,m_1)}^T\right] \oplus \left[\Pi_{(\ell,m_2)}(X \otimes Y_2)\Pi_{(\ell,m_2)}^T\right]\right\}\Pi_{(m_1+m_2,\ell)}^T \quad \text{(Lemma 1.1)}
\end{aligned}$$

$$= \Pi_{(m_1+m_2,\ell)} \left\{ (\Pi_{(\ell,m_1)} \oplus \Pi_{(\ell,m_2)}) \left[(X \otimes Y_1) \oplus (X \otimes Y_2) \right] (\Pi_{(\ell,m_1)} \oplus \Pi_{(\ell,m_2)})^T \right\} \Pi^T_{(m_1+m_2,\ell)}$$
$$= Q_{\ell,\boldsymbol{m}} \left[(X \otimes Y_1) \oplus (X \otimes Y_2) \right] Q^T_{\ell,\boldsymbol{m}},$$

where $Q_{\ell,\boldsymbol{m}} := \Pi_{(m_1+m_2,\ell)}(\Pi_{(\ell,m_1)} \oplus \Pi_{(\ell,m_2)})$ is a permutation matrix, being a product of two permutation matrices. □

Lemma 1.4. *For all $\ell \in \mathbb{N}$ and $\boldsymbol{m} \in \mathbb{N}^d$ there exists a permutation matrix $Q_{\ell,\boldsymbol{m}} \in \mathbb{C}^{\ell(m_1+\ldots+m_d)\times\ell(m_1+\ldots+m_d)}$ such that*

$$X \otimes (Y_1 \oplus \cdots \oplus Y_d) = Q_{\ell,\boldsymbol{m}} [(X \otimes Y_1) \oplus \cdots \oplus (X \otimes Y_d)] Q^T_{\ell,\boldsymbol{m}}, \quad \forall X \in \mathbb{C}^{\ell\times\ell},\ \forall Y_1 \in \mathbb{C}^{m_1\times m_1}, \ldots,\ \forall Y_d \in \mathbb{C}^{m_d\times m_d}.$$

Proof. The case $d = 1$ is trivial. For $d = 2$ the result has been proved in Lemma 1.3. Now we fix $d \geq 3$, we assume the result is true for $d-1$, and we prove that it is true also for d. Let $\ell \in \mathbb{N}$, $\boldsymbol{m} \in \mathbb{N}^d$. Then, for all $X \in \mathbb{C}^{\ell\times\ell}$ and all $Y_1, \ldots, Y_d$ with $Y_j \in \mathbb{C}^{m_j\times m_j}$, $j = 1, \ldots, d$, we have

$$\begin{aligned}
&X \otimes (Y_1 \oplus \cdots \oplus Y_d) = Q_{\ell,(m_1,m_2+\ldots+m_d)} \{ (X \otimes Y_1) \oplus [X \otimes (Y_2 \oplus \cdots \oplus Y_d)] \} Q^T_{\ell,(m_1,m_2+\ldots+m_d)} \quad \text{(Lemma 1.3)} \\
&= Q_{\ell,(m_1,m_2+\ldots+m_d)} \left\{ (X \otimes Y_1) \oplus [Q_{\ell,(m_2,\ldots,m_d)}((X \otimes Y_2) \oplus \cdots \oplus (X \otimes Y_d)) Q^T_{\ell,(m_2,\ldots,m_d)}] \right\} Q^T_{\ell,(m_1,m_2+\ldots+m_d)} \quad \text{(induction hypothesis)} \\
&= Q_{\ell,(m_1,m_2+\ldots+m_d)} \left\{ (I_{\ell m_1} \oplus Q_{\ell,(m_2,\ldots,m_d)})[(X \otimes Y_1) \oplus (X \otimes Y_2) \oplus \cdots \oplus (X \otimes Y_d)](I_{\ell m_1} \oplus Q_{\ell,(m_2,\ldots,m_d)})^T \right\} Q^T_{\ell,(m_1,m_2+\ldots+m_d)} \\
&= Q_{\ell,\boldsymbol{m}} [(X \otimes Y_1) \oplus \cdots \oplus (X \otimes Y_d)] Q^T_{\ell,\boldsymbol{m}},
\end{aligned}$$

where $Q_{\ell,\boldsymbol{m}} := Q_{\ell,(m_1,m_2+\ldots+m_d)}(I_{\ell m_1} \oplus Q_{\ell,(m_2,\ldots,m_d)})$. □

Lemma 1.5. *For all $n_1^{(k)}, n_2^{(k)} \in \mathbb{N}$, $k = 1, \ldots, d$, there exists a permutation matrix $P_{n_1^{(1)},n_2^{(1)},n_1^{(2)},n_2^{(2)},\ldots,n_1^{(d)},n_2^{(d)}}$ of dimension $\prod_{k=1}^d (n_1^{(k)} + n_2^{(k)})$ such that*

$$\bigotimes_{k=1}^d (X_1^{(k)} \oplus X_2^{(k)}) = P_{n_1^{(1)},n_2^{(1)},n_1^{(2)},n_2^{(2)},\ldots,n_1^{(d)},n_2^{(d)}} \left[\bigoplus_{i_1=1}^2 \cdots \bigoplus_{i_d=1}^2 (X_{i_1}^{(1)} \otimes \cdots \otimes X_{i_d}^{(d)}) \right] P^T_{n_1^{(1)},n_2^{(1)},n_1^{(2)},n_2^{(2)},\ldots,n_1^{(d)},n_2^{(d)}},$$

for all matrices $X_1^{(k)}, X_2^{(k)}$, $k = 1, \ldots, d$, with $X_1^{(k)} \in \mathbb{C}^{n_1^{(k)}\times n_1^{(k)}}$ and $X_2^{(k)} \in \mathbb{C}^{n_2^{(k)}\times n_2^{(k)}}$.

Proof. For $d = 1$ the result is clear. Fix $d \geq 2$, assume the result holds for $d-1$, and let us prove it for d. We have

$$\begin{aligned}
&\bigotimes_{k=1}^d (X_1^{(k)} \oplus X_2^{(k)}) = (X_1^{(1)} \oplus X_2^{(1)}) \otimes \left[\bigotimes_{k=2}^d (X_1^{(k)} \oplus X_2^{(k)}) \right] \\
&= (X_1^{(1)} \oplus X_2^{(1)}) \otimes \left\{ P_{n_1^{(2)},n_2^{(2)},\ldots,n_1^{(d)},n_2^{(d)}} \left[\bigoplus_{i_2=1}^2 \cdots \bigoplus_{i_d=1}^2 (X_{i_2}^{(2)} \otimes \cdots \otimes X_{i_d}^{(d)}) \right] P^T_{n_1^{(2)},n_2^{(2)},\ldots,n_1^{(d)},n_2^{(d)}} \right\} \quad \text{(induction hypothesis)} \\
&= \left(I_{n_1^{(1)}+n_2^{(1)}} \otimes P_{n_1^{(2)},n_2^{(2)},\ldots,n_1^{(d)},n_2^{(d)}} \right) \left\{ (X_1^{(1)} \oplus X_2^{(1)}) \otimes \left[\bigoplus_{i_2=1}^2 \cdots \bigoplus_{i_d=1}^2 (X_{i_2}^{(2)} \otimes \cdots \otimes X_{i_d}^{(d)}) \right] \right\} \left(I_{n_1^{(1)}+n_2^{(1)}} \otimes P_{n_1^{(2)},n_2^{(2)},\ldots,n_1^{(d)},n_2^{(d)}} \right)^T \\
&= \left(I_{n_1^{(1)}+n_2^{(1)}} \otimes P_{n_1^{(2)},n_2^{(2)},\ldots,n_1^{(d)},n_2^{(d)}} \right) \left\{ \left(X_1^{(1)} \otimes \left[\bigoplus_{i_2=1}^2 \cdots \bigoplus_{i_d=1}^2 (X_{i_2}^{(2)} \otimes \cdots \otimes X_{i_d}^{(d)}) \right] \right) \right. \\
&\qquad \left. \oplus \left(X_2^{(1)} \otimes \left[\bigoplus_{i_2=1}^2 \cdots \bigoplus_{i_d=1}^2 (X_{i_2}^{(2)} \otimes \cdots \otimes X_{i_d}^{(d)}) \right] \right) \right\} \left(I_{n_1^{(1)}+n_2^{(1)}} \otimes P_{n_1^{(2)},n_2^{(2)},\ldots,n_1^{(d)},n_2^{(d)}} \right)^T \quad \text{(Remark 1.1)} \\
&= \left(I_{n_1^{(1)}+n_2^{(1)}} \otimes P_{n_1^{(2)},n_2^{(2)},\ldots,n_1^{(d)},n_2^{(d)}} \right) \left\{ \left(Q_{n_1^{(1)},\boldsymbol{\eta}} \left[\bigoplus_{i_2=1}^2 \cdots \bigoplus_{i_d=1}^2 (X_1^{(1)} \otimes X_{i_2}^{(2)} \otimes \cdots \otimes X_{i_d}^{(d)}) \right] Q^T_{n_1^{(1)},\boldsymbol{\eta}} \right) \right.
\end{aligned}$$

$$\oplus\left(Q_{n_2^{(1)},\boldsymbol{\eta}}\left[\bigoplus_{i_2=1}^{2}\cdots\bigoplus_{i_d=1}^{2}(X_2^{(1)}\otimes X_{i_2}^{(2)}\otimes\cdots\otimes X_{i_d}^{(d)})\right]Q_{n_2^{(1)},\boldsymbol{\eta}}^T\right)\Bigg\}\left(I_{n_1^{(1)}+n_2^{(1)}}\otimes P_{n_1^{(2)},n_2^{(2)},\ldots,n_1^{(d)},n_2^{(d)}}\right)^T$$

(we used Lemma 1.4; $\boldsymbol{\eta}:=(n_{i_2}^{(2)}\cdots n_{i_d}^{(d)})_{(i_2,\ldots,i_d)=(1,\ldots,1),\ldots,(2,\ldots,2)}$ is a multi-index, recall the multi-index notation)

$$=\left(I_{n_1^{(1)}+n_2^{(1)}}\otimes P_{n_1^{(2)},n_2^{(2)},\ldots,n_1^{(d)},n_2^{(d)}}\right)\left(Q_{n_1^{(1)},\boldsymbol{\eta}}\oplus Q_{n_2^{(1)},\boldsymbol{\eta}}\right)\Bigg\{\Bigg[\bigoplus_{i_2=1}^{2}\cdots\bigoplus_{i_d=1}^{2}(X_1^{(1)}\otimes X_{i_2}^{(2)}\otimes\cdots\otimes X_{i_d}^{(d)})\Bigg]$$

$$\oplus\left[\bigoplus_{i_2=1}^{2}\cdots\bigoplus_{i_d=1}^{2}(X_2^{(1)}\otimes X_{i_2}^{(2)}\otimes\cdots\otimes X_{i_d}^{(d)})\right]\Bigg\}\left(Q_{n_1^{(1)},\boldsymbol{\eta}}\oplus Q_{n_2^{(1)},\boldsymbol{\eta}}\right)^T\left(I_{n_1^{(1)}+n_2^{(1)}}\otimes P_{n_1^{(2)},n_2^{(2)},\ldots,n_1^{(d)},n_2^{(d)}}\right)^T$$

$$=P_{n_1^{(1)},n_2^{(1)},n_1^{(2)},n_2^{(2)},\ldots,n_1^{(d)},n_2^{(d)}}\left[\bigoplus_{i_1=1}^{2}\cdots\bigoplus_{i_d=1}^{2}(X_{i_1}^{(1)}\otimes\cdots\otimes X_{i_d}^{(d)})\right]P^T_{n_1^{(1)},n_2^{(1)},n_1^{(2)},n_2^{(2)},\ldots,n_1^{(d)},n_2^{(d)}},$$

where $P_{n_1^{(1)},n_2^{(1)},n_1^{(2)},n_2^{(2)},\ldots,n_1^{(d)},n_2^{(d)}}:=\left(I_{n_1^{(1)}+n_2^{(1)}}\otimes P_{n_1^{(2)},n_2^{(2)},\ldots,n_1^{(d)},n_2^{(d)}}\right)\left(Q_{n_1^{(1)},\boldsymbol{\eta}}\oplus Q_{n_2^{(1)},\boldsymbol{\eta}}\right)$. □

Before concluding this subsection, we stress that a lot of other properties involving tensor products and direct sums can be proved by using techniques similar to those illustrated above. Here we have supplied only the results needed later on.

1.2.2 Hadamard product

The Hadamard product of two matrices X, Y of the same dimensions, say $X, Y \in \mathbb{C}^{m\times\ell}$, is denoted by $X\circ Y$ and is nothing else than the componentwise product of X, Y:

$$(X\circ Y)_{ij}=x_{ij}y_{ij},\qquad i=1,\ldots,m,\ \ j=1,\ldots,\ell.$$

If X, Y are Hermitian, then $X\circ Y$ is Hermitian as well. This property does not hold for the usual matrix product, because, if X, Y are Hermitian, XY may fail to be Hermitian. Moreover, if X, Y are square matrices, then $X\circ Y$ is a principal submatrix of $X\otimes Y$. More precisely, if $X, Y\in\mathbb{C}^{m\times m}$, then $X\circ Y$ is the principal submatrix of $X\otimes Y$ corresponding to the set of indices $F=\{1, m+1, 2m+1,\ldots,(m-1)m+1\}$. From this observation, some important properties of the Hadamard product can be deduced. We collect some of them in Lemma 1.6 for future purposes.

Lemma 1.6. *The Hadamard product possesses the following properties.*

1. $\|X\circ Y\|\le\|X\|\,\|Y\|$ *for all square matrices* $X, Y\in\mathbb{C}^{m\times m}$.
2. *If* X, Y *are HPD (resp. HPSD), then* $X\circ Y$ *is HPD (resp. HPSD).*
3. *If* $X\ge X'\ge O$ *and* $Y\ge Y'\ge O$, *then* $X\circ Y\ge X'\circ Y'$.

Proof. 1. Since $X\circ Y$ is a principal submatrix of $X\otimes Y$, we have $\{\|(X\circ Y)\mathbf{x}\| : \|\mathbf{x}\|=1\}\subseteq\{\|(X\otimes Y)\mathbf{y}\| : \|\mathbf{y}\|=1\}$. Indeed, fixed $\mathbf{x}\in\mathbb{C}^m$ with $\|\mathbf{x}\|=1$, if we take $\mathbf{y}\in\mathbb{C}^{m^2}$ such that $y_i=x_i$ if $i\in F=\{1, m+1, 2m+1,\ldots,(m-1)m+1\}$ and $y_i=0$ otherwise, then $\|\mathbf{y}\|=1$ and $\|(X\otimes Y)\mathbf{y}\|=\|(X\circ Y)\mathbf{x}\|$. Therefore,

$$\|X\circ Y\|=\max_{\|\mathbf{x}\|=1}\|(X\circ Y)\mathbf{x}\|\le\max_{\|\mathbf{y}\|=1}\|(X\otimes Y)\mathbf{y}\|=\|X\otimes Y\|\le\|X\|\|Y\|,$$

where the last inequality holds by (1.13).

2. If X, Y are HPD (HPSD), then $X\otimes Y$ is HPD (HPSD) and $X\circ Y$ is also HPD (HPSD), being a principal submatrix of the HPD (HPSD) matrix $X\otimes Y$.

3. If $X\ge X'\ge O$ and $Y\ge Y'\ge O$, then $X\circ Y-X'\circ Y'=(X-X')\circ Y+X'\circ(Y-Y')\ge O$ by item 2, and so $X\circ Y\ge X'\circ Y'$. □

1.3 Spectral distribution, spectral symbol, clustering

We introduce in this section the fundamental definitions and tools for analyzing the asymptotic spectrum of matrix-sequences. All the notions defined here can be found in [37].

Definition 1.1 (**Spectral distribution of a matrix-sequence**, **spectral symbol**). Let $\{X_n\}$ be a sequence of matrices, with X_n of size d_n tending to infinity, and let $\mathbf{f} : D \to \mathbb{C}^{s\times s}$ be a measurable matrix-valued function, defined on a measurable set $D \subset \mathbb{R}^d$ with $0 < m_d(D) < \infty$.

- We say that $\{X_n\}$ is distributed like $\mathbf{f}$ in the sense of the eigenvalues, in symbols $\{X_n\} \sim_\lambda \mathbf{f}$, if

$$\lim_{n\to\infty} \frac{1}{d_n} \sum_{j=1}^{d_n} F(\lambda_j(X_n)) = \frac{1}{m_d(D)} \int_D \frac{\sum_{i=1}^{s} F\left(\lambda_i(\mathbf{f}(\mathbf{x}))\right)}{s} d\mathbf{x}, \qquad \forall F \in C_c(\mathbb{C}), \tag{1.20}$$

  where $\mathbf{x} = (x_1, \ldots, x_d)$ and $\lambda_i(\mathbf{f}(\mathbf{x}))$, $i = 1, \ldots, s$, are the eigenvalues of $\mathbf{f}(\mathbf{x})$. In this case, $\mathbf{f}$ is referred to as the spectral symbol (or simply the symbol) of the matrix-sequence $\{X_n\}$.

- We say that $\{X_n\}$ is distributed like $\mathbf{f}$ in the sense of the singular values, in symbols $\{X_n\} \sim_\sigma \mathbf{f}$, if

$$\lim_{n\to\infty} \frac{1}{d_n} \sum_{j=1}^{d_n} F(\sigma_j(X_n)) = \frac{1}{m_d(D)} \int_D \frac{\sum_{i=1}^{s} F\left(\sigma_i(\mathbf{f}(\mathbf{x}))\right)}{s} d\mathbf{x}, \qquad \forall F \in C_c(\mathbb{R}), \tag{1.21}$$

  where $\sigma_i(\mathbf{f}(\mathbf{x}))$, $i = 1, \ldots, s$, are the singular values of $\mathbf{f}(\mathbf{x})$.

Note that, in the case $s = 1$, the function $\mathbf{f} : D \to \mathbb{C}$ is scalar-valued (so it will be denoted by f instead of $\mathbf{f}$), and the limit relations (1.20)–(1.21) become

$$\lim_{n\to\infty} \frac{1}{d_n} \sum_{j=1}^{d_n} F(\lambda_j(X_n)) = \frac{1}{m_d(D)} \int_D F(f(\mathbf{x})) d\mathbf{x}, \qquad \forall F \in C_c(\mathbb{C}), \tag{1.22}$$

$$\lim_{n\to\infty} \frac{1}{d_n} \sum_{j=1}^{d_n} F(\sigma_j(X_n)) = \frac{1}{m_d(D)} \int_D F(|f(\mathbf{x})|) d\mathbf{x}, \qquad \forall F \in C_c(\mathbb{R}). \tag{1.23}$$

Remark 1.2. The informal meaning behind (1.20) is the following. Assuming that $\mathbf{f}$ is continuous, if (1.20) holds, then a suitable ordering of the eigenvalues $\{\lambda_j(X_n)\}_{j=1,\ldots,d_n}$, assigned in correspondence of an equispaced grid on D, reconstructs approximately the s hypersurfaces $\mathbf{x} \to \lambda_i(\mathbf{f}(\mathbf{x}))$, $i = 1, \ldots, s$, when n is large. In particular, we may think about the eigenvalues of X_n as if they were split into s different subsets (or 'branches') of the same cardinality, in which the i-th subset is given by a uniform sampling over D of the i-th eigenvalue function $\lambda_i(\mathbf{f}(\mathbf{x}))$. For instance, if $\mathbf{f}$ is continuous, $d = 1$, $d_n = ns$, and $D = [a, b]$, then the eigenvalues of X_n are approximately equal to $\lambda_i(\mathbf{f}(a + j(b-a)/n))$, $j = 1, \ldots, n$, $i = 1, \ldots, s$. Analogously, if $\mathbf{f}$ is continuous, $d = 2$, $d_n = n^2 s$, and $D = [a_1, b_1] \times [a_2, b_2]$, then the eigenvalues of X_n are approximately equal to $\lambda_i(\mathbf{f}(a_1 + j_1(b_1 - a_1)/n, a_2 + j_2(b_2 - a_2)/n))$, $j_1, j_2 = 1, \ldots, n$, $i = 1, \ldots, s$ (and so on in a d-dimensional setting).

For the convenience of the reader, we also report the definition of spectral distribution (and singular value distribution) of a matrix-family $\{X_{\boldsymbol{n}}\}_{\boldsymbol{n}\in\mathbb{N}^d}$ parameterized by a multi-index.

Definition 1.2 (**Spectral distribution of a matrix-family**, **spectral symbol**). Let $\{X_{\boldsymbol{n}}\}_{\boldsymbol{n}\in\mathbb{N}^d}$ be a family of matrices, with $X_{\boldsymbol{n}}$ of size $d_{\boldsymbol{n}}$ tending to infinity as $\boldsymbol{n} \to \infty$, and let $\mathbf{f} : D \to \mathbb{C}^{s\times s}$ be a measurable matrix-valued function, defined on a measurable set $D \subset \mathbb{R}^d$ with $0 < m_d(D) < \infty$.

- We say that $\{X_{\boldsymbol{n}}\}_{\boldsymbol{n}\in\mathbb{N}^d}$ is distributed like $\mathbf{f}$ in the sense of the eigenvalues, in symbols $\{X_{\boldsymbol{n}}\}_{\boldsymbol{n}\in\mathbb{N}^d} \sim_\lambda \mathbf{f}$, if

$$\lim_{\boldsymbol{n}\to\infty} \frac{1}{d_{\boldsymbol{n}}} \sum_{j=1}^{d_{\boldsymbol{n}}} F(\lambda_j(X_{\boldsymbol{n}})) = \frac{1}{m_d(D)} \int_D \frac{\sum_{i=1}^{s} F\left(\lambda_i(\mathbf{f}(\mathbf{x}))\right)}{s} d\mathbf{x}, \qquad \forall F \in C_c(\mathbb{C}). \tag{1.24}$$

  In this case, $\mathbf{f}$ is referred to as the spectral symbol (or the symbol) of the matrix-family $\{X_{\boldsymbol{n}}\}_{\boldsymbol{n}\in\mathbb{N}^d}$. Note that (1.24) says that $\mathbf{f}$ is indeed the symbol, in the sense of Definition 1.1, of any matrix-sequence of the form $\{X_{\boldsymbol{n}(n)}(\mathbf{f})\}_n$, with $\boldsymbol{n}(n) \to \infty$ as $n \to \infty$ (recall from Subsection 1.1.1 that a multi-index tends to infinity when all its components tend to infinity; in particular, $\boldsymbol{n}(n) \to \infty$ means $\min_{j=1,\ldots,d} n_j(n) \to \infty$).

- We say that $\{X_{\boldsymbol{n}}\}_{\boldsymbol{n}\in\mathbb{N}^d}$ is distributed like $\mathbf{f}$ in the sense of the singular values, in symbols $\{X_{\boldsymbol{n}}\}_{\boldsymbol{n}\in\mathbb{N}^d} \sim_\sigma \mathbf{f}$, if $\{X_{\boldsymbol{n}(n)}\}_n \sim_\sigma \mathbf{f}$ in the sense of Definition 1.1 for every sequence of multi-indices $\{\boldsymbol{n}(n)\}_n$ such that $\boldsymbol{n}(n) \to \infty$. Equivalently, $\{X_{\boldsymbol{n}}\}_{\boldsymbol{n}\in\mathbb{N}^d} \sim_\sigma \mathbf{f}$ if

$$\lim_{\boldsymbol{n}\to\infty} \frac{1}{d_{\boldsymbol{n}}} \sum_{j=1}^{d_{\boldsymbol{n}}} F(\sigma_j(X_{\boldsymbol{n}})) = \frac{1}{m_d(D)} \int_D \frac{\sum_{i=1}^{s} F\left(\sigma_i(\mathbf{f}(\mathbf{x}))\right)}{s} d\mathbf{x}, \qquad \forall F \in C_c(\mathbb{R}). \tag{1.25}$$

Now we turn to the definition of clustering. Recall that, according to our notation (see Section 1.1), $D(S,\epsilon)$ denotes the ϵ-expansion of the subset $S \subseteq \mathbb{C}$.

Definition 1.3 (**Clustering of a matrix-sequence at a closed subset of** $\mathbb{C}$). Let $\{X_n\}$ be a sequence of matrices, with X_n of size d_n tending to infinity, and let $S \subseteq \mathbb{C}$ be a nonempty closed subset of $\mathbb{C}$. We say that $\{X_n\}$ is strongly clustered at S in the sense of the eigenvalues if, for every $\epsilon > 0$, the number of eigenvalues of X_n outside $D(S,\epsilon)$ is bounded by a constant C_ϵ independent of n. In other words,

$$q_\epsilon(n,S) := \#\{j \in \{1,\ldots,d_n\} : \lambda_j(X_n) \notin D(S,\epsilon)\} = O(1), \quad \text{as } n \to \infty. \tag{1.26}$$

We say that $\{X_n\}$ is weakly clustered at S in the sense of the eigenvalues if, for every $\epsilon > 0$,

$$q_\epsilon(n,S) = o(d_n), \quad \text{as } n \to \infty.$$

If $\{X_n\}$ is strongly or weakly clustered at S and S is not connected, then the connected components of S are called sub-clusters.

By replacing 'eigenvalues' with 'singular values' and $\lambda_j(X_n)$ with $\sigma_j(X_n)$ in (1.26), we obtain the definitions of a matrix-sequence strongly or weakly clustered at a closed subset of $\mathbb{C}$ in the sense of the singular values.

Throughout the thesis, when we speak of strong/weak cluster, matrix-sequence strongly/weakly clustered, etc., without further specifications, it is always understood 'in the sense of the eigenvalues' (when the clustering is intended in the sense of the singular values, this is specified every time).

It is worth noting that, since the singular values are always nonnegative, any matrix-sequence is strongly clustered in the sense of the singular values at a certain $S \subseteq [0,\infty)$. Similarly, any matrix-sequence formed by matrices with only real eigenvalues (e.g., by Hermitian matrices) is strongly clustered at some $S \subseteq \mathbb{R}$ in the sense of the eigenvalues.

Definition 1.4 (**Spectral attraction**). Let $\{X_n\}$ be a sequence of matrices, with X_n of size d_n tending to infinity, and let $z \in \mathbb{C}$. We say that z strongly attracts the spectrum $\Lambda(X_n)$ with infinite order if, once we have ordered the eigenvalues of X_n according to their distance from z, i.e.

$$|\lambda_1(X_n) - z| \le |\lambda_2(X_n) - z| \le \ldots \le |\lambda_{d_n}(X_n) - z|,$$

the following limit relation holds for each fixed j:

$$\lim_{n\to\infty} |\lambda_j(X_n) - z| = 0.$$

It is now time to introduce the notion of essential range of a matrix-valued function $\mathbf{f}$. For a measurable scalar function $f : D \to \mathbb{C}$, defined on a measurable set $D \subseteq \mathbb{R}^d$, the essential range of f, $\mathcal{ER}(f)$, is defined as the set of points $z \in \mathbb{C}$ such that, for every $\epsilon > 0$, the measure of $\{f \in D(z,\epsilon)\} := \{\mathbf{x} \in D : f(\mathbf{x}) \in D(z,\epsilon)\}$ is positive. In symbols,

$$\mathcal{ER}(f) := \{z \in \mathbb{C} : m_d(\{f \in D(z,\epsilon)\}) > 0, \quad \forall \epsilon > 0\}.$$

Note that $\mathcal{ER}(f)$ is always closed (the complement is open). Moreover, it can be shown that $f(x) \in \mathcal{ER}(f)$ for almost every $\mathbf{x} \in D$, i.e., $f \in \mathcal{ER}(f)$ a.e. In addition, whenever f is continuous and D is sufficiently regular (say, D is contained in the closure of its interior), then $\mathcal{ER}(f)$ coincides with the closure of the image of f.

Definition 1.5 (**Essential range of a matrix-valued function**). Given a measurable matrix-valued function $\mathbf{f} : D \to \mathbb{C}^{s\times s}$, defined on some measurable set $D \subseteq \mathbb{R}^d$, the essential range of $\mathbf{f}$, denoted by $\mathcal{ER}(\mathbf{f})$, is defined as

$$\mathcal{ER}(\mathbf{f}) := \{z \in \mathbb{C} : m_d(\{\exists j : \lambda_j(\mathbf{f}) \in D(z,\epsilon)\}) > 0, \quad \forall \epsilon > 0\},$$

where $\{\exists j : \lambda_j(\mathbf{f}) \in D(z,\epsilon)\} := \{\mathbf{x} \in D : \exists j \in \{1,\dots,s\} \text{ such that } \lambda_j(\mathbf{f}(\mathbf{x})) \in D(z,\epsilon)\}$.

We point out that $\mathcal{ER}(\mathbf{f})$ is well-defined, because the set $\{\exists j : \lambda_j(\mathbf{f}) \in D(z,\epsilon)\}$ is measurable for every $z \in \mathbb{C}$ and $\epsilon > 0$. Moreover, $\mathcal{ER}(\mathbf{f})$ is closed, since its complement is open. Finally, in the case where the eigenvalue functions $\lambda_j(\mathbf{f}) : D \to \mathbb{C}$, $j = 1,\dots,s$, are measurable, we have

$$\mathcal{ER}(f) = \bigcup_{j=1}^{s} \mathcal{ER}(\lambda_j(\mathbf{f})).$$

The following result is stated in [37, Theorem 4.2] and can be proved by using the same arguments shown in the proof of [37, Theorem 2.4].

Theorem 1.5. *Assume that $\{X_n\} \sim_\lambda \mathbf{f}$, with $\{X_n\}$, $\mathbf{f}$ as in Definition 1.1. Then $\{X_n\}$ is weakly clustered at $\mathcal{ER}(\mathbf{f})$ and every point $z \in \mathcal{ER}(\mathbf{f})$ strongly attracts $\Lambda(X_n)$ with infinite order.*

We end this section by providing some useful theorems for proving asymptotic spectral distribution and clustering results. For their proof, see [37, Theorems 3.4 and 3.5]. In Chapeter 2 (Theorem 2.7), we will prove a generalization of Theorem 1.6 to the case where the scalar function f is replaced by a matrix-valued function $\mathbf{f}$.

Theorem 1.6. *Let $\{X_n\}$, $\{Y_n\}$ be sequences of matrices with X_n, $Y_n \in \mathbb{C}^{d_n\times d_n}$ and d_n tending to infinity, and assume the following.*

- *Every X_n is Hermitian and $\{X_n\} \sim_\lambda f$, where $f : D \subset \mathbb{R}^d \to \mathbb{R}$ is a measurable function defined on a measurable set D with $0 < m_d(D) < \infty$;*
- *$\|X_n\|$, $\|Y_n\| \le C$ for all n, with C a constant independent of n;*
- *$|||Y_n|||_1 = o(d_n)$ as $n \to \infty$.*

Then, setting $Z_n := X_n + Y_n$, we have $\{Z_n\} \sim_\lambda f$.

Theorem 1.7. *Let $\{X_n\}$, $\{Y_n\}$ be sequences of matrices with X_n, $Y_n \in \mathbb{C}^{d_n\times d_n}$ and d_n tending to infinity, and assume the following.*

- *Every X_n is Hermitian and $\{X_n\} \sim_\lambda f$, where $f : D \subset \mathbb{R}^d \to \mathbb{R}$ is a measurable function defined on a measurable set D with $0 < m_d(D) < \infty$;*
- *$\|X_n\|$, $|||Y_n|||_1 \le C$ for all n, with C a constant independent of n.*

Then, setting $Z_n := X_n + Y_n$, we have $\{Z_n\} \sim_\lambda f$ and $\{Z_n\}$ is strongly clustered at $\mathcal{ER}(f)$.

1.4 Toeplitz matrices and related topics

In this section, we provide the definition and some properties of multilevel block Toeplitz and circulant matrices. We first focus on multilevel block Toeplitz matrices [11] in Subsection 1.4.1, and then we consider the case of multilevel block circulant matrices [20] in Subsection 1.4.2. Concerning circulant matrices, besides the classical reference book by Davis [20], the reader is referred to [16] for an applicative viewpoint, in particular in connection with the approximation/preconditioning of Toeplitz matrices. We end in Subsection 1.4.3 by reporting some properties of the so-called Generalized Locally Toeplitz (GLT) sequences [68, 63, 64], which will be used in Chapter 5 together with Theorem 1.6 in order to derive an important spectral distribution result.

1.4.1 Multilevel block Toeplitz matrices

Given $\boldsymbol{m} \in \mathbb{N}^d$, a matrix of the form

$$[A_{\boldsymbol{i}-\boldsymbol{j}}]_{\boldsymbol{i},\boldsymbol{j}=\boldsymbol{1}}^{\boldsymbol{m}} \in \mathbb{C}^{N(\boldsymbol{m})s \times N(\boldsymbol{m})s}, \tag{1.27}$$

with blocks $A_{\boldsymbol{k}} \in \mathbb{C}^{s\times s}$, $\boldsymbol{k} = -(\boldsymbol{m}-\boldsymbol{1}), \ldots, \boldsymbol{m}-\boldsymbol{1}$, is called a multilevel block Toeplitz matrix, or, more precisely, a d-level block Toeplitz matrix. Given a function $\mathbf{f} : [-\pi,\pi]^d \to \mathbb{C}^{s\times s}$ in $L^1([-\pi,\pi]^d)$, we denote its Fourier coefficients by

$$\mathbf{f}_{\boldsymbol{k}} = \frac{1}{(2\pi)^d} \int_{[-\pi,\pi]^d} \mathbf{f}(\boldsymbol{\theta}) \mathrm{e}^{-\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{\theta}} d\boldsymbol{\theta} \in \mathbb{C}^{s\times s}, \qquad \boldsymbol{k} \in \mathbb{Z}^d, \tag{1.28}$$

where the integrals are computed componentwise and $\boldsymbol{k}\cdot\boldsymbol{\theta} = k_1\theta_1 + \ldots + k_d\theta_d$. For every $\boldsymbol{m} \in \mathbb{N}^d$, the $\boldsymbol{m}$-th Toeplitz matrix associated with $\mathbf{f}$ is defined as

$$T_{\boldsymbol{m}}(\mathbf{f}) := [\mathbf{f}_{\boldsymbol{i}-\boldsymbol{j}}]_{\boldsymbol{i},\boldsymbol{j}=\boldsymbol{1}}^{\boldsymbol{m}}. \tag{1.29}$$

We call $\{T_{\boldsymbol{m}}(\mathbf{f})\}_{\boldsymbol{m}\in\mathbb{N}^d}$ the family of (multilevel block) Toeplitz matrices associated with $\mathbf{f}$, which, in turn, is called the generating function of $\{T_{\boldsymbol{m}}(\mathbf{f})\}_{\boldsymbol{m}\in\mathbb{N}^d}$.

For each fixed $s \geq 1$ and $\boldsymbol{m} \in \mathbb{N}^d$, the map $T_{\boldsymbol{m}}(\cdot) : L^1([-\pi,\pi]^d, \mathbb{C}^{s\times s}) \to \mathbb{C}^{N(\boldsymbol{m})s\times N(\boldsymbol{m})s}$ is linear: for all $\alpha,\beta \in \mathbb{C}$ and $\mathbf{f},\mathbf{g} \in L^1([-\pi,\pi]^d, \mathbb{C}^{s\times s})$,

$$T_{\boldsymbol{m}}(\alpha\mathbf{f} + \beta\mathbf{g}) = \alpha T_{\boldsymbol{m}}(\mathbf{f}) + \beta T_{\boldsymbol{m}}(\mathbf{g}).$$

This follows from the relation $(\alpha\mathbf{f}+\beta\mathbf{g})_{\boldsymbol{k}} = \alpha\mathbf{f}_{\boldsymbol{k}} + \beta\mathbf{g}_{\boldsymbol{k}}$, $\boldsymbol{k} \in \mathbb{Z}^d$, which is a consequence of the linearity of the integral in (1.28). We now observe that, in general, for every $\mathbf{f} \in L^1([-\pi,\pi]^d, \mathbb{C}^{s\times s})$, the Fourier coefficients of $\mathbf{f}$ are related to those of $\mathbf{f}^*$ by

$$(\mathbf{f}_{\boldsymbol{j}})^* = \left(\frac{1}{(2\pi)^d} \int_{[-\pi,\pi]^d} \mathbf{f}(\boldsymbol{\theta}) \mathrm{e}^{-\mathrm{i}\boldsymbol{j}\cdot\boldsymbol{\theta}} d\boldsymbol{\theta}\right)^* = \frac{1}{(2\pi)^d} \int_{[-\pi,\pi]^d} \mathbf{f}(\boldsymbol{\theta})^* \mathrm{e}^{\mathrm{i}\boldsymbol{j}\cdot\boldsymbol{\theta}} d\boldsymbol{\theta} = (\mathbf{f}^*)_{-\boldsymbol{j}}, \qquad \boldsymbol{j} \in \mathbb{Z}^d.$$

Therefore, for all $\boldsymbol{i},\boldsymbol{j} = \boldsymbol{1}, \ldots, \boldsymbol{m}$,

$$[T_{\boldsymbol{m}}(\mathbf{f}^*)]_{\boldsymbol{i}\boldsymbol{j}} = (\mathbf{f}^*)_{\boldsymbol{i}-\boldsymbol{j}} = (\mathbf{f}_{\boldsymbol{j}-\boldsymbol{i}})^* = [T_{\boldsymbol{m}}(\mathbf{f})^*]_{\boldsymbol{i}\boldsymbol{j}},$$

i.e.,

$$T_{\boldsymbol{m}}(\mathbf{f}^*) = T_{\boldsymbol{m}}(\mathbf{f})^*.$$

From this identity, which holds for all $\boldsymbol{m} \in \mathbb{N}^d$ and all $\mathbf{f} \in L^1([-\pi,\pi]^d, \mathbb{C}^{s\times s})$, we infer that, if $\mathbf{f}$ is a Hermitian matrix-valued function, i.e. $\mathbf{f}(\boldsymbol{\theta})$ is Hermitian for almost every $\boldsymbol{\theta} \in [-\pi,\pi]^d$, then all the matrices $T_{\boldsymbol{m}}(\mathbf{f})$, $\boldsymbol{m} \in \mathbb{N}^d$, are Hermitian.

Theorem 1.8 is a fundamental result concerning multilevel block Toeplitz matrices generated by Hermitian matrix-valued functions. In particular, item 3 in Theorem 1.8 is the Szegö–Tilli theorem; see [11] for a rich account concerning the history of the Szegö theorem, originally appeared in [39]. Item 4 is actually a consequence of item 3, while items 1, 2 can be proved by using the minimax principle.

Theorem 1.8. *Let $\mathbf{f} : [-\pi,\pi]^d \to \mathbb{C}^{s\times s}$ be a Hermitian matrix-valued function in $L^1([-\pi,\pi]^d)$. Define*

$$m_{\mathbf{f}} := \operatorname*{ess\,inf}_{\boldsymbol{\theta}\in[-\pi,\pi]^d} \lambda_{\min}(\mathbf{f}(\boldsymbol{\theta})), \qquad M_{\mathbf{f}} := \operatorname*{ess\,sup}_{\boldsymbol{\theta}\in[-\pi,\pi]^d} \lambda_{\max}(\mathbf{f}(\boldsymbol{\theta})).$$

Then the following properties hold.

1. *$T_{\boldsymbol{m}}(\mathbf{f})$ is Hermitian and $\Lambda(T_{\boldsymbol{m}}(\mathbf{f})) \subseteq [m_{\mathbf{f}}, M_{\mathbf{f}}]$ for all $\boldsymbol{m} \in \mathbb{N}^d$.*
2. *If $\lambda_{\min}(\mathbf{f}(\boldsymbol{\theta}))$ is not a.e. constant then $\Lambda(T_{\boldsymbol{m}}(\mathbf{f})) \subset (m_{\mathbf{f}}, M_{\mathbf{f}}]$ for all $\boldsymbol{m} \in \mathbb{N}^d$.*
   If $\lambda_{\max}(\mathbf{f}(\boldsymbol{\theta}))$ is not a.e. constant then $\Lambda(T_{\boldsymbol{m}}(\mathbf{f})) \subset [m_{\mathbf{f}}, M_{\mathbf{f}})$ for all $\boldsymbol{m} \in \mathbb{N}^d$.
   In particular, if $\mathbf{f} \geq O$ a.e. and $m_d(\{\boldsymbol{\theta} \in [-\pi,\pi]^d : \mathbf{f}(\boldsymbol{\theta}) > O\}) > 0$, then $T_{\boldsymbol{m}}(\mathbf{f}) > O$ for all $\boldsymbol{m} \in \mathbb{N}^d$.
3. *We have $\{T_{\boldsymbol{m}}(\mathbf{f})\}_{\boldsymbol{m}\in\mathbb{N}^d} \sim_\lambda \mathbf{f}$, i.e.*

$$\lim_{\boldsymbol{m}\to\infty} \frac{1}{N(\boldsymbol{m})s} \sum_{k=1}^{N(\boldsymbol{m})s} F(\lambda_k(T_{\boldsymbol{m}}(\mathbf{f}))) = \frac{1}{(2\pi)^d} \int_{[-\pi,\pi]^d} \frac{\sum_{j=1}^s F(\lambda_j(\mathbf{f}(\boldsymbol{\theta})))}{s} d\boldsymbol{\theta}, \qquad \forall F \in C_c(\mathbb{C}). \tag{1.30}$$

   Hence, $\mathbf{f}$ is the symbol of the Toeplitz family $\{T_{\boldsymbol{m}}(\mathbf{f})\}_{\boldsymbol{m}\in\mathbb{N}^d}$.
4. *For each fixed $j \geq 1$, the j-th largest and smallest eigenvalue of $T_{\boldsymbol{m}}(\mathbf{f})$ satisfy*

$$\lambda_j(T_{\boldsymbol{m}}(\mathbf{f})) \to M_{\mathbf{f}}, \qquad \lambda_{N(\boldsymbol{m})s-j+1}(T_{\boldsymbol{m}}(\mathbf{f})) \to m_{\mathbf{f}}$$

   when $\boldsymbol{m} \to \infty$.

Proof. 1. By the minimax principle, since every $T_{\boldsymbol{m}}(\mathbf{f})$ is Hermitian (because $\mathbf{f}$ is a Hermitian matrix-valued function), in order to prove that $\Lambda(T_{\boldsymbol{m}}(\mathbf{f})) \subseteq [m_{\mathbf{f}}, M_{\mathbf{f}}]$ it suffices to show that

$$m_{\mathbf{f}}\|\mathbf{x}\|_2^2 \leq \mathbf{x}^* T_{\boldsymbol{m}}(\mathbf{f})\mathbf{x} \leq M_{\mathbf{f}}\|\mathbf{x}\|_2^2, \qquad \forall \mathbf{x} \in \mathbb{C}^{N(\boldsymbol{m})s}. \tag{1.31}$$

Let $\mathbf{x} \in \mathbb{C}^{N(\boldsymbol{m})s}$ and partition $\mathbf{x}$ as follows: $\mathbf{x} = [\mathbf{x}_i]_{i=1}^{\boldsymbol{m}}$, where each $\mathbf{x}_i \in \mathbb{C}^s$. Then

$$\begin{aligned}
\mathbf{x}^* T_{\boldsymbol{m}}(\mathbf{f})\mathbf{x} = \mathbf{x}^*[\mathbf{f}_{i-j}]_{i,j=1}^{\boldsymbol{m}}\mathbf{x} &= \sum_{i,j=1}^{\boldsymbol{m}} \mathbf{x}_i^* \mathbf{f}_{i-j}\mathbf{x}_j = \sum_{i,j=1}^{\boldsymbol{m}} \frac{1}{(2\pi)^d} \mathbf{x}_i^* \left(\int_{[-\pi,\pi]^d} \mathbf{f}(\boldsymbol{\theta}) \mathrm{e}^{-\mathrm{i}(i-j)\cdot\boldsymbol{\theta}} d\boldsymbol{\theta}\right) \mathbf{x}_j \\
&= \sum_{i,j=1}^{\boldsymbol{m}} \frac{1}{(2\pi)^d} \sum_{k,\ell=1}^{s} \left(\int_{[-\pi,\pi]^d} f_{k\ell}(\boldsymbol{\theta}) \mathrm{e}^{-\mathrm{i}(i-j)\cdot\boldsymbol{\theta}} d\boldsymbol{\theta}\right) (\mathbf{x}_i^*)_k (\mathbf{x}_j)_\ell = \sum_{i,j=1}^{\boldsymbol{m}} \frac{1}{(2\pi)^d} \int_{[-\pi,\pi]^d} \left(\sum_{k,\ell=1}^{s} f_{k\ell}(\boldsymbol{\theta})(\mathbf{x}_i^*)_k (\mathbf{x}_j)_\ell\right) \mathrm{e}^{-\mathrm{i}i\cdot\boldsymbol{\theta}} \mathrm{e}^{\mathrm{i}j\cdot\boldsymbol{\theta}} d\boldsymbol{\theta} \\
&= \sum_{i,j=1}^{\boldsymbol{m}} \frac{1}{(2\pi)^d} \int_{[-\pi,\pi]^d} \mathbf{x}_i^* \mathbf{f}(\boldsymbol{\theta}) \mathbf{x}_j \mathrm{e}^{-\mathrm{i}i\cdot\boldsymbol{\theta}} \mathrm{e}^{\mathrm{i}j\cdot\boldsymbol{\theta}} d\boldsymbol{\theta} = \frac{1}{(2\pi)^d} \int_{[-\pi,\pi]^d} \left(\sum_{i,j=1}^{\boldsymbol{m}} \mathbf{x}_i^* \mathbf{f}(\boldsymbol{\theta}) \mathbf{x}_j \mathrm{e}^{-\mathrm{i}i\cdot\boldsymbol{\theta}} \mathrm{e}^{\mathrm{i}j\cdot\boldsymbol{\theta}}\right) d\boldsymbol{\theta} \\
&= \frac{1}{(2\pi)^d} \int_{[-\pi,\pi]^d} \mathbf{q}_{\mathbf{x}}(\boldsymbol{\theta})^* \mathbf{f}(\boldsymbol{\theta}) \mathbf{q}_{\mathbf{x}}(\boldsymbol{\theta}) d\boldsymbol{\theta},
\end{aligned} \tag{1.32}$$

where $\mathbf{q}_{\mathbf{x}}(\boldsymbol{\theta}) := \sum_{j=1}^{\boldsymbol{m}} \mathbf{x}_j \mathrm{e}^{\mathrm{i}j\cdot\boldsymbol{\theta}}$ satisfies

$$\begin{aligned}
\frac{1}{(2\pi)^d} \int_{[-\pi,\pi]^d} \|\mathbf{q}_{\mathbf{x}}(\boldsymbol{\theta})\|_2^2 d\boldsymbol{\theta} &= \frac{1}{(2\pi)^d} \int_{[-\pi,\pi]^d} \left(\sum_{i=1}^{\boldsymbol{m}} \mathbf{x}_i \mathrm{e}^{\mathrm{i}i\cdot\boldsymbol{\theta}}\right)^* \left(\sum_{j=1}^{\boldsymbol{m}} \mathbf{x}_j \mathrm{e}^{\mathrm{i}j\cdot\boldsymbol{\theta}}\right) d\boldsymbol{\theta} \\
&= \frac{1}{(2\pi)^d} \int_{[-\pi,\pi]^d} \sum_{i,j=1}^{\boldsymbol{m}} \mathbf{x}_i^* \mathbf{x}_j \mathrm{e}^{-\mathrm{i}i\cdot\boldsymbol{\theta}} \mathrm{e}^{\mathrm{i}j\cdot\boldsymbol{\theta}} d\boldsymbol{\theta} = \sum_{i,j=1}^{\boldsymbol{m}} \mathbf{x}_i^* \mathbf{x}_j \frac{1}{(2\pi)^d} \int_{[-\pi,\pi]^d} \mathrm{e}^{-\mathrm{i}i\cdot\boldsymbol{\theta}} \mathrm{e}^{\mathrm{i}j\cdot\boldsymbol{\theta}} d\boldsymbol{\theta} \\
&= \sum_{i,j=1}^{\boldsymbol{m}} \mathbf{x}_i^* \mathbf{x}_j \frac{1}{(2\pi)^d} \left(\mathrm{e}^{\mathrm{i}i\cdot\boldsymbol{\theta}}, \mathrm{e}^{\mathrm{i}j\cdot\boldsymbol{\theta}}\right)_{L^2([-\pi,\pi]^d)} = \sum_{i=1}^{\boldsymbol{m}} \|\mathbf{x}_i\|_2^2 = \|\mathbf{x}\|_2^2.
\end{aligned} \tag{1.33}$$

In the last passages, we have used the fact that $\left(e^{\mathrm{i}\boldsymbol{i}\cdot\boldsymbol{\theta}}, e^{\mathrm{i}\boldsymbol{j}\cdot\boldsymbol{\theta}}\right)_{L^2([-\pi,\pi]^d)}$ equals $(2\pi)^d$ if $\boldsymbol{i} = \boldsymbol{j}$ and 0 otherwise, due to the L^2-orthogonality of the Fourier frequencies $e^{\mathrm{i}\boldsymbol{i}\cdot\boldsymbol{\theta}}$, $\boldsymbol{i} \in \mathbb{Z}^d$. From (1.32) and the minimax principle, we get

$$\frac{1}{(2\pi)^d}\int_{[-\pi,\pi]^d} \lambda_{\min}(\mathbf{f}(\boldsymbol{\theta}))\|\mathbf{q}_{\mathbf{x}}(\boldsymbol{\theta})\|_2^2 d\boldsymbol{\theta} \le \mathbf{x}^* T_{\boldsymbol{m}}(\mathbf{f})\mathbf{x} \le \frac{1}{(2\pi)^d}\int_{[-\pi,\pi]^d} \lambda_{\max}(\mathbf{f}(\boldsymbol{\theta}))\|\mathbf{q}_{\mathbf{x}}(\boldsymbol{\theta})\|_2^2 d\boldsymbol{\theta}, \tag{1.34}$$

and (1.31) follows from (1.33)–(1.34) and from the definitions of $m_{\mathbf{f}}$ and $M_{\mathbf{f}}$.

2. Assume that $\lambda_{\min}(\mathbf{f}(\boldsymbol{\theta}))$ is not a.e. constant and fix $\boldsymbol{m} \in \mathbb{N}^d$. We show that

$$\mathbf{x}^* T_{\boldsymbol{m}}(\mathbf{f})\mathbf{x} > m_{\mathbf{f}}, \qquad \forall \mathbf{x} \in \mathbb{C}^{N(\boldsymbol{m})s} \text{ with } \|\mathbf{x}\|_2 = 1. \tag{1.35}$$

Once we have proved (1.35), from the minimax principle we have

$$\lambda_{\min}(T_{\boldsymbol{m}}(\mathbf{f})) = \min_{\|\mathbf{x}\|_2=1} \mathbf{x}^* T_{\boldsymbol{m}}(\mathbf{f})\mathbf{x} > m_{\mathbf{f}},$$

which, in combination with item 1, yields the first statement in item 2. The second statement is proved in the same way, while the third statement follows from the first one. To prove (1.35), assume by contradiction that there exists a vector $\hat{\mathbf{x}}$ with $\|\hat{\mathbf{x}}\|_2 = 1$, such that

$$\hat{\mathbf{x}}^* T_{\boldsymbol{m}}(\mathbf{f})\hat{\mathbf{x}} = m_{\mathbf{f}}.$$

Note that $\hat{\mathbf{x}}^* T_{\boldsymbol{m}}(\mathbf{f})\hat{\mathbf{x}}$ cannot be less than $m_{\mathbf{f}}$ by item 1. Since $\|\hat{\mathbf{x}}\|_2 = 1$, by (1.33)–(1.34) we have

$$m_{\mathbf{f}} = \frac{1}{(2\pi)^d}\int_{[-\pi,\pi]^d} m_{\mathbf{f}}\|\mathbf{q}_{\hat{\mathbf{x}}}(\boldsymbol{\theta})\|_2^2 d\boldsymbol{\theta} \le \frac{1}{(2\pi)^d}\int_{[-\pi,\pi]^d} \lambda_{\min}(\mathbf{f}(\boldsymbol{\theta}))\|\mathbf{q}_{\hat{\mathbf{x}}}(\boldsymbol{\theta})\|_2^2 d\boldsymbol{\theta} \le \hat{\mathbf{x}}^* T_{\boldsymbol{m}}(\mathbf{f})\hat{\mathbf{x}}.$$

Recalling that $\hat{\mathbf{x}}^* T_{\boldsymbol{m}}(\mathbf{f})\hat{\mathbf{x}} = m_{\mathbf{f}}$, all the previous inequalities are actually equalities and we obtain

$$\int_{[-\pi,\pi]^d} (\lambda_{\min}(\mathbf{f}(\boldsymbol{\theta})) - m_{\mathbf{f}})\|\mathbf{q}_{\hat{\mathbf{x}}}(\boldsymbol{\theta})\|_2^2 d\boldsymbol{\theta} = 0. \tag{1.36}$$

Now, since $\mathbf{q}_{\hat{\mathbf{x}}}(\boldsymbol{\theta})$ is a d-variate trigonometric polynomial, it vanishes at most in a set of zero Lebesgue measure (we omit the details of this proof). Therefore, from (1.36) we deduce that $\lambda_{\min}(\mathbf{f}(\boldsymbol{\theta})) - m_{\mathbf{f}} = 0$ a.e., i.e., $\lambda_{\min}(\mathbf{f}(\boldsymbol{\theta})) = m_{\mathbf{f}}$ a.e. This is a contradiction to the assumption that $\lambda_{\min}(\mathbf{f}(\boldsymbol{\theta}))$ is not a.e. constant.

3. For the proof of item 3, see [67] (see also Subsection 2.1.1, where we provide a proof in the case $d = s = 1$, which can be extended to the general case without significant difficulties).

4. Fixed $j \ge 1$, we prove that $\lambda_j(T_{\boldsymbol{m}}(\mathbf{f})) \to M_{\mathbf{f}}$ as $\boldsymbol{m} \to \infty$ (the proof that $\lambda_{N(\boldsymbol{m})s-j+1} \to m_{\mathbf{f}}$ is similar). Assume by contradiction that $\lambda_j(T_{\boldsymbol{m}}(\mathbf{f}))$ does not converge to $M_{\mathbf{f}}$ as $\boldsymbol{m} \to \infty$. This means that there exists a sequence $\{T_{\boldsymbol{m}(n)}(\mathbf{f})\}_n$ such that $\boldsymbol{m}(n) \to \infty$ and $\lambda_j(T_{\boldsymbol{m}(n)}(\mathbf{f})) \le M < M_{\mathbf{f}}$ and for all n. By definition of $M_{\mathbf{f}}$, we can choose an interval $[\alpha,\beta] \subset (M,\infty)$ such that $m_d(\{\boldsymbol{\theta} \in [-\pi,\pi]^d : \ \alpha \le \lambda_{\max}(\mathbf{f}(\boldsymbol{\theta})) \le \beta\}) > 0$ and a test function $F \in C_c(\mathbb{C})$ such that $0 \le F \le 1$ on $\mathbb{R}$, $F = 0$ on $(-\infty, M]$ and $F = 1$ on $[\alpha,\beta]$. For this test function we have $F(\lambda_i(T_{\boldsymbol{m}(n)}(\mathbf{f}))) = 0$ for $i = j, \dots, N(\boldsymbol{m}(n))s$, and so

$$0 \le \frac{1}{N(\boldsymbol{m}(n))s}\sum_{i=1}^{N(\boldsymbol{m})s} F(\lambda_i(T_{\boldsymbol{m}(n)}(\mathbf{f}))) \le \frac{j-1}{N(\boldsymbol{m}(n))s} \to 0.$$

This is a contradiction to the fact that, by item 3, we have

$$\begin{aligned}\lim_{\boldsymbol{m}\to\infty}\frac{1}{N(\boldsymbol{m})s}\sum_{i=1}^{N(\boldsymbol{m})s}F(\lambda_i(T_{\boldsymbol{m}}(\mathbf{f})))&=\frac{1}{(2\pi)^d}\int_{[-\pi,\pi]^d}\frac{\sum_{i=1}^{s}F(\lambda_i(\mathbf{f}(\boldsymbol{\theta})))}{s}d\boldsymbol{\theta}\\&\geq\frac{1}{(2\pi)^d}\int_{\{\boldsymbol{\theta}\in[-\pi,\pi]^d:\ \alpha\leq\lambda_{\max}(\mathbf{f}(\boldsymbol{\theta}))\leq\beta\}}\frac{F(\lambda_{\max}(\mathbf{f}(\boldsymbol{\theta})))}{s}d\boldsymbol{\theta}\\&=\frac{1}{(2\pi)^d s}\int_{\{\boldsymbol{\theta}\in[-\pi,\pi]^d:\ \alpha\leq\lambda_{\max}(\mathbf{f}(\boldsymbol{\theta}))\leq\beta\}}d\boldsymbol{\theta}\\&=\frac{m_d(\{\boldsymbol{\theta}\in[-\pi,\pi]^d:\ \alpha\leq\lambda_{\max}(\mathbf{f}(\boldsymbol{\theta}))\leq\beta\})}{(2\pi)^d s}>0.\end{aligned}$$

□

From Theorem 1.8 we derive the following proposition, which states that the operator $T_{\boldsymbol{m}}(\cdot)$ is monotone.

Proposition 1.1. *Let $\mathbf{f},\mathbf{g}:[-\pi,\pi]^d\to\mathbb{C}^{s\times s}$ be Hermitian matrix-valued functions in $L^1([-\pi,\pi]^d)$ with $\mathbf{f}(\boldsymbol{\theta})\geq\mathbf{g}(\boldsymbol{\theta})$ a.e. Then $T_{\boldsymbol{m}}(\mathbf{f})\geq T_{\boldsymbol{m}}(\mathbf{g})$ for all $\boldsymbol{m}\in\mathbb{N}^d$.*

Proof. We just observe that, by linearity, $T_{\boldsymbol{m}}(\mathbf{f})\geq T_{\boldsymbol{m}}(\mathbf{g})$ is equivalent to $T_{\boldsymbol{m}}(\mathbf{f}-\mathbf{g})\geq O$. The latter is satisfied by Theorem 1.8, since $\mathbf{f}(\boldsymbol{\theta})-\mathbf{g}(\boldsymbol{\theta})\geq O$ a.e. by hypothesis and hence $m_{\mathbf{f}-\mathbf{g}}\geq 0$. □

Important inequalities involving Toeplitz matrices and Schatten p-norms can be found in [61, Corollary 3.5]. In Lemma 1.7, we report one of these inequalities of interest later on.

Lemma 1.7. *For $\mathbf{f}\in L^\infty([-\pi,\pi]^d,\mathbb{C}^{s\times s})$ and $\boldsymbol{m}\in\mathbb{N}^d$,*

$$\|T_{\boldsymbol{m}}(\mathbf{f})\|\leq\operatorname*{ess\,sup}_{\boldsymbol{\theta}\in[-\pi,\pi]^d}\|\mathbf{f}(\boldsymbol{\theta})\|. \tag{1.37}$$

Proof. From (1.4) we know that

$$\|T_{\boldsymbol{m}}(\mathbf{f})\|=\max_{\|\mathbf{u}\|=\|\mathbf{v}\|=1}\mathbf{u}^*T_{\boldsymbol{m}}(\mathbf{f})\mathbf{v}.$$

By performing some computations completely analogous to the ones in the proof of Theorem 1.8, see in particular the chain of equalities (1.32)–(1.33), we see that

$$\mathbf{u}^*T_{\boldsymbol{m}}(\mathbf{f})\mathbf{v}=\frac{1}{(2\pi)^d}\int_{[-\pi,\pi]^d}\mathbf{q}_{\mathbf{u}}(\boldsymbol{\theta})^*\mathbf{f}(\boldsymbol{\theta})\mathbf{q}_{\mathbf{v}}(\boldsymbol{\theta})d\boldsymbol{\theta},$$

where

$$\frac{1}{(2\pi)^d}\int_{[-\pi,\pi]^d}\|\mathbf{q}_{\mathbf{u}}(\boldsymbol{\theta})\|^2d\boldsymbol{\theta}=\|\mathbf{u}\|^2=1,\qquad\frac{1}{(2\pi)^d}\int_{[-\pi,\pi]^d}\|\mathbf{q}_{\mathbf{v}}(\boldsymbol{\theta})\|^2d\boldsymbol{\theta}=\|\mathbf{v}\|^2=1.$$

Using the Cauchy-Schwarz inequality, we obtain

$$\begin{aligned}|\mathbf{u}^*T_{\boldsymbol{m}}(\mathbf{f})\mathbf{v}|&\leq\frac{1}{(2\pi)^d}\int_{[-\pi,\pi]^d}|\mathbf{q}_{\mathbf{u}}(\boldsymbol{\theta})^*\mathbf{f}(\boldsymbol{\theta})\mathbf{q}_{\mathbf{v}}(\boldsymbol{\theta})|\,d\boldsymbol{\theta}\leq\frac{1}{(2\pi)^d}\int_{[-\pi,\pi]^d}\|\mathbf{q}_{\mathbf{u}}(\boldsymbol{\theta})\|\,\|\mathbf{f}(\boldsymbol{\theta})\|\,\|\mathbf{q}_{\mathbf{v}}(\boldsymbol{\theta})\|d\boldsymbol{\theta}\\&\leq\operatorname*{ess\,sup}_{\boldsymbol{\theta}\in[-\pi,\pi]^d}\|\mathbf{f}(\boldsymbol{\theta})\|\,\frac{1}{(2\pi)^d}\int_{[-\pi,\pi]^d}\|\mathbf{q}_{\mathbf{u}}(\boldsymbol{\theta})\|\,\|\mathbf{q}_{\mathbf{v}}(\boldsymbol{\theta})\|d\boldsymbol{\theta}\\&\leq\operatorname*{ess\,sup}_{\boldsymbol{\theta}\in[-\pi,\pi]^d}\|\mathbf{f}(\boldsymbol{\theta})\|\,\frac{1}{(2\pi)^d}\left(\int_{[-\pi,\pi]^d}\|\mathbf{q}_{\mathbf{u}}(\boldsymbol{\theta})\|^2d\boldsymbol{\theta}\right)^{1/2}\left(\int_{[-\pi,\pi]^d}\|\mathbf{q}_{\mathbf{v}}(\boldsymbol{\theta})\|^2d\boldsymbol{\theta}\right)^{1/2}\\&=\operatorname*{ess\,sup}_{\boldsymbol{\theta}\in[-\pi,\pi]^d}\|\mathbf{f}(\boldsymbol{\theta})\|,\end{aligned}$$

and the thesis follows. □

Note that (1.37) can be reformulated in terms of the Schatten ∞-norm as follows:

$$\|\|T_{\boldsymbol{m}}(\mathbf{f})\|\|_\infty \le \| \, \|\|\mathbf{f}(\boldsymbol{\theta})\|\|_\infty \, \|_{L^\infty([-\pi,\pi]^d)}.$$

From this reformulation, we see that (1.37) coincides with the inequality (28) in [61] for $p = \infty$.

Theorem 1.9. *Let*

$$X = \begin{bmatrix} b & c & & \\ a & \ddots & \ddots & \\ & \ddots & \ddots & c \\ & & a & b \end{bmatrix} = \text{Tridiagonal}(a, b, c)$$

be an $m \times m$ real Toeplitz tridiagonal matrix such that $ac > 0$. Then, X has m real distinct eigenvalues

$$\lambda_j(X) = b + 2\sqrt{ac}\,\cos\frac{j\pi}{m+1}, \quad j = 1, \ldots, m.$$

Proof. See [9, p. 35] or [66, p. 154]. □

The next result concerns the exact asymptotics of the j-th smallest eigenvalue of $T_m(f)$, for j fixed and $m \to \infty$. This result is due to Parter [45] (see also [46] for a generalization). It shows that, under the assumption that f is continuous and $f - \min f$ has a unique zero $\theta_{\min}$, $\lambda_{m-j+1}(T_m(f))$ converges to $m_f = \min f$ as $m \to \infty$ with asymptotic speed dictated by the order of the zero $\theta_{\min}$.

Theorem 1.10 (Parter). *Let $f : \mathbb{R} \to \mathbb{R}$ be continuous and 2π-periodic. Let $m_f := \min_{\theta\in\mathbb{R}} f(\theta) = f(\theta_{\min})$ and let $\theta_{\min}$ be the unique zero of $f - m_f$ in $(-\pi, \pi]$. Assume there exists $s \ge 1$ such that f has $2s$ continuous derivatives in $(\theta_{\min} - \epsilon, \theta_{\min} + \epsilon)$ for some $\epsilon > 0$ and $f^{(2s)}(\theta_{\min}) > 0$ is the first non-vanishing derivative of f at $\theta_{\min}$. Then, for each fixed $j \ge 1$,*

$$\lambda_{m-j+1}(T_m(f)) - m_f \sim c_{s,j}\,\frac{f^{(2s)}(\theta_{\min})}{(2s)!}\,\frac{1}{m^{2s}}, \qquad \text{as} \quad m \to \infty, \tag{1.38}$$

i.e., $\lim_{m\to\infty} m^{2s}\left(\lambda_{m-j+1}(T_m(f)) - m_f\right) = c_{s,j}\,\dfrac{f^{(2s)}(\theta_{\min})}{(2s)!}$, *where $c_{s,j} > 0$ is a constant depending only on s and j.*

Remark 1.3. The constant $c_{s,j}$ is the j-th smallest eigenvalue of the boundary value problem

$$\begin{cases} (-1)^s u^{(2s)}(x) = f(x), \quad \text{for} \quad 0 < x < 1, \\ u(0) = u'(0) = \ldots = u^{(s-1)}(0) = 0, \quad u(1) = u'(1) = \ldots = u^{(s-1)}(1) = 0; \end{cases} \tag{1.39}$$

see [45, p. 191]. This means that $c_{s,j}$ is the j-th smallest number satisfying $(-1)^s u^{(2s)}(x) = c_{s,j}\,u(x)$ for some (nonzero) function u belonging to an 'appropriate functional space' associated with (1.39). In particular, $c_{s,1}$ is the minimum eigenvalue of (1.39). The sequence $\{c_{s,1}\}$ was investigated in [12], where it was shown that the numbers $c_{1,1}, c_{2,1}, c_{3,1}, \ldots$ appear in many situations and the following asymptotic formula holds:

$$c_{s,1} = \sqrt{8\pi s}\left(\frac{4s}{\mathrm{e}}\right)^{2s}\left[1 + O\left(\frac{1}{\sqrt{s}}\right)\right], \qquad \text{as} \quad s \to \infty.$$

Remark 1.4. When $s = 1$, the boundary value problem (1.39) becomes

$$\begin{cases} -u''(x) = f(x), \qquad 0 < x < 1, \\ u(0) = u(1) = 0, \end{cases} \tag{1.40}$$

and its eigenvalues can be computed explicitly, because they coincide with the eigenvalues of the unidimensional negative Laplacian operator $-\frac{d^2}{dx^2}$ with homogeneous Dirichlet boundary conditions:

$$-\frac{d^2}{dx^2} : H_0^2(0,1) \subset L^2(0,1) \to L^2(0,1). \tag{1.41}$$

The mentioned 'appropriate functional space' is in this case $H_0^2(0,1)$. The eigenvalues of (1.41) are $j^2\pi^2$, $j = 1,2,\ldots$, and an eigenfunction corresponding to the j-th eigenvalue $j^2\pi^2$ is $u_j(x) = \sin(j\pi x)$: $-u_j''(x) = j^2\pi^2 u_j(x)$. Thus, by Remark 1.3, we find that $c_{1,j} = j^2\pi^2$ for all $j \geq 1$.

Remark 1.5. Parter's theorem applies to the function $f(\theta) = (2 - 2\cos\theta)^s$, $s \geq 1$. Indeed, it can be proved that this function satisfies all the hypotheses of Theorem 1.10 with $m_f = 0$, $\theta_{\min} = 0$, and the number s appearing in Theorem 1.10 being exactly the exponent s in the definition of f. Moreover, $f^{(2s)}(\theta_{\min}) = (2s)!$ Therefore, by (1.38) we obtain that, for each fixed $j \geq 1$,

$$\lambda_{m-j+1}(T_m((2-2\cos\theta)^s)) \sim \frac{c_{s,j}}{m^{2s}}, \qquad \text{as } m \to \infty.$$

On the other hand, for the case $s = 1$, noting that $T_m(2 - 2\cos\theta) = \text{Tridiagonal}(-1, 2, -1)$ and using Theorem 1.9, we get

$$\lambda_{m-j+1}(T_m(2-2\cos\theta)) = 4\left(\sin\frac{j\pi}{2(m+1)}\right)^2 \sim \frac{j^2\pi^2}{m^2}, \qquad \text{as } m \to \infty,$$

and so we find again $c_{1,j} = j^2\pi^2$ for all $j \geq 1$.

The last results relate tensor products and Toeplitz matrices. In particular, in Lemma 1.9 we show that a tensor product of unilevel block Toeplitz matrices generated by (matrix-valued) trigonometric polynomials coincides (modulo permutation transformations) with the multilevel block Toeplitz matrix generated by the tensor product of the trigonometric polynomials.

Lemma 1.8. *Let $f_1, \ldots, f_d \in L^1([-\pi,\pi])$ and let $\boldsymbol{m} = (m_1, \ldots, m_d) \in \mathbb{N}^d$. Then,*

$$T_{m_1}(f_1) \otimes \cdots \otimes T_{m_d}(f_d) = T_{\boldsymbol{m}}(f_1 \otimes \cdots \otimes f_d) \tag{1.42}$$

(note that the tensor-product function $f_1 \otimes \cdots \otimes f_d : [-\pi,\pi]^d \to \mathbb{C}$ belongs to $L^1([-\pi,\pi]^d)$ by Fubini's theorem).

Proof. The proof is simple if we use the fundamental property (1.12). Noting that the Fourier coefficients of $f_1 \otimes \cdots \otimes f_d$ are given by

$$(f_1 \otimes \cdots \otimes f_d)_{\boldsymbol{k}} = (f_1)_{k_1} \cdots (f_d)_{k_d}, \qquad \boldsymbol{k} \in \mathbb{Z}^d,$$

for all $\boldsymbol{i}, \boldsymbol{j} = \boldsymbol{1}, \ldots, \boldsymbol{m}$ we have

$$\begin{aligned}[T_{m_1}(f_1) \otimes \cdots \otimes T_{m_d}(f_d)]_{\boldsymbol{ij}} &= [T_{m_1}(f_1)]_{i_1 j_1} \cdots [T_{m_d}(f_d)]_{i_d j_d} = (f_1)_{i_1-j_1} \cdots (f_d)_{i_d-j_d} = (f_1 \otimes \cdots \otimes f_d)_{\boldsymbol{i}-\boldsymbol{j}} \\ &= [T_{\boldsymbol{m}}(f_1 \otimes \cdots \otimes f_d)]_{\boldsymbol{ij}},\end{aligned}$$

and so (1.42) holds. □

A matrix-valued function of the form $\mathbf{p}(\theta) = \sum_{k=-N}^{N} A_k e^{\mathrm{i}k\theta}$, with $A_k \in \mathbb{C}^{s\times s}$, $k = -N, \ldots, N$, is called (matrix-valued) trigonometric polynomial.

Lemma 1.9. *For every $\boldsymbol{m}, \boldsymbol{s} \in \mathbb{N}^d$ there exists a permutation matrix $\Gamma_{\boldsymbol{m},\boldsymbol{s}}$ of size $\prod_{j=1}^{d}(m_j s_j)$ such that*

$$T_{m_1}(\mathbf{p}_1) \otimes \cdots \otimes T_{m_d}(\mathbf{p}_d) = \Gamma_{\boldsymbol{m},\boldsymbol{s}}\,[T_{\boldsymbol{m}}(\mathbf{p}_1 \otimes \cdots \otimes \mathbf{p}_d)]\,\Gamma_{\boldsymbol{m},\boldsymbol{s}}^T,$$

for any choice of trigonometric polynomials $\mathbf{p}_j : [-\pi,\pi] \to \mathbb{C}^{s_j \times s_j}$, $j = 1, \ldots, d$.

Proof. For $\boldsymbol{k} \in \mathbb{Z}^d$ and $A \in \mathbb{C}^{s\times s}$, it can be shown by direct computation that

$$T_{\boldsymbol{m}}(A\mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{\theta}}) = T_{\boldsymbol{m}}(\mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{\theta}}) \otimes A = T_{m_1}(\mathrm{e}^{\mathrm{i}k_1\theta_1}) \otimes \cdots \otimes T_{m_d}(\mathrm{e}^{\mathrm{i}k_d\theta_d}) \otimes A.$$

Therefore, for any choice of the trigonometric polynomials

$$\mathbf{p}_j(\theta) := \sum_{k=-N_j}^{N_j} A_k^{(j)}\mathrm{e}^{\mathrm{i}k\theta}, \quad j = 1,\ldots,d \qquad (A_k^{(j)} \in \mathbb{C}^{s_j\times s_j}, \quad j = 1,\ldots,d, \quad k = -N_j,\ldots,N_j),$$

by the bilinearity of $\otimes$ and the linearity of $T_{\boldsymbol{m}}(\cdot)$, we have

$$\begin{aligned} T_{\boldsymbol{m}}(\mathbf{p}_1(\theta_1) \otimes \cdots \otimes \mathbf{p}_d(\theta_d)) &= T_{\boldsymbol{m}}\left(\sum_{k_1=-N_1}^{N_1} \cdots \sum_{k_d=-N_d}^{N_d} A_{k_1}^{(1)} \otimes \cdots \otimes A_{k_d}^{(d)}\mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{\theta}}\right) = \sum_{k_1=-N_1}^{N_1} \cdots \sum_{k_d=-N_d}^{N_d} T_{\boldsymbol{m}}(A_{k_1}^{(1)} \otimes \cdots \otimes A_{k_d}^{(d)}\mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{\theta}}) \\ &= \sum_{\boldsymbol{k}=-\boldsymbol{N}}^{\boldsymbol{N}} T_{m_1}(\mathrm{e}^{\mathrm{i}k_1\theta_1}) \otimes \cdots \otimes T_{m_d}(\mathrm{e}^{\mathrm{i}k_d\theta_d}) \otimes A_{k_1}^{(1)} \otimes \cdots \otimes A_{k_d}^{(d)}. \end{aligned}$$

On the other hand,

$$\begin{aligned} T_{m_1}(\mathbf{p}_1(\theta_1)) \otimes \cdots \otimes T_{m_d}(\mathbf{p}_d(\theta_d)) &= T_{m_1}\left(\sum_{k_1=-N_1}^{N_1} A_{k_1}^{(1)}\mathrm{e}^{\mathrm{i}k_1\theta_1}\right) \otimes \cdots \otimes T_{m_d}\left(\sum_{k_d=-N_d}^{N_d} A_{k_d}^{(d)}\mathrm{e}^{\mathrm{i}k_d\theta_d}\right) \\ &= \left(\sum_{k_1=-N_1}^{N_1} T_{m_1}(\mathrm{e}^{\mathrm{i}k_1\theta_1}) \otimes A_{k_1}^{(1)}\right) \otimes \cdots \otimes \left(\sum_{k_d=-N_d}^{N_d} T_{m_d}(\mathrm{e}^{\mathrm{i}k_d\theta_d}) \otimes A_{k_d}^{(d)}\right) \\ &= \sum_{\boldsymbol{k}=-\boldsymbol{N}}^{\boldsymbol{N}} T_{m_1}(\mathrm{e}^{\mathrm{i}k_1\theta_1}) \otimes A_{k_1}^{(1)} \otimes \cdots \otimes T_{m_d}(\mathrm{e}^{\mathrm{i}k_d\theta_d}) \otimes A_{k_d}^{(d)}. \end{aligned}$$

By Lemma 1.2, there exists the permutation matrix $\Gamma_{\boldsymbol{m},\boldsymbol{s}} := \Pi_{(\boldsymbol{m},\boldsymbol{s});\sigma}$, where $\sigma := [1, d+1, 2, d+2, \ldots, d, 2d]$, which depends only on $\boldsymbol{m}, \boldsymbol{s}$ and satisfies

$$T_{m_1}(\mathrm{e}^{\mathrm{i}k_1\theta_1}) \otimes A_{k_1}^{(1)} \otimes \cdots \otimes T_{m_d}(\mathrm{e}^{\mathrm{i}k_d\theta_d}) \otimes A_{k_d}^{(d)} = \Gamma_{\boldsymbol{m},\boldsymbol{s}}\left[T_{m_1}(\mathrm{e}^{\mathrm{i}k_1\theta_1}) \otimes \cdots \otimes T_{m_d}(\mathrm{e}^{\mathrm{i}k_d\theta_d}) \otimes A_{k_1}^{(1)} \otimes \cdots \otimes A_{k_d}^{(d)}\right]\Gamma_{\boldsymbol{m},\boldsymbol{s}}^T.$$

Hence,

$$\begin{aligned} T_{m_1}(\mathbf{p}_1(\theta_1)) \otimes \cdots \otimes T_{m_d}(\mathbf{p}_d(\theta_d)) &= \sum_{\boldsymbol{k}=-\boldsymbol{N}}^{\boldsymbol{N}} \Gamma_{\boldsymbol{m},\boldsymbol{s}}\left[T_{m_1}(\mathrm{e}^{\mathrm{i}k_1\theta_1}) \otimes \cdots \otimes T_{m_d}(\mathrm{e}^{\mathrm{i}k_d\theta_d}) \otimes A_{k_1}^{(1)} \otimes \cdots \otimes A_{k_d}^{(d)}\right]\Gamma_{\boldsymbol{m},\boldsymbol{s}}^T \\ &= \Gamma_{\boldsymbol{m},\boldsymbol{s}}\left[\sum_{\boldsymbol{k}=-\boldsymbol{N}}^{\boldsymbol{N}} T_{m_1}(\mathrm{e}^{\mathrm{i}k_1\theta_1}) \otimes \cdots \otimes T_{m_d}(\mathrm{e}^{\mathrm{i}k_d\theta_d}) \otimes A_{k_1}^{(1)} \otimes \cdots \otimes A_{k_d}^{(d)}\right]\Gamma_{\boldsymbol{m},\boldsymbol{s}}^T \\ &= \Gamma_{\boldsymbol{m},\boldsymbol{s}}T_{\boldsymbol{m}}(\mathbf{p}_1(\theta_1) \otimes \cdots \otimes \mathbf{p}_d(\theta_d))\Gamma_{\boldsymbol{m},\boldsymbol{s}}^T. \end{aligned}$$

□

Lemma 1.9 shows that the operators $T_{\boldsymbol{m}}(\cdot)$ and $\otimes$ are interchangeable, modulo permutation transformations which only depend on the dimensions of the involved matrices. The same result is true also for the operators $T_{\boldsymbol{m}}(\cdot)$ and $\oplus$; see [29, Theorem 2].

1.4.2 Multilevel block circulant matrices

Given $\boldsymbol{m} \in \mathbb{N}^d$, a matrix of the form

$$[A_{(\boldsymbol{i}-\boldsymbol{j}) \bmod \boldsymbol{m}}]_{\boldsymbol{i},\boldsymbol{j}=\boldsymbol{1}}^{\boldsymbol{m}} \in \mathbb{C}^{N(\boldsymbol{m})s \times N(\boldsymbol{m})s}, \tag{1.43}$$

with blocks $A_{\boldsymbol{k}} \in \mathbb{C}^{s\times s}$, $\boldsymbol{k} = \boldsymbol{0}, \dots, \boldsymbol{m}-\boldsymbol{1}$, is called a multilevel block circulant matrix, or, more precisely, a d-level block circulant matrix. The fundamental theorem concerning multilevel block circulant matrices is the following. For $\boldsymbol{m} \in \mathbb{N}^d$ we denote by $F_{\boldsymbol{m}}$ the unitary d-level Fourier transform, i.e. $F_{\boldsymbol{m}} := F_{m_1} \otimes \cdots \otimes F_{m_d}$, where $F_m := \frac{1}{\sqrt{m}}(\mathrm{e}^{-2\pi \mathrm{i} jk/m})_{j,k=0}^{m-1} = \frac{1}{\sqrt{m}}(\mathrm{e}^{-2\pi \mathrm{i}(j-1)(k-1)/m})_{j,k=1}^{m}$ is the standard unitary Fourier transform of order m ($F_m^* F_m = I_m$).

Theorem 1.11. *The matrix* (1.43) *has the following block spectral decomposition:*

$$[A_{(\boldsymbol{i}-\boldsymbol{j}) \bmod \boldsymbol{m}}]_{\boldsymbol{i},\boldsymbol{j}=\boldsymbol{1}}^{\boldsymbol{m}} = (F_{\boldsymbol{m}} \otimes I_s) \operatorname*{diag}_{\boldsymbol{j}=\boldsymbol{0},\dots,\boldsymbol{m}-\boldsymbol{1}} \left[\mathbf{g}\left(\frac{2\pi \boldsymbol{j}}{\boldsymbol{m}}\right)\right] (F_{\boldsymbol{m}} \otimes I_s)^*, \tag{1.44}$$

where $\mathbf{g}(\boldsymbol{\theta}) := \sum_{\boldsymbol{k}=\boldsymbol{0}}^{\boldsymbol{m}-\boldsymbol{1}} A_{\boldsymbol{k}} \mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{\theta}}$. *In particular, the spectrum of* $[A_{(\boldsymbol{i}-\boldsymbol{j}) \bmod \boldsymbol{m}}]_{\boldsymbol{i},\boldsymbol{j}=\boldsymbol{1}}^{\boldsymbol{m}}$ *is given by the union of the spectra of the diagonal blocks* $\mathbf{g}(2\pi \boldsymbol{j}/\boldsymbol{m}) \in \mathbb{C}^{s\times s}$, $\boldsymbol{j} = \boldsymbol{0}, \dots, \boldsymbol{m}-\boldsymbol{1}$.

Proof. The proof of this theorem is a good exercise on the multi-index notation. It consists of four steps.

1. Consider the $m \times m$ matrix

$$Z_m := \begin{bmatrix} 0 & & & & 1 \\ 1 & \ddots & & & \\ & \ddots & \ddots & & \\ & & \ddots & \ddots & \\ & & & 1 & 0 \end{bmatrix} = [\delta_{(i-j-1) \bmod m}]_{i,j=1}^{m},$$

where $\delta_r := 1$ if $r = 0$ and $\delta_r := 0$ otherwise. The matrix Z_m is called the generator of unilevel circulant matrices of order m. This name is due to the fact that the powers of Z_m are

$$Z_m^2 = \begin{bmatrix} 0 & & & & 1 & 0 \\ 0 & \ddots & & & & 1 \\ 1 & \ddots & \ddots & & & \\ & \ddots & \ddots & \ddots & & \\ & & \ddots & \ddots & \ddots & \\ & & & 1 & 0 & 0 \end{bmatrix}, \quad Z_m^3 = \begin{bmatrix} 0 & & & 1 & 0 & 0 \\ 0 & \ddots & & & 1 & 0 \\ 0 & \ddots & \ddots & & & 1 \\ 1 & \ddots & \ddots & \ddots & & \\ & \ddots & \ddots & \ddots & \ddots & \\ & & 1 & 0 & 0 & 0 \end{bmatrix}, \quad \dots, \quad Z_m^m = I_m,$$

or, in formulas,

$$(Z_m^k)_{ij} = \delta_{(i-j-k) \bmod m}, \quad i, j = 1, \dots, m, \quad k = 0, \dots, m-1. \tag{1.45}$$

Therefore, any unilevel circulant matrix of order m can be written as a linear combination of powers of Z_m:

$$[a_{(i-j) \bmod m}]_{i,j=1}^{m} = \begin{bmatrix} a_0 & a_{m-1} & & & a_2 & a_1 \\ a_1 & \ddots & \ddots & & & a_2 \\ a_2 & \ddots & \ddots & \ddots & & \\ & \ddots & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & \ddots & a_{m-1} \\ a_{m-1} & & & a_2 & a_1 & a_0 \end{bmatrix} = \sum_{k=0}^{m-1} a_k Z_m^k.$$

Note also that any unilevel block circulant matrix $[A_{(i-j) \bmod m}]_{i,j=1}^{m}$ can be written as

$$[A_{(i-j) \bmod m}]_{i,j=1}^{m} = \begin{bmatrix} A_0 & A_{m-1} & & & A_2 & A_1 \\ A_1 & \ddots & \ddots & & & A_2 \\ A_2 & \ddots & \ddots & \ddots & & \\ & \ddots & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & \ddots & A_{m-1} \\ A_{m-1} & & & A_2 & A_1 & A_0 \end{bmatrix} = \sum_{k=0}^{m-1} Z_m^k \otimes A_k.$$

2. The spectral decomposition of Z_m is explicitly known and is given by

$$Z_m = F_m D_m F_m^*, \qquad D_m := \operatorname*{diag}_{j=0,\ldots,m-1} (\mathrm{e}^{2\pi \mathrm{i} j/m}) = \operatorname*{diag}_{j=1,\ldots,m} (\mathrm{e}^{2\pi \mathrm{i}(j-1)/m}).$$

This can be verified by direct computation: for all $i, j = 1, \ldots, m$, we have

$$(F_m^* Z_m)_{ij} = \sum_{k=1}^{m} \mathrm{e}^{2\pi \mathrm{i}(i-1)(k-1)/m} \delta_{(k-j-1) \bmod m} = \mathrm{e}^{2\pi \mathrm{i}(i-1)j/m} = (D_m F_m^*)_{ij}.$$

Therefore, defining $Z_{\boldsymbol{m}}^{\boldsymbol{k}} := Z_{m_1}^{k_1} \otimes Z_{m_2}^{k_2} \otimes \cdots \otimes Z_{m_d}^{k_d}$, $\boldsymbol{k}, \boldsymbol{m} \in \mathbb{N}^d$, also the spectral decomposition of $Z_{\boldsymbol{m}}^{\boldsymbol{k}}$ is known. Indeed, using the properties of tensor products in Subsection 1.2.1, we obtain

$$\begin{aligned} Z_{\boldsymbol{m}}^{\boldsymbol{k}} &= Z_{m_1}^{k_1} \otimes Z_{m_2}^{k_2} \otimes \cdots \otimes Z_{m_d}^{k_d} = (F_{m_1} D_{m_1}^{k_1} F_{m_1}^*) \otimes (F_{m_2} D_{m_2}^{k_2} F_{m_2}^*) \otimes \cdots \otimes (F_{m_d} D_{m_d}^{k_d} F_{m_d}^*) \\ &= (F_{m_1} \otimes F_{m_2} \otimes \cdots \otimes F_{m_d})(D_{m_1}^{k_1} \otimes D_{m_2}^{k_2} \otimes \cdots \otimes D_{m_d}^{k_d})(F_{m_1} \otimes F_{m_2} \otimes \cdots \otimes F_{m_d})^* = F_{\boldsymbol{m}} D_{\boldsymbol{m}}^{\boldsymbol{k}} F_{\boldsymbol{m}}^*, \end{aligned} \tag{1.46}$$

where

$$D_{\boldsymbol{m}}^{\boldsymbol{k}} := D_{m_1}^{k_1} \otimes D_{m_2}^{k_2} \otimes \cdots \otimes D_{m_d}^{k_d} = \operatorname*{diag}_{\boldsymbol{j}=\boldsymbol{0},\ldots,\boldsymbol{m}-\boldsymbol{1}} (\mathrm{e}^{2\pi \mathrm{i} \sum_{r=1}^{d} j_r k_r / m_r}) = \operatorname*{diag}_{\boldsymbol{j}=\boldsymbol{0},\ldots,\boldsymbol{m}-\boldsymbol{1}} (\mathrm{e}^{2\pi \mathrm{i}(\boldsymbol{j}/\boldsymbol{m}) \cdot \boldsymbol{k}}).$$

3. The multilevel block circulant matrix $[A_{(\boldsymbol{i}-\boldsymbol{j}) \bmod \boldsymbol{m}}]_{\boldsymbol{i},\boldsymbol{j}=\boldsymbol{1}}^{\boldsymbol{m}}$ in (1.43) has the following expression:

$$[A_{(\boldsymbol{i}-\boldsymbol{j}) \bmod \boldsymbol{m}}]_{\boldsymbol{i},\boldsymbol{j}=\boldsymbol{1}}^{\boldsymbol{m}} = \sum_{\boldsymbol{k}=\boldsymbol{0}}^{\boldsymbol{m}-\boldsymbol{1}} Z_{\boldsymbol{m}}^{\boldsymbol{k}} \otimes A_{\boldsymbol{k}}. \tag{1.47}$$

To prove (1.47), we first notice that, by the fundamental property (1.12) and by (1.45), for all $\boldsymbol{i}, \boldsymbol{j} = \boldsymbol{1}, \ldots, \boldsymbol{m}$ we have

$$(Z_{\boldsymbol{m}}^{\boldsymbol{k}})_{\boldsymbol{ij}} = (Z_{m_1}^{k_1})_{i_1 j_1} (Z_{m_2}^{k_2})_{i_2 j_2} \cdots (Z_{m_d}^{k_d})_{i_d j_d} = \delta_{(i_1 - j_1 - k_1) \bmod m_1} \delta_{(i_2 - j_2 - k_2) \bmod m_2} \cdots \delta_{(i_d - j_d - k_d) \bmod m_d} = \delta_{(\boldsymbol{i}-\boldsymbol{j}-\boldsymbol{k}) \bmod \boldsymbol{m}},$$

where $\delta_{\boldsymbol{r}} = 1$ if $\boldsymbol{r} = \boldsymbol{0}$ and $\delta_{\boldsymbol{r}} = 0$ otherwise. The equality (1.47) is then proved 'blockwise', by showing that, for all $\boldsymbol{i}, \boldsymbol{j} = \boldsymbol{1}, \ldots, \boldsymbol{m}$, the block in position $(\boldsymbol{i}, \boldsymbol{j})$ of the first matrix is equal to the block in position $(\boldsymbol{i}, \boldsymbol{j})$ of the second matrix. Indeed, for all $\boldsymbol{i}, \boldsymbol{j} = \boldsymbol{1}, \ldots, \boldsymbol{m}$, we have

$$\left(\sum_{\boldsymbol{k}=\boldsymbol{0}}^{\boldsymbol{m}-\boldsymbol{1}} Z_{\boldsymbol{m}}^{\boldsymbol{k}} \otimes A_{\boldsymbol{k}} \right)_{\boldsymbol{ij}} = \sum_{\boldsymbol{k}=\boldsymbol{0}}^{\boldsymbol{m}-\boldsymbol{1}} (Z_{\boldsymbol{m}}^{\boldsymbol{k}} \otimes A_{\boldsymbol{k}})_{\boldsymbol{ij}} = \sum_{\boldsymbol{k}=\boldsymbol{0}}^{\boldsymbol{m}-\boldsymbol{1}} (Z_{\boldsymbol{m}}^{\boldsymbol{k}})_{\boldsymbol{ij}} A_{\boldsymbol{k}} = \sum_{\boldsymbol{k}=\boldsymbol{0}}^{\boldsymbol{m}-\boldsymbol{1}} \delta_{(\boldsymbol{i}-\boldsymbol{j}-\boldsymbol{k}) \bmod \boldsymbol{m}} A_{\boldsymbol{k}} = A_{(\boldsymbol{i}-\boldsymbol{j}) \bmod \boldsymbol{m}}.$$

4. Using the identity (1.47), the spectral decomposition (1.46), and the properties of tensor products in Subsection 1.2.1, we obtain

$$\begin{aligned}[A_{(i-j)\bmod \boldsymbol{m}}]_{i,j=\boldsymbol{0}}^{\boldsymbol{m}-\boldsymbol{1}} &= \sum_{k=0}^{m-1} Z_m^k \otimes A_k = \sum_{k=0}^{m-1} (F_m D_m F_m^*) \otimes A_k = \sum_{k=0}^{m-1} (F_m \otimes I_s)(D_m^k \otimes A_k)(F_m \otimes I_s)^* \\ &= (F_m \otimes I_s)\left(\sum_{k=0}^{m-1} D_m^k \otimes A_k\right)(F_m \otimes I_s)^* = (F_m \otimes I_s)\left(\sum_{k=0}^{m-1} \operatorname*{diag}_{j=0,\ldots,m-1} (\mathrm{e}^{2\pi \mathrm{i}(j/m)\cdot k}) \otimes A_k\right)(F_m \otimes I_s)^* \\ &= (F_m \otimes I_s)\left(\sum_{k=0}^{m-1} \operatorname*{diag}_{j=0,\ldots,m-1} (\mathrm{e}^{2\pi \mathrm{i}(j/m)\cdot k} A_k)\right)(F_m \otimes I_s)^* = (F_m \otimes I_s) \operatorname*{diag}_{j=0,\ldots,m-1}\left(\sum_{k=0}^{m-1} \mathrm{e}^{2\pi \mathrm{i}(j/m)\cdot k} A_k\right)(F_m \otimes I_s)^*,\end{aligned}$$

and the thesis follows, since $\sum_{k=0}^{m-1} \mathrm{e}^{2\pi \mathrm{i}(j/m)\cdot k} A_k = \mathbf{g}(2\pi \boldsymbol{j}/\boldsymbol{m})$. □

We remark that an identity like (1.47) also holds for multilevel block Toeplitz matrices. Indeed, it can be shown that the multilevel block Toeplitz matrix in (1.27) has the following expression:

$$[A_{i-j}]_{i,j=\boldsymbol{0}}^{\boldsymbol{m}-\boldsymbol{1}} = \sum_{k=-(m-1)}^{m-1} J_m^{(k)} \otimes A_k, \tag{1.48}$$

where

$$J_{\boldsymbol{m}}^{(\boldsymbol{k})} := J_{m_1}^{(k_1)} \otimes J_{m_2}^{(k_2)} \otimes \cdots \otimes J_{m_d}^{(k_d)}$$

and $J_m^{(k)}$ is the $m \times m$ matrix such that $(J_m^{(k)})_{ij} = 1$ if $i - j = k$ and $(J_m^{(k)})_{ij} = 0$ otherwise:

$$(J_m^{(k)})_{ij} = \delta_{i-j-k}, \quad i, j = 1, \ldots, m, \quad k = -(m-1), \ldots, m-1.$$

1.4.3 GLT sequences

We now focus on the spectral distribution of sequences of matrices obtained from a combination of some algebraic operations on multilevel block Toeplitz matrices and diagonal sampling matrices. These matrix-sequences are particular instances of Generalized Locally Toeplitz (GLT) sequences and, consequently, they belong to the noteworthy GLT algebra. We do not pretend to cover here all the details of this fascinating subject, and so we refer the reader to [68, 63, 64]. We just say that the GLT algebra virtually includes all the matrix-sequences coming from 'regular' discretizations of PDE. We should also say, however, that the spectral distribution of GLT sequences is still under investigation; the latest findings in this direction can be found in [28, 29] (see in particular Theorem 5 in [28] and Theorems 9–10 in [29]).

For every Riemann integrable function $a : [0,1]^d \to \mathbb{C}$ and every $\boldsymbol{m} \in \mathbb{N}^d$, we define the d-level diagonal sampling matrix $D_{\boldsymbol{m}}(a) \in \mathbb{C}^{N(\boldsymbol{m}) \times N(\boldsymbol{m})}$ in the following way:

$$D_{\boldsymbol{m}}(a) := \operatorname*{diag}_{\boldsymbol{j}=\boldsymbol{0},\ldots,\boldsymbol{m}-\boldsymbol{1}} a\left(\frac{\boldsymbol{j}}{\boldsymbol{m}}\right) = \operatorname*{diag}_{\boldsymbol{j}=\boldsymbol{1},\ldots,\boldsymbol{m}} a\left(\frac{\boldsymbol{j}-\boldsymbol{1}}{\boldsymbol{m}}\right), \tag{1.49}$$

where, as always, the multi-index $\boldsymbol{j}$ varies from $\boldsymbol{1}$ to $\boldsymbol{m}-\boldsymbol{1}$ following the lexicographic ordering (1.1).

Theorem 1.12. *For every $k \in \{1, \ldots, \eta\}$ and $l \in \{1, \ldots, \zeta_k\}$, with $\eta, \zeta_1, \ldots, \zeta_\eta$ positive integers, let $\{B_{\boldsymbol{m}}^{(k,l)}\}_{\boldsymbol{m} \in \mathbb{N}^d}$ be either $\{D_{\boldsymbol{m}}(a^{(k,l)})\}_{\boldsymbol{m} \in \mathbb{N}^d}$ or $\{T_{\boldsymbol{m}}(f^{(k,l)})\}_{\boldsymbol{m} \in \mathbb{N}^d}$.[2] Then, setting*

$$X_{\boldsymbol{m}} := \sum_{k=1}^{\eta} \prod_{l=1}^{\zeta_k} B_{\boldsymbol{m}}^{(k,l)},$$

[2] It is understood that $a^{(k,l)}$ is a Riemann integrable function defined on $[0,1]^d$ and $f^{(k,l)} \in L^1([-\pi,\pi]^d)$.

we have

$$\{X_{\boldsymbol{m}}\}_{\boldsymbol{m}\in\mathbb{N}^d} \sim_\sigma \sum_{k=1}^{\eta}\prod_{l=1}^{\zeta_k} b^{(k,l)},$$

where $b^{(k,l)} : [0,1]^d \times [-\pi,\pi]^d \to \mathbb{C}$ *is defined as*

$$b^{(k,l)}(x_1,\dots,x_d;\theta_1,\dots,\theta_d) := \begin{cases} a^{(k,l)}(x_1,\dots,x_d) & \text{if } \{B_{\boldsymbol{m}}^{(k,l)}\} = \{D_{\boldsymbol{m}}(a^{(k,l)})\}, \\ f^{(k,l)}(\theta_1,\dots,\theta_d) & \text{if } \{B_{\boldsymbol{m}}^{(k,l)}\} = \{T_{\boldsymbol{m}}(f^{(k,l)})\}. \end{cases}$$

Moreover, if every $X_{\boldsymbol{m}}$ *is Hermitian, then*

$$\{X_{\boldsymbol{m}}\}_{\boldsymbol{m}\in\mathbb{N}^d} \sim_\lambda \sum_{k=1}^{\eta}\prod_{l=1}^{\zeta_k} b^{(k,l)}.$$

The above theorem combines results from [64, Theorem 2.2] and [63, Theorems 4.5 and 4.8]. These results are formulated in the more general setting of GLT sequences and are based on the a.c.s. notion given in [59] and reported in Section 2.1, but already present in the seminal work by Tilli [68]. Theorem 1.12 could also be extended by including the (pseudo-)inverse of matrices under mild assumptions on the function $b^{(k,l)}$, namely that the set where $b^{(k,l)}$ vanishes has zero Lebesgue measure; see [64, Theorem 2.2].

We now focus on a specific application of Theorem 1.12 which will be of interest in Chapter 5. Given a d-level diagonal sampling matrix $D_{\boldsymbol{m}}(a)$ associated with a Riemann integrable function $a : [0,1]^d \to \mathbb{C}$, we define the symmetric matrix $\tilde{D}_{\boldsymbol{m}}(a)$ as

$$[\tilde{D}_{\boldsymbol{m}}(a)]_{i,j} := [D_{\boldsymbol{m}}(a)]_{\min(i,j),\min(i,j)} = \begin{cases} [D_{\boldsymbol{m}}(a)]_{i,i} & \text{if } i \le j, \\ [D_{\boldsymbol{m}}(a)]_{j,j} & \text{if } i > j, \end{cases} \qquad i,j = 1,\dots,N(\boldsymbol{m}). \tag{1.50}$$

In multi-index notation,

$$[\tilde{D}_{\boldsymbol{m}}(a)]_{\boldsymbol{i},\boldsymbol{j}} = [D_{\boldsymbol{m}}(a)]_{\boldsymbol{i}\wedge\boldsymbol{j},\boldsymbol{i}\wedge\boldsymbol{j}} = \begin{cases} [D_{\boldsymbol{m}}(a)]_{\boldsymbol{i},\boldsymbol{i}} & \text{if } \boldsymbol{i} \preceq \boldsymbol{j}, \\ [D_{\boldsymbol{m}}(a)]_{\boldsymbol{j},\boldsymbol{j}} & \text{if } \boldsymbol{i} \succ \boldsymbol{j}, \end{cases}, \qquad \boldsymbol{i},\boldsymbol{j} = \boldsymbol{1},\dots,\boldsymbol{m}. \tag{1.51}$$

We recall that a d-variate trigonometric polynomial is just a finite linear combination of the Fourier frequencies $\{\mathrm{e}^{\mathrm{i}\boldsymbol{j}\cdot\boldsymbol{\theta}} : \boldsymbol{j} \in \mathbb{Z}^d\}$.

Corollary 1.1. *Let* $\{T_{\boldsymbol{m}}(f_i)\}_{\boldsymbol{m}\in\mathbb{N}^d}$, $i = 1,\dots,\mu$, *be families of* d*-level Toeplitz matrices associated with* d*-variate trigonometric polynomials* f_i, $i = 1,\dots,\mu$, *and let* $\{\tilde{D}_{\boldsymbol{m}}(a_i)\}_{\boldsymbol{m}\in\mathbb{N}^d}$, $i = 1,\dots,\mu$, *be families of matrices associated with Riemann integrable functions* $a_i : [0,1]^d \to \mathbb{C}$, *with* $\tilde{D}_{\boldsymbol{m}}(a_i)$ *defined as in* (1.50)–(1.51). *Then,*

$$\left\{\sum_{i=1}^{\mu} \tilde{D}_{\boldsymbol{m}}(a_i)\circ T_{\boldsymbol{m}}(f_i)\right\}_{\boldsymbol{m}\in\mathbb{N}^d} \sim_\sigma \sum_{i=1}^{\mu} a_i \otimes f_i. \tag{1.52}$$

Moreover, if a_i *and* f_i *are real-valued for all* $i = 1,\dots,\mu$, *then*

$$\left\{\sum_{i=1}^{\mu} \tilde{D}_{\boldsymbol{m}}(a_i)\circ T_{\boldsymbol{m}}(f_i)\right\}_{\boldsymbol{m}\in\mathbb{N}^d} \sim_\lambda \sum_{i=1}^{\mu} a_i \otimes f_i. \tag{1.53}$$

Proof. We decompose the Toeplitz matrix $T_{\boldsymbol{m}}(f_i)$ as

$$T_{\boldsymbol{m}}(f_i) = T_{\boldsymbol{m}}^D(f_i) + T_{\boldsymbol{m}}^L(f_i) + T_{\boldsymbol{m}}^U(f_i),$$

where $T_{\boldsymbol{m}}^D(f_i)$, $T_{\boldsymbol{m}}^L(f_i)$ and $T_{\boldsymbol{m}}^U(f_i)$ form the diagonal, lower and upper triangular matrix of $T_{\boldsymbol{m}}(f_i)$, respectively. The matrices $T_{\boldsymbol{m}}^D(f_i)$, $T_{\boldsymbol{m}}^L(f_i)$ and $T_{\boldsymbol{m}}^U(f_i)$ are also d-level Toeplitz matrices associated with certain trigonometric polynomials f_i^D, f_i^L and f_i^U such that $f_i = f_i^D + f_i^L + f_i^U$. More precisely, we have

$$T_{\boldsymbol{m}}^D(f_i) = [a_{\boldsymbol{i}-\boldsymbol{j}}]_{\boldsymbol{i},\boldsymbol{j}=\boldsymbol{1}}^{\boldsymbol{m}}, \quad \text{where} \quad a_{\boldsymbol{k}} := \begin{cases} (f_i)_{\boldsymbol{k}} & \text{if } \boldsymbol{k} = \boldsymbol{0}, \\ 0 & \text{otherwise} \end{cases} \quad \Rightarrow \quad T_{\boldsymbol{m}}^D(f_i) = T_{\boldsymbol{m}}(f_i^D), \quad \text{with} \quad f_i^D(\boldsymbol{\theta}) = (f_i)_{\boldsymbol{0}},$$

$$T_{\boldsymbol{m}}^L(f_i) = [b_{\boldsymbol{i}-\boldsymbol{j}}]_{\boldsymbol{i},\boldsymbol{j}=\boldsymbol{1}}^{\boldsymbol{m}}, \quad \text{where} \quad b_{\boldsymbol{k}} := \begin{cases} (f_i)_{\boldsymbol{k}} & \text{if } \boldsymbol{k} < \boldsymbol{0}, \\ 0 & \text{otherwise} \end{cases} \quad \Rightarrow \quad T_{\boldsymbol{m}}^L(f_i) = T_{\boldsymbol{m}}(f_i^L), \quad \text{with} \quad f_i^L(\boldsymbol{\theta}) = \sum_{\boldsymbol{k}<\boldsymbol{0}} (f_i)_{\boldsymbol{k}} \mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{\theta}},$$

$$T_{\boldsymbol{m}}^U(f_i) = [c_{\boldsymbol{i}-\boldsymbol{j}}]_{\boldsymbol{i},\boldsymbol{j}=\boldsymbol{1}}^{\boldsymbol{m}}, \quad \text{where} \quad c_{\boldsymbol{k}} := \begin{cases} (f_i)_{\boldsymbol{k}} & \text{if } \boldsymbol{k} > \boldsymbol{0}, \\ 0 & \text{otherwise} \end{cases} \quad \Rightarrow \quad T_{\boldsymbol{m}}^U(f_i) = T_{\boldsymbol{m}}(f_i^U), \quad \text{with} \quad f_i^U(\boldsymbol{\theta}) = \sum_{\boldsymbol{k}>\boldsymbol{0}} (f_i)_{\boldsymbol{k}} \mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{\theta}};$$

since f_i is a trigonometric polynomial, the number of nonzero Fourier coefficients is finite, f_i^L and f_i^U are well-defined and $f_i = f_i^D + f_i^L + f_i^U$.

Now, the matrix $\tilde{D}_{\boldsymbol{m}}(a_i) \circ T_{\boldsymbol{m}}(f_i)$ can be decomposed as

$$\begin{aligned} \tilde{D}_{\boldsymbol{m}}(a_i) \circ T_{\boldsymbol{m}}(f_i) &= \tilde{D}_{\boldsymbol{m}}(a_i) \circ T_{\boldsymbol{m}}^L(f_i) + \tilde{D}_{\boldsymbol{m}}(a_i) \circ T_{\boldsymbol{m}}^D(f_i) + \tilde{D}_{\boldsymbol{m}}(a_i) \circ T_{\boldsymbol{m}}^U(f_i) \\ &= T_{\boldsymbol{m}}^L(f_i) D_{\boldsymbol{m}}(a_i) + D_{\boldsymbol{m}}(a_i) T_{\boldsymbol{m}}^D(f_i) + D_{\boldsymbol{m}}(a_i) T_{\boldsymbol{m}}^U(f_i) \\ &= T_{\boldsymbol{m}}(f_i^L) D_{\boldsymbol{m}}(a_i) + D_{\boldsymbol{m}}(a_i) T_{\boldsymbol{m}}(f_i^D) + D_{\boldsymbol{m}}(a_i) T_{\boldsymbol{m}}(f_i^U), \end{aligned} \tag{1.54}$$

where $D_{\boldsymbol{m}}(a_i)$ is the diagonal sampling matrix associated with a_i, defined in (1.49), and used in the definition of $\tilde{D}_{\boldsymbol{m}}(a_i)$. Because of this decomposition, (1.52) follows from Theorem 1.12 (we have $\eta = 3\mu$ and $\zeta_k = 2$ for all $k = 1, \ldots, 3\mu$). In addition, if a_i and f_i are real-valued for all $i = 1, \ldots, \mu$, then $\tilde{D}_{\boldsymbol{m}}(a_i)$ and $T_{\boldsymbol{m}}(f_i)$ are Hermitian, and so $\tilde{D}_{\boldsymbol{m}}(a_i) \circ T_{\boldsymbol{m}}(f_i)$ is Hermitian as well. Hence, again by the decomposition (1.54), the spectral distribution result (1.53) follows from Theorem 1.12. □

Chapter 2

Some new tools for computing spectral distributions

In this chapter, we present new tools, taken from [34, 35], for determining the asymptotic spectral distribution and the symbol of a sequence of matrices. In Section 2.1 we focus on Hermitian matrix-sequences, while in Section 2.2 we address the non-Hermitian case.

2.1 Tools for determining the spectral distribution of Hermitian matrix-sequences and applications

In this section, we provide a general tool for deducing the spectral distribution of a 'difficult' sequence $\{A_n\}_n$ formed by Hermitian matrices, starting from the one of 'simpler' sequences $\{B_{n,m}\}_n$, again formed by Hermitian matrices, that approximate $\{A_n\}_n$ when $m \to \infty$. The tool is based on the notion of approximating class of sequences (a.c.s.), which was inspired by the work of Paolo Tilli and Stefano Serra-Capizzano, and is applied here in a more general setting. An a.c.s.-based proof of the famous Szegö theorem on the spectral distribution of Toeplitz matrices (item 3 of Theorem 1.8 in the case $d = s = 1$) is finally presented in Subsection 2.1.1. We begin by introducing the notion of approximating class of sequences in the next definition; see [59, Definition 2.1].

Definition 2.1 (**approximating class of sequences**). Let $\{A_n\}_n$ be a matrix-sequence, with A_n of size d_n tending to infinity. An approximating class of sequences (a.c.s.) for $\{A_n\}_n$ is a sequence of matrix-sequences $\{\{B_{n,m}\}_n : m\}$ such that, for every m,

$$A_n = B_{n,m} + R_{n,m} + N_{n,m} \qquad \forall n \geq n_m \tag{2.1}$$

where $\text{rank}(R_{n,m}) \leq \varrho(m)d_n$, $\|N_{n,m}\| \leq \nu(m)$, the quantities n_m, $\varrho(m)$, $\nu(m)$ depend only on m and $\lim_{m\to\infty} \varrho(m) = \lim_{m\to\infty} \nu(m) = 0$.

Roughly speaking, saying that $\{\{B_{n,m}\}_n : m\}$ is an a.c.s. for $\{A_n\}_n$ means that A_n is equal to $B_{n,m}$ plus a small-rank matrix (with respect to the size d_n) plus a small-norm matrix. Lemma 2.1 shows that, if A_n and $B_{n,m}$ are Hermitian, then the small-rank matrix $R_{n,m}$ and the small-norm matrix $N_{n,m}$ in the splitting (2.1) may be supposed Hermitian.

Lemma 2.1. *Let $\{A_n\}_n$ be a sequence of Hermitian matrices, with A_n of size $d_n \to \infty$, and let $\{\{B_{n,m}\}_n : m\}$ be an a.c.s. for $\{A_n\}_n$ formed by Hermitian matrices (i.e. every $B_{n,m}$ is Hermitian). Then, for every m, we have*

$$A_n = B_{n,m} + R_{n,m} + N_{n,m} \qquad \forall n \geq n_m$$

where $R_{n,m}, N_{n,m}$ are Hermitian, $\text{rank}(R_{n,m}) \leq \varrho(m)d_n$, $\|N_{n,m}\| \leq \nu(m)$, *the quantities n_m, $\varrho(m)$, $\nu(m)$ depend only on m and $\lim_{m\to\infty} \varrho(m) = \lim_{m\to\infty} \nu(m) = 0$.*

Proof. Take the real part in (2.1) and use the inequalities $\operatorname{rank}(\Re(X)) \le 2\operatorname{rank}(X)$ and $\|\Re(X)\| \le \|X\|$ to conclude that, by replacing $R_{n,m}, N_{n,m}$ with $\Re(R_{n,m}), \Re(N_{n,m})$ (if necessary), we can assume $R_{n,m}, N_{n,m}$ to be Hermitian. □

Now we turn to the main theorems of this section (Theorems 2.1 and 2.3), which provide a general tool for determining the spectral distribution of a 'difficult' matrix-sequence $\{A_n\}_n$ formed by Hermitian matrices, starting from the knowledge of the spectral distribution of simpler matrix-sequences $\{B_{n,m}\}_n$, $m = 1, 2, 3, \ldots$, again formed by Hermitian matrices. Recall that, for any Hermitian matrix $X \in \mathbb{C}^{m\times m}$, the eigenvalues of X are arranged in non-increasing order: $\lambda_1(X) \ge \ldots \ge \lambda_m(X)$; moreover, $\lambda_j(X) = +\infty$ if $j \le 0$ and $\lambda_j(X) = -\infty$ if $j \ge m+1$ (see Section 1.1). If $H : \mathbb{R} \to \mathbb{R}$, we define $H(\infty) := \lim_{x\to\infty} H(x)$ (whenever the limit exists). Similarly, $H(-\infty) := \lim_{x\to-\infty} H(x)$.

Theorem 2.1. *Let $\{A_n\}_n$ be a sequence of Hermitian matrices, with A_n of size $d_n \to \infty$. Assume that*

1. *$\{\{B_{n,m}\}_n : m\}$ is an a.c.s. for $\{A_n\}_n$ formed by Hermitian matrices;*
2. *for every m and every $F \in C_c^1(\mathbb{R})$, there exists $\lim_{n\to\infty} \frac{1}{d_n}\sum_{j=1}^{d_n} F(\lambda_j(B_{n,m})) =: \phi_m(F) \in \mathbb{C}$;*
3. *for every $F \in C_c^1(\mathbb{R})$, there exists $\lim_{m\to\infty} \phi_m(F) =: \phi(F) \in \mathbb{C}$.*

Then, for all $F \in C_c^1(\mathbb{R})$,

$$\exists \lim_{n\to\infty} \frac{1}{d_n}\sum_{j=1}^{d_n} F(\lambda_j(A_n)) = \phi(F). \tag{2.2}$$

Proof. The technique of this proof is taken from [59, Proposition 2.3], where an analogous result was proved for the singular values instead of the eigenvalues. We first observe that it suffices to prove (2.2) for those test functions $F \in C_c^1(\mathbb{R})$ that are real-valued. Indeed, any (complex-valued) $F \in C_c^1(\mathbb{R})$ can be decomposed as $F = \Re(F) + \mathrm{i}\Im(F)$, where $\Re(F), \Im(F) \in C_c^1(\mathbb{R})$. Thus, once we have proved (2.2) for all real-valued functions in $C_c^1(\mathbb{R})$, we have

$$\lim_{n\to\infty} \frac{1}{d_n}\sum_{j=1}^{d_n} F(\lambda_j(A_n)) = \lim_{n\to\infty} \frac{1}{d_n}\sum_{j=1}^{d_n} \Big[\Re(F(\lambda_j(A_n))) + \mathrm{i}\Im(F(\lambda_j(A_n)))\Big] = \phi(\Re(F)) + \mathrm{i}\phi(\Im(F)) = \phi(F),$$

where the last equality holds by the linearity of the functional ϕ, which follows from its definition.

Now, let $F \in C_c^1(\mathbb{R})$ be real-valued. For all n, m we have

$$\left|\frac{1}{d_n}\sum_{j=1}^{d_n} F(\lambda_j(A_n)) - \phi(F)\right| \le \left|\frac{1}{d_n}\sum_{j=1}^{d_n} F(\lambda_j(A_n)) - \frac{1}{d_n}\sum_{j=1}^{d_n} F(\lambda_j(B_{n,m}))\right| + \left|\frac{1}{d_n}\sum_{j=1}^{d_n} F(\lambda_j(B_{n,m})) - \phi_m(F)\right| + |\phi_m(F) - \phi(F)|. \tag{2.3}$$

By hypothesis, the second term in the right-hand side tends to 0 for $n \to \infty$, while the third one tends to 0 for $m \to \infty$. Therefore, if we prove that

$$\lim_{m\to\infty} \limsup_{n\to\infty} \left|\frac{1}{d_n}\sum_{j=1}^{d_n} F(\lambda_j(A_n)) - \frac{1}{d_n}\sum_{j=1}^{d_n} F(\lambda_j(B_{n,m}))\right| = 0, \tag{2.4}$$

then, passing first to the $\limsup_{n\to\infty}$ and then to the $\lim_{m\to\infty}$ in (2.3), we get the thesis.

In conclusion, we only have to prove (2.4). To this end, we recall that $\{\{B_{n,m}\}_n : m\}$ is an a.c.s. for $\{A_n\}_n$ and that $A_n, B_{n,m}$ are Hermitian as in Lemma 2.1. Hence, for every m,

$$A_n = B_{n,m} + R_{n,m} + N_{n,m} \qquad \forall n \geq n_m$$

where $R_{n,m}, N_{n,m}$ are Hermitian, $\operatorname{rank}(R_{n,m}) \leq \varrho(m)d_n$, $\|N_{n,m}\| \leq \nu(m)$, the quantities n_m, $\varrho(m)$, $\nu(m)$ depend only on m and $\lim_{m\to\infty} \varrho(m) = \lim_{m\to\infty} \nu(m) = 0$. We can then write, for every m and every $n \geq n_m$,

$$\begin{aligned}
&\left|\frac{1}{d_n}\sum_{j=1}^{d_n} F(\lambda_j(A_n)) - \frac{1}{d_n}\sum_{j=1}^{d_n} F(\lambda_j(B_{n,m}))\right| \\
&\leq \left|\frac{1}{d_n}\sum_{j=1}^{d_n} F(\lambda_j(A_n)) - \frac{1}{d_n}\sum_{j=1}^{d_n} F(\lambda_j(B_{n,m}+R_{n,m}))\right| + \left|\frac{1}{d_n}\sum_{j=1}^{d_n} F(\lambda_j(B_{n,m}+R_{n,m})) - \frac{1}{d_n}\sum_{j=1}^{d_n} F(\lambda_j(B_{n,m}))\right|. \qquad (2.5)
\end{aligned}$$

We will consider separately the two terms in the right-hand side of (2.5), and we will show that each of them is bounded from above by a quantity depending only on m and tending to 0 as $m \to \infty$. After this, (2.4) is proved and the thesis follows.

In order to estimate the first term in the right-hand side of (2.5), we use the Weyl's perturbation theorem; see [7, p. 63]. We have

$$\begin{aligned}
&\left|\frac{1}{d_n}\sum_{j=1}^{d_n} F(\lambda_j(A_n)) - \frac{1}{d_n}\sum_{j=1}^{d_n} F(\lambda_j(B_{n,m}+R_{n,m}))\right| \leq \frac{1}{d_n}\sum_{j=1}^{d_n} \left|F(\lambda_j(A_n)) - F(\lambda_j(B_{n,m}+R_{n,m}))\right| \\
&\leq \frac{1}{d_n}\sum_{j=1}^{d_n} \|F'\|_\infty \left|\lambda_j(A_n) - \lambda_j(B_{n,m}+R_{n,m})\right| \leq \|F'\|_\infty \|A_n - B_{n,m} - R_{n,m}\| = \|F'\|_\infty \|N_{n,m}\| \leq \|F'\|_\infty \nu(m),
\end{aligned}$$

which tends to 0 as $m \to \infty$.

In order to estimate the second term in the right-hand side of (2.5), we will use the interlacing Theorem 1.4. We first observe that F can be expressed as the difference between two nonnegative, non-decreasing, bounded functions:

$$F = H - K, \qquad H(x) := \int_{-\infty}^{x} (F')_+(t)dt, \qquad K(x) := \int_{-\infty}^{x} (F')_-(t)dt,$$

where $(F')_+ := \max(F', 0)$ and $(F')_- := \max(-F', 0)$. Hence, for the second term in the right-hand side of (2.5) we have

$$\begin{aligned}
&\left|\frac{1}{d_n}\sum_{j=1}^{d_n} F(\lambda_j(B_{n,m}+R_{n,m})) - \frac{1}{d_n}\sum_{j=1}^{d_n} F(\lambda_j(B_{n,m}))\right| \\
&\leq \left|\frac{1}{d_n}\sum_{j=1}^{d_n} H(\lambda_j(B_{n,m}+R_{n,m})) - \frac{1}{d_n}\sum_{j=1}^{d_n} H(\lambda_j(B_{n,m}))\right| + \left|\frac{1}{d_n}\sum_{j=1}^{d_n} K(\lambda_j(B_{n,m}+R_{n,m})) - \frac{1}{d_n}\sum_{j=1}^{d_n} K(\lambda_j(B_{n,m}))\right|. \qquad (2.6)
\end{aligned}$$

Defining $r_{n,m} := \operatorname{rank}(R_{n,m}) \leq \varrho(m)d_n$, Theorem 1.4 gives

$$\lambda_{j-r_{n,m}}(B_{n,m}) \geq \lambda_j(B_{n,m}+R_{n,m}) \geq \lambda_{j+r_{n,m}}(B_{n,m}), \qquad \forall j = 1, \ldots, d_n,$$

and, moreover, it is clear from our notation that

$$\lambda_{j-r_{n,m}}(B_{n,m}) \geq \lambda_j(B_{n,m}) \geq \lambda_{j+r_{n,m}}(B_{n,m}), \qquad \forall j = 1, \ldots, d_n.$$

Therefore, recalling the monotonicity and nonnegativity of H,

$$\left|\frac{1}{d_n}\sum_{j=1}^{d_n} H(\lambda_j(B_{n,m}+R_{n,m})) - \frac{1}{d_n}\sum_{j=1}^{d_n} H(\lambda_j(B_{n,m}))\right| \le \frac{1}{d_n}\sum_{j=1}^{d_n}\left|H(\lambda_j(B_{n,m}+R_{n,m})) - H(\lambda_j(B_{n,m}))\right|$$
$$\le \frac{1}{d_n}\sum_{j=1}^{d_n}\left|H(\lambda_{j-r_{n,m}}(B_{n,m})) - H(\lambda_{j+r_{n,m}}(B_{n,m}))\right| = \frac{1}{d_n}\sum_{j=1}^{d_n} H(\lambda_{j-r_{n,m}}(B_{n,m})) - \frac{1}{d_n}\sum_{j=1}^{d_n} H(\lambda_{j+r_{n,m}}(B_{n,m}))$$
$$= \frac{1}{d_n}\sum_{j=1-r_{n,m}}^{d_n-r_{n,m}} H(\lambda_j(B_{n,m})) - \frac{1}{d_n}\sum_{j=1+r_{n,m}}^{d_n+r_{n,m}} H(\lambda_j(B_{n,m})) = \frac{1}{d_n}\sum_{j=1-r_{n,m}}^{r_{n,m}} H(\lambda_j(B_{n,m})) - \frac{1}{d_n}\sum_{j=d_n-r_{n,m}+1}^{d_n+r_{n,m}} H(\lambda_j(B_{n,m}))$$
$$\le \frac{1}{d_n}\sum_{j=1-r_{n,m}}^{r_{n,m}} H(\lambda_j(B_{n,m})) \le \frac{2r_{n,m}H(\infty)}{d_n} \le 2\varrho(m)\|H\|_\infty.$$

Similarly, one can show that the second term in the right-hand side of (2.6) is bounded from above by $2\varrho(m)\|K\|_\infty$, implying that the quantity in (2.6), namely the second term in the right-hand side of (2.5), is less than or equal to $2(\|H\|_\infty + \|K\|_\infty)\varrho(m)$. Since the latter tends to 0 as $m \to \infty$, the thesis is proved. □

The only unpleasant point about Theorem 2.1 is that, in traditional formulations of asymptotic spectral distribution results, the usual set of test functions F is $C_c(\mathbb{C})$ or $C_c(\mathbb{R})$, but not $C_c^1(\mathbb{R})$; see also Definitions 1.1–1.2. However, this point is readily settled in Theorem 2.3, where we prove that, under the same hypotheses of Theorem 2.1, if the second and third assumptions are met for every $F \in C_c(\mathbb{R})$, then (2.2) holds for every $F \in C_c(\mathbb{R})$. For the proof of Theorem 2.3, we shall use the following corollary of the Banach-Steinhaus theorem [50].

Theorem 2.2. *Let $\mathscr{E}, \mathscr{F}$ be normed vector spaces, with $\mathscr{E}$ a Banach space, and let $T_n : \mathscr{E} \to \mathscr{F}$ be a sequence of continuous linear operators. Assume that, for all $x \in \mathscr{E}$, there exists $\lim_{n\to\infty} T_n x =: Tx \in \mathscr{F}$. Then,*

- $\sup \|T_n\| < \infty$;
- *$T : \mathscr{E} \to \mathscr{F}$ is a continuous linear operator with $\|T\| \le \liminf_{n\to\infty} \|T_n\|$.*

Theorem 2.3. *Let $\{A_n\}_n$ be a sequence of Hermitian matrices, with A_n of size $d_n \to \infty$. Assume that*

1. *$\{\{B_{n,m}\}_n : m\}$ is an a.c.s. for $\{A_n\}_n$ formed by Hermitian matrices;*
2. *for every m and every $F \in C_c(\mathbb{R})$, there exists $\lim_{n\to\infty} \frac{1}{d_n}\sum_{j=1}^{d_n} F(\lambda_j(B_{n,m})) =: \phi_m(F) \in \mathbb{C}$;*
3. *for every $F \in C_c(\mathbb{R})$, there exists $\lim_{m\to\infty} \phi_m(F) =: \phi(F) \in \mathbb{C}$.*

Then $\phi : (C_c(\mathbb{R}), \|\cdot\|_\infty) \to \mathbb{C}$ is a continuous linear functional with $\|\phi\| \le 1$, and, for all $F \in C_c(\mathbb{R})$,

$$\exists \lim_{n\to\infty} \frac{1}{d_n}\sum_{j=1}^{d_n} F(\lambda_j(A_n)) = \phi(F). \tag{2.7}$$

Proof. For fixed n, m, let

$$\phi_{n,m}(F) := \frac{1}{d_n}\sum_{j=1}^{d_n} F(\lambda_j(B_{n,m})) : (C_c(\mathbb{R}), \|\cdot\|_\infty) \to \mathbb{C}.$$

It is clear that each $\phi_{n,m}$ is a continuous linear functional on $(C_c(\mathbb{R}), \|\cdot\|_\infty)$ with $\|\phi_{n,m}\| \leq 1$. Indeed, the linearity of $\phi_{n,m}$ is obvious and the inequality $|\phi_{n,m}(F)| \leq \|F\|_\infty$, which is satisfied for all $F \in C_c(\mathbb{R})$, yields the continuity of $\phi_{n,m}$ as well as the bound $\|\phi_{n,m}\| \leq 1$. The functional ϕ_m is the pointwise limit of $\phi_{n,m}$ as $n \to \infty$. Hence, by Theorem 2.2, $\phi_m : (C_c(\mathbb{R}), \|\cdot\|_\infty) \to \mathbb{C}$ is a continuous linear functional on $(C_c(\mathbb{R}), \|\cdot\|_\infty)$ with $\|\phi_m\| \leq 1$. The functional ϕ is the pointwise limit of ϕ_m as $m \to \infty$. Hence, again by Theorem 2.2, ϕ is a continuous linear functional on $(C_c(\mathbb{R}), \|\cdot\|_\infty)$ with $\|\phi\| \leq 1$.

Now, fix $F \in C_c(\mathbb{R})$. For all $\epsilon > 0$ we can find $F_\epsilon \in C_c^1(\mathbb{R})$ such that $\|F - F_\epsilon\|_\infty \leq \epsilon$. As a consequence, for all $\epsilon > 0$ and all n we have

$$\left|\frac{1}{d_n}\sum_{j=1}^{d_n} F(\lambda_j(A_n)) - \phi(F)\right| \leq \left|\frac{1}{d_n}\sum_{j=1}^{d_n} F(\lambda_j(A_n)) - \frac{1}{d_n}\sum_{j=1}^{d_n} F_\epsilon(\lambda_j(A_n))\right| + \left|\frac{1}{d_n}\sum_{j=1}^{d_n} F_\epsilon(\lambda_j(A_n)) - \phi(F_\epsilon)\right| + |\phi(F_\epsilon) - \phi(F)|$$

$$\leq \|F - F_\epsilon\|_\infty + \left|\frac{1}{d_n}\sum_{j=1}^{d_n} F_\epsilon(\lambda_j(A_n)) - \phi(F_\epsilon)\right| + |\phi(F_\epsilon) - \phi(F)|.$$

Considering that (2.7) holds for F_ϵ by Theorem 2.1, we have

$$\limsup_{n\to\infty} \left|\frac{1}{d_n}\sum_{j=1}^{d_n} F(\lambda_j(A_n)) - \phi(F)\right| \leq \epsilon + |\phi(F_\epsilon) - \phi(F)|.$$

Passing to the limit as $\epsilon \to 0$ and taking into account the continuity of ϕ, we obtain

$$\limsup_{n\to\infty} \left|\frac{1}{d_n}\sum_{j=1}^{d_n} F(\lambda_j(A_n)) - \phi(F)\right| = 0,$$

which means that (2.7) holds for every $F \in C_c(\mathbb{R})$. □

2.1.1 An a.c.s.-based proof of the Szegö theorem on the spectral distribution of Toeplitz matrices

As an application of Theorem 2.3, we present in this subsection a new proof of the famous Szegö theorem on the spectral distribution of Toeplitz matrices, which is nothing else than item 3 of Theorem 1.8 in the case $d = s = 1$. This theorem, originally appeared in [39], has undergone several extensions [70, 11] until the final version by Tilli [67], which includes all the others as particular cases. For the proof of the various extensions, other arguments, different from the one used in [39], have been proposed. In particular, Tilli's argument [67] is similar to the one that we are going to present, but it does not make use of the concept of a.c.s., which was introduced later. To our knowledge, an a.c.s.-based proof, like the one that we are going to see in the following, has never appeared in the literature. Such proof is particularly useful to understand how the a.c.s. notion can be seen as a fundamental definition that sets the basis for an approximation theory for matrix-sequences, of which Theorems 2.1 and 2.3 are fundamental stones.

Let us start with reformulating Definition 1.1 in terms of functionals ϕ and in the case where the symbol $\mathbf{f}$ is a univariate scalar function f (i.e., in the case $d = s = 1$).

Definition 2.2. Let $\{A_n\}_n$ be a sequence of matrices, with A_n of size $d_n \to \infty$, and let $f : D \to \mathbb{C}$ be a measurable function, defined on a measurable set $D \subset \mathbb{R}$ with $0 < m(D) < \infty$. We say that $\{A_n\}_n$ has an asymptotic spectral distribution described by f, in symbols $\{A_n\}_n \sim_\lambda f$, if

$$\lim_{n\to\infty} \frac{1}{d_n}\sum_{j=1}^{d_n} F(\lambda_j(A_n)) = \phi_f(F), \qquad \forall F \in C_c(\mathbb{C}), \tag{2.8}$$

where

$$\phi_f(F) := \frac{1}{m(D)}\int_D F(f(x))dx. \tag{2.9}$$

In the case where $\{A_n\}_n$ is formed by Hermitian matrices and f is real-valued, all the eigenvalues of A_n are real and writing $\{A_n\}_n \sim_\lambda f$ is equivalent to saying that (2.8) is verified for every test function $F \in C_c(\mathbb{R})$, with ϕ_f still defined by (2.9). Concerning the functional ϕ_f, we record the following property, of interest later on.

Lemma 2.2. *Let $f_m : D \to \mathbb{C}$ be a sequence of measurable functions, defined on a measurable set $D \subset \mathbb{R}$ with $0 < m(D) < \infty$, and assume that f_m converges in measure to some measurable function $f : D \to \mathbb{C}$. Then,*

$$\phi_{f_m}(F) \to \phi_f(F), \qquad \forall F \in C_c(\mathbb{C}). \tag{2.10}$$

In particular, if f_m, f are real-valued then

$$\phi_{f_m}(F) \to \phi_f(F), \qquad \forall F \in C_c(\mathbb{R}). \tag{2.11}$$

Proof. Let $F \in C_c(\mathbb{C})$ and $\epsilon > 0$. Defining $\{|f_m - f| \geq \epsilon\} := \{x \in D : |f_m(x) - f(x)| \geq \epsilon\}$ and $\{|f_m - f| < \epsilon\} := \{x \in D : |f_m(x) - f(x)| < \epsilon\}$, we have

$$\begin{aligned}
|\phi_{f_m}(F) - \phi_f(F)| &\leq \frac{1}{m(D)} \int_D |F(f_m(x)) - F(f(x))| dx \\
&= \frac{1}{m(D)} \int_{\{|f_m - f| \geq \epsilon\}} |F(f_m(x)) - F(f(x))| dx + \frac{1}{m(D)} \int_{\{|f_m - f| < \epsilon\}} |F(f_m(x)) - F(f(x))| dx \\
&\leq \frac{2\|F\|_\infty m(\{|f_m - f| \geq \epsilon\})}{m(D)} + \omega_F(\epsilon),
\end{aligned} \tag{2.12}$$

where ω_F is the modulus of continuity of F. Note that $\lim_{m\to\infty} m(\{|f_m - f| \geq \epsilon\}) = 0$ (because $f_m \to f$ in measure) and $\lim_{\epsilon\to 0} \omega_F(\epsilon) = 0$ (because F is uniformly continuous by the Heine-Cantor theorem). Hence, passing first to the $\limsup_{m\to\infty}$ and then to the $\lim_{\epsilon\to 0}$ in (2.12), we get (2.10). In the case where f_m, f are real-valued, (2.10) immediately implies (2.11), because every $F \in C_c(\mathbb{R})$ is obtained as the restriction to $\mathbb{R}$ of some $\tilde{F} \in C_c(\mathbb{C})$. □

Now, let $f : [-\pi, \pi] \to \mathbb{C}$ be a function in $L^1([-\pi, \pi])$, and denote its Fourier coefficients by

$$f_j := \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x) \mathrm{e}^{-\mathrm{i}jx} dx, \qquad j \in \mathbb{Z}.$$

We recall from Subsection 1.4.1 that, for every $n \geq 1$, the n-th Toeplitz matrix associated with f is defined as

$$T_n(f) := [f_{i-j}]_{i,j=1}^n.$$

In the case where f is real-valued, all the matrices $T_n(f)$ are Hermitian and the following result holds.

Theorem 2.4 (**Szegö**). *Let f be a real-valued function in $L^1([-\pi, \pi])$, then $\{T_n(f)\}_n \sim_\lambda f$.*

Our goal is to provide a proof of Theorem 2.4 based on the notion of a.c.s. and, especially, on Theorem 2.3. To this end, we need some auxiliary lemmas. If $f \in L^1([-\pi, \pi])$, we set

$$\|f\|_{L^1([-\pi,\pi])} := \int_{-\pi}^{\pi} |f(x)| dx.$$

Lemma 2.3. *Let $f \in L^1([-\pi, \pi])$ and $n \in \mathbb{N}$, then*

$$|||T_n(f)|||_1 \leq Cn\|f\|_{L^1([-\pi,\pi])}, \tag{2.13}$$

where $C = 1/\pi$.

Proof. The proof is taken from [67, Lemma 3.1]. We first observe that, if $f \geq 0$, then $T_n(f)$ is HPSD by Theorem 1.8 and, consequently, the singular values and the eigenvalues of $T_n(f)$ coincide. Thus,

$$|||T_n(f)|||_1 = \sum_{j=1}^{n} \lambda_j(T_n(f)) = \mathrm{trace}(T_n(f)) = nf_0 = \frac{n}{2\pi}\|f\|_{L^1([-\pi,\pi])}, \tag{2.14}$$

which proves the thesis whenever f is nonnegative.

Now suppose that $f \in L^1([-\pi,\pi])$ is arbitrary, and consider the following nonnegative functions:

$$\begin{aligned} \Re(f)^+(x) &= \max(\Re(f(x)), 0), & \Re(f)^-(x) &= \max(-\Re(f(x)), 0), \\ \Im(f)^+(x) &= \max(\Im(f(x)), 0), & \Im(f)^-(x) &= \max(-\Im(f(x)), 0). \end{aligned}$$

Then

$$f = \Re(f)^+ - \Re(f)^- + \mathrm{i}\,\Im(f)^+ - \mathrm{i}\,\Im(f)^-$$

and

$$T_n(f) = T_n(\Re(f)^+) - T_n(\Re(f)^-) + \mathrm{i}\,T_n(\Im(f)^+) - \mathrm{i}\,T_n(\Im(f)^-).$$

Since $\Re(f)^+$, $\Re(f)^-$, $\Im(f)^+$, $\Im(f)^-$ are nonnegative, by (2.14) we have

$$\begin{aligned} |||T_n(f)|||_1 &\leq \left|\left|\left|T_n(\Re(f)^+)\right|\right|\right|_1 + \left|\left|\left|T_n(\Re(f)^-)\right|\right|\right|_1 + \left|\left|\left|T_n(\Im(f)^+)\right|\right|\right|_1 + \left|\left|\left|T_n(\Im(f)^-)\right|\right|\right|_1 \\ &= \frac{n}{2\pi}\int_{-\pi}^{\pi} \Re(f)^+(x)dx + \frac{n}{2\pi}\int_{-\pi}^{\pi} \Re(f)^-(x)dx + \frac{n}{2\pi}\int_{-\pi}^{\pi} \Im(f)^+(x)dx + \frac{n}{2\pi}\int_{-\pi}^{\pi} \Im(f)^-(x)dx \\ &= \frac{n}{2\pi}\int_{-\pi}^{\pi} \left[\Re(f)^+(x) + \Re(f)^-(x) + \Im(f)^+(x) + \Im(f)^-(x)\right] dx = \frac{n}{2\pi}\int_{-\pi}^{\pi} \left[|\Re(f(x))| + |\Im(f(x))|\right] dx \\ &\leq \frac{n}{2\pi}\int_{-\pi}^{\pi} 2|f(x)|dx. \end{aligned}$$

□

The inequality (2.13) is part of a large family of inequalities involving Toeplitz matrices and Schatten p-norms. In particular, in a finer version of (2.13), the constant $C = 1/\pi$ is replaced by $C = 1/(2\pi)$, which is precisely the constant obtained in (2.14), in the case where f is nonnegative. We refer the interested reader to [61, Corollary 3.5].

Lemma 2.4. *Let $Z_{n,m}$ be a matrix of size n, and assume that, for every m,*

$$\left|\left|\left|Z_{n,m}\right|\right|\right|_1 \leq \alpha(m)n, \qquad \forall n \geq n_m,$$

where $\alpha(m)$, n_m depend only on m. Then, for every m,

$$Z_{n,m} = R_{n,m} + N_{n,m}, \qquad \forall n \geq n_m,$$

where $\mathrm{rank}(R_{n,m}) \leq \sqrt{\alpha(m)}\,n$ *and* $\|N_{n,m}\| \leq \sqrt{\alpha(m)}$.

Proof. The thesis may be somehow derived from the results in [60] (see in particular Theorem 4.4 and Corollaries 4.1–4.2). However, since the derivation is not so plain, we include a short and direct proof for the sake of the reader.

Fix m and $n \geq n_m$. Since $\left|\left|\left|Z_{n,m}\right|\right|\right|_1 \leq \alpha(m)n$, the number of singular values of $Z_{n,m}$ that exceed $\sqrt{\alpha(m)}$ cannot be larger than $\sqrt{\alpha(m)}\,n$. Let $Z_{n,m} = U_{n,m}\Sigma_{n,m}V_{n,m}^*$ be a singular value decomposition of $Z_{n,m}$ and write

$$Z_{n,m} = U_{n,m}\Sigma_{n,m}V_{n,m}^* = U_{n,m}\Sigma_{n,m}^{(1)}V_{n,m}^* + U_{n,m}\Sigma_{n,m}^{(2)}V_{n,m}^*,$$

where $\Sigma^{(1)}_{n,m}$ is obtained from $\Sigma_{n,m}$ by setting to 0 all the singular values that are less than or equal to $\sqrt{\alpha(m)}$, while $\Sigma^{(2)}_{n,m} := \Sigma_{n,m} - \Sigma^{(1)}_{n,m}$ is obtained from $\Sigma_{n,m}$ by setting to 0 all the singular values that exceed $\sqrt{\alpha(m)}$. Then

$$Z_{n,m} = R_{n,m} + N_{n,m},$$

where $R_{n,m} := U_{n,m}\Sigma^{(1)}_{n,m}V^*_{n,m}$ and $N_{n,m} := U_{n,m}\Sigma^{(2)}_{n,m}V^*_{n,m}$ satisfy $\operatorname{rank}(R_{n,m}) \le \sqrt{\alpha(m)}\,n$ and $\|N_{n,m}\| \le \sqrt{\alpha(m)}$. □

The next lemma shows that Theorem 2.4 holds in the case where f is a trigonometric polynomial.

Lemma 2.5. *Let p be a real-valued trigonometric polynomial, then $\{T_n(p)\}_n \sim_\lambda p$.*

Proof. Let $p(x) := \sum_{j=-s}^{s} p_j \mathrm{e}^{\mathrm{i}jx}$ be a real-valued trigonometric polynomial. Note that $p_{-j} = \overline{p_j}$ for all $j = -s, \dots, s$, because p is real. For every $n \ge 2s+1$, consider the following decomposition of $T_n(p)$:

$$T_n(p) = \begin{bmatrix} p_0 & \cdots & p_{-s} & & & p_s & \cdots & p_1 \\ \vdots & \ddots & & \ddots & & & \ddots & \vdots \\ p_s & & \ddots & & \ddots & & & p_s \\ & \ddots & & \ddots & & \ddots & & \\ & & \ddots & & \ddots & & \ddots & \\ p_{-s} & & & \ddots & & \ddots & & p_{-s} \\ \vdots & \ddots & & & \ddots & & \ddots & \vdots \\ p_{-1} & \cdots & p_{-s} & & & p_s & \cdots & p_0 \end{bmatrix} - \begin{bmatrix} & & & & & & p_s & \cdots & p_1 \\ & & & & & & & \ddots & \vdots \\ & & & & & & & & p_s \\ & & & & & & & & \\ & & & & & & & & \\ & & & & & & & & \\ p_{-s} & & & & & & & & \\ \vdots & \ddots & & & & & & & \\ p_{-1} & \cdots & p_{-s} & & & & & & \end{bmatrix} =: C_n(p) - Z_n(p). \tag{2.15}$$

$C_n(p)$ is a (Hermitian) circulant matrix and hence its eigenvalues are explicitly known (see Theorem 1.11):

$$\lambda_j(C_n(p)) = p\left(\frac{2\pi(j-1)}{n}\right), \qquad j = 1, \dots, n.$$

Therefore, for every test function $F \in C_c(\mathbb{R})$, we have

$$\lim_{n\to\infty} \frac{1}{n}\sum_{j=1}^{n} F(\lambda_j(C_n(p))) = \lim_{n\to\infty} \frac{1}{n}\sum_{j=0}^{n-1} F\left(p\left(\frac{2\pi j}{n}\right)\right) = \frac{1}{2\pi}\int_0^{2\pi} F(p(x))dx = \frac{1}{2\pi}\int_{-\pi}^{\pi} F(p(x))dx,$$

where the last equality holds because p is periodic with period 2π, while the second equality is due to the fact that $\frac{2\pi}{n}\sum_{j=0}^{n-1} F(p(\frac{2\pi j}{n}))$ is a quadrature formula for approximating $\int_0^{2\pi} F(p(x))dx$ and converges to this integral as $n \to \infty$, because the function $F(p(x))$ is continuous on $[0, 2\pi]$. Thus, $\{C_n(p)\}_n \sim_\lambda p$.

Now, for every n, m, set $A_n := T_n(p)$ and $B_{n,m} := C_n(p)$. We have just proved that $\{B_{n,m}\}_n \sim_\lambda p$ for every m. All the hypotheses of Theorem 2.3 are then satisfied (with $\phi_m = \phi = \phi_p$, as given by (2.9) for $f = p$) if $\{\{B_{n,m}\}_n : m\}$ is an a.c.s. for $\{A_n\}_n$. But this is clearly true, because, in view of (2.15), for every m we have

$$A_n = B_{n,m} + R_{n,m} + N_{n,m}, \qquad \forall n \ge n_m,$$

where $N_{n,m}$ is the zero matrix and $R_{n,m} := -Z_n(p)$ satisfies $\operatorname{rank}(R_{n,m}) \le 2s \le \varrho(m)n$ for all $n \ge n_m$, provided that we choose $n_m = m$ and $\varrho(m) = 2s/m$. All the hypotheses of Theorem 2.3 are then satisfied and so $\{T_n(p)\}_n \sim_\lambda p$. □

Proof of Theorem 2.4. Take a sequence of real trigonometric polynomials p_m such that $p_m \to f$ in $L^1([-\pi,\pi])$. We prove that the assumptions of Theorem 2.3 are satisfied with

$$A_n = T_n(f), \qquad B_{n,m} = T_n(p_m), \qquad \phi_m = \phi_{p_m}, \qquad \phi = \phi_p.$$

We first note that $T_n(f)$ and $T_n(p_m)$ are Hermitian, because f and p_m are real. By Lemma 2.5, for every m we have $\{T_n(p_m)\}_n \sim_\lambda p_m$. By Lemma 2.2, $\phi_{p_m}(F) \to \phi_p(F)$ for all $F \in C_c(\mathbb{R})$, because $p_m \to f$ in $L^1([-\pi,\pi])$ and hence, a fortiori, $p_m \to f$ in measure. It remains to show that $\{\{T_n(p_m)\}_n : m\}$ is an a.c.s. for $\{T_n(f)\}_n$.

By Lemma 2.3, for every n, m we have

$$|||T_n(f) - T_n(p_m)|||_1 = |||T_n(f - p_m)|||_1 \le (n/\pi)\|f - p_m\|_{L^1([-\pi,\pi])} = \alpha(m)n,$$

where $\alpha(m) := (1/\pi)\|f - p_m\|_{L^1([-\pi,\pi])}$. Thus, by Lemma 2.4, for every n, m we have

$$T_n(f) - T_n(p_m) = R_{n,m} + N_{n,m},$$

where $\operatorname{rank}(R_{n,m}) \le \sqrt{\alpha(m)}\,n$ and $\|N_{n,m}\| \le \sqrt{\alpha(m)}$. Since $\alpha(m) \to 0$ as $m \to \infty$, $\{\{T_n(p_m)\}_n : m\}$ is an a.c.s. for $\{T_n(f)\}_n$. The thesis now follows from Theorem 2.3. □

We conclude by saying that a completely analogous proof as the one presented in this subsection can be given also for the multilevel block version of the Szegö theorem stated in item 3 of Theorem 1.8. Here, we decided to address only the monolevel scalar case in order to avoid technicalities and notational complications, so as to make more clear the 'a.c.s. idea' and the way in which Theorem 2.3 is applied in practice.

2.2 Tools for determining the spectral distribution of non-Hermitian perturbations of Hermitian matrix-sequences and applications

The tools presented in this section serve to determine the spectral distribution of a matrix-sequence of the form $\{X_n+Y_n\}$, where X_n is Hermitian and Y_n is a perturbation of X_n with small trace-norm with respect to the matrix size d_n. More precisely, given a matrix-sequence $\{X_n\}$, with X_n Hermitian of size d_n tending to infinity, we consider the sequence $\{X_n + Y_n\}$, where $\{Y_n\}$ is an arbitrary (non-Hermitian) perturbation of $\{X_n\}$. In this section, we prove that $\{X_n + Y_n\}$ has an asymptotic spectral distribution if: $\{X_n\}$ has an asymptotic spectral distribution, the spectral norms $\|X_n\|, \|Y_n\|$ are uniformly bounded with respect to n, and $|||Y_n|||_1 = o(d_n)$. Furthermore, under the above assumptions, the functional ϕ identifying the asymptotic spectral distribution is the same for $\{X_n + Y_n\}$ and $\{X_n\}$. This result extends Theorem 1.6, where the functional ϕ identifying the asymptotic spectral distribution of both $\{X_n\}$ and $\{Y_n\}$ is given by $\phi(F) := \frac{1}{m_d(D)} \int_D F(f(\mathbf{x}))d\mathbf{x}$, to the case where the spectral distribution of $\{X_n\}$ and $\{Y_n\}$ is described by more general functionals ϕ. We mention some examples of applications, including the case of matrix-sequences with spectral distributions described by matrix-valued functions and the approximation by $\mathbb{Q}_p$ Finite Element Methods of convection-diffusion equations. The latter application will be developed in full details in Chapter 3.

2.2.1 Main results

Our main result, briefly summarized above, is Theorem 2.6. In order to prove it, we need some preliminary work. If $S \subset \mathbb{C}$ is compact and F is continuous over S, we set $\|F\|_{\infty,S} := \max_{z\in S} |F(z)|$.

Theorem 2.5. *Let $\{Z_n\}$ be a sequence of matrices, with Z_n of size d_n tending to infinity, and assume the following.*

1. *$\{Z_n\}$ is weakly clustered at a compact set $S \subset \mathbb{C}$ with $\mathbb{C}\backslash S$ connected.*

2. *$\rho(Z_n) \leq C$ for all n, with C a constant independent of n.*

3. *For some radius R and for all functions $p \in C_c(\mathbb{C})$ coinciding over $\overline{D(0,R)}$ with a complex polynomial in $\mathbb{C}[z]$, there exists $\lim_{n\to\infty} \frac{1}{d_n} \sum_{j=1}^{d_n} p(\lambda_j(Z_n)) = \phi(p)$, where $\phi : C_c(\mathbb{C}) \to \mathbb{C}$ satisfies the following 'continuity property':*

$$\forall F \in C_c(\mathbb{C}),\ \forall \epsilon > 0 \quad \exists \delta := \delta_{\epsilon,F} > 0 : \quad |\phi(F) - \phi(G)| \leq \epsilon \quad \forall G \in C_c(\mathbb{C}) \text{ with } \|F - G\|_{\infty,S} \leq \delta. \tag{2.16}$$

Then, for all $F \in C_c(\mathbb{C})$ holomorphic in the interior of S there exists

$$\lim_{n\to\infty} \frac{1}{d_n} \sum_{j=1}^{d_n} F(\lambda_j(Z_n)) = \phi(F). \tag{2.17}$$

In particular, if the interior of S is empty, (2.17) *holds for all $F \in C_c(\mathbb{C})$.*

Proof. Let $F \in C_c(\mathbb{C})$ be holomorphic in the interior of S and let $\epsilon \in (0,1)$. By the hypothesis on ϕ, there exists $\delta := \delta_{\epsilon,F} > 0$ such that $|\phi(F)-\phi(G)| \leq \epsilon$ for all $G \in C_c(\mathbb{C})$ with $\|F-G\|_{\infty,S} \leq \delta$. Without loss of generality, we may assume $\delta \leq \epsilon$. By the Mergelyan theorem [50], there exists a polynomial $q(z) := q_{\epsilon,F}(z) \in \mathbb{C}[z]$ such that $\|q - F\|_{\infty,S} \leq \delta$. Let $p := p_{\epsilon,F}$ be a function in $C_c(\mathbb{C})$ coinciding with q over $S \cup \overline{D(0,R)}$, and note that $\|p - F\|_{\infty,S} \leq \delta \leq \epsilon$ and $|\phi(p) - \phi(F)| \leq \epsilon$. Then, for all n we have

$$\left|\frac{1}{d_n}\sum_{j=1}^{d_n} F(\lambda_j(Z_n)) - \phi(F)\right| \leq \left|\frac{1}{d_n}\sum_{j=1}^{d_n} F(\lambda_j(Z_n)) - \frac{1}{d_n}\sum_{j=1}^{d_n} p(\lambda_j(Z_n))\right| + \left|\frac{1}{d_n}\sum_{j=1}^{d_n} p(\lambda_j(Z_n)) - \phi(p)\right| + |\phi(p) - \phi(F)|. \tag{2.18}$$

The second term in the right-hand side tends to 0 as $n \to \infty$ (by the third assumption), while the third term is bounded from above by ϵ. For the first term we have

$$\begin{aligned}
&\left|\frac{1}{d_n}\sum_{j=1}^{d_n} F(\lambda_j(Z_n)) - \frac{1}{d_n}\sum_{j=1}^{d_n} p(\lambda_j(Z_n))\right| \leq \frac{1}{d_n}\sum_{j=1}^{d_n} |F(\lambda_j(Z_n)) - p(\lambda_j(Z_n))| \\
&= \frac{1}{d_n} \sum_{j:\, \lambda_j(Z_n)\notin D(S,\epsilon)} |F(\lambda_j(Z_n)) - p(\lambda_j(Z_n))| + \frac{1}{d_n} \sum_{j:\, \lambda_j(Z_n)\in D(S,\epsilon)\backslash S} |F(\lambda_j(Z_n)) - p(\lambda_j(Z_n))| \\
&\quad + \frac{1}{d_n} \sum_{j:\, \lambda_j(Z_n)\in S} |F(\lambda_j(Z_n)) - p(\lambda_j(Z_n))|.
\end{aligned} \tag{2.19}$$

Now observe that the spectrum $\Lambda(Z_n)$ is contained in $\overline{D(0,C)}$ for all n, because the spectral radii $\rho(Z_n)$ are all bounded from above by C (second assumption). Moreover, by definition of $D(S,\epsilon)$ (see Section 1.1), for all n and all $j \in \{1,\dots,d_n\}$ such that $\lambda_j(Z_n) \in D(S,\epsilon)$ we can find a point $\mu_{j,n} \in S$ such that $|\lambda_j(Z_n) - \mu_{j,n}| \leq \epsilon$. Let ω_F and ω_p be the moduli of continuity of F and p over $\overline{D(S,1)}$, and note that $\overline{D(S,1)} \supseteq D(S,\epsilon)$, because we have fixed $\epsilon \in (0,1)$. Then, the three summands in (2.19) can be bounded as follows:

$$\frac{1}{d_n} \sum_{j:\, \lambda_j(Z_n)\notin D(S,\epsilon)} |F(\lambda_j(Z_n)) - p(\lambda_j(Z_n))| \leq \frac{\|F - p\|_{\infty,\overline{D(0,C)}}\ \#\{j \in \{1,\dots,d_n\} :\ \lambda_j(Z_n) \notin D(S,\epsilon)\}}{d_n} \leq c\,\frac{q_\epsilon(n,S)}{d_n},$$

with $q_\epsilon(n,S) := \#\{j \in \{1,\dots,d_n\} :\ \lambda_j(Z_n) \notin D(S,\epsilon)\}$ as in Definition 1.3 and c a constant independent of n;

$$\begin{aligned}
&\frac{1}{d_n} \sum_{j:\, \lambda_j(Z_n)\in D(S,\epsilon)\setminus S} |F(\lambda_j(Z_n)) - p(\lambda_j(Z_n))| \\
&\qquad \le \frac{1}{d_n} \sum_{j:\, \lambda_j(Z_n)\in D(S,\epsilon)\setminus S} \Big(|F(\lambda_j(Z_n)) - F(\mu_{j,n})| + |F(\mu_{j,n}) - p(\mu_{j,n})| + |p(\mu_{j,n}) - p(\lambda_j(Z_n))|\Big) \\
&\qquad \le \frac{1}{d_n} \sum_{j:\, \lambda_j(Z_n)\in D(S,\epsilon)\setminus S} \Big(\omega_F(\epsilon) + \delta + \omega_p(\epsilon)\Big) \le \omega_F(\epsilon) + \epsilon + \omega_p(\epsilon); \\
&\frac{1}{d_n} \sum_{j:\, \lambda_j(Z_n)\in S} |F(\lambda_j(Z_n)) - p(\lambda_j(Z_n))| \le \|F - p\|_{\infty,S} \le \delta \le \epsilon.
\end{aligned}$$

Passing to the limit as $n \to \infty$ in (2.18) and recalling that $\{Z_n\}$ is weakly clustered at S, we get

$$\limsup_{n\to\infty} \left| \frac{1}{d_n} \sum_{j=1}^{d_n} F(\lambda_j(Z_n)) - \phi(F) \right| \le \omega_F(\epsilon) + \epsilon + \omega_p(\epsilon) + \epsilon + \epsilon,$$

and the thesis follows from the fact that the right-hand side tends to 0 as $\epsilon \to 0$, since F, p are continuous (and hence uniformly continuous) over $\overline{D(S,1)}$. □

Remark 2.1. We note that, if $\phi : C_c(\mathbb{C}) \to \mathbb{C}$ is a functional satisfying (2.16) for a compact set S and if K is a compact set containing S, then ϕ satisfies (2.16) also for the compact set K.

Lemma 2.6. $|\mathrm{trace}(Z)| \le |||Z|||_1$ *for all square matrices* Z.

Proof. This is Weyl's majorization theorem for $p = 1$; see, e.g., [7, Theorem II.3.6, Eq. (II.23)]. For the reader's convenience, we include a short and direct proof. Let $Z \in \mathbb{C}^{m\times m}$ be a square matrix and let $Z = U\Sigma V$ be a singular value decomposition of Z. Then,

$$|\mathrm{trace}(Z)| = \left|\sum_{i=1}^m (U\Sigma V)_{ii}\right| = \left|\sum_{i=1}^m \sum_{k=1}^m u_{ik}\sigma_k(Z)v_{ki}\right| = \left|\sum_{k=1}^m \sigma_k(Z) \sum_{i=1}^m u_{ik}v_{ki}\right| \le \sum_{k=1}^m \sigma_k(Z) \sum_{i=1}^m |u_{ik}||v_{ki}| \le \sum_{k=1}^m \sigma_k(Z) = |||Z|||_1,$$

where the latter inequality follows from the Cauchy-Schwarz inequality and from the fact that the Euclidean norm of the vectors $\mathbf{u}_k := [u_{1k},\dots,u_{mk}]$ and $\mathbf{v}_k := [v_{k1},\dots,v_{km}]$ is 1 (the matrices U, V are unitary). □

Lemma 2.7. *Let* $\{Z_n\}$ *be a sequence of matrices, with* Z_n *of size* d_n *tending to infinity, and assume that* $|||\Im(Z_n)|||_1 = o(d_n)$ *and* $\Lambda(\Re(Z_n)) \subseteq [c,d]$ *for all* n*, with* c, d *independent of* n*. Then* $\{Z_n\}$ *is weakly clustered at* $[c,d]$.

Proof. The Lemma follows from [37, Corollary 3.3]. We may also derive it directly from [37, Lemma 3.2]. □

Theorem 2.6. *Let* $\{X_n\}$, $\{Y_n\}$ *be sequences of matrices with* $X_n, Y_n \in \mathbb{C}^{d_n\times d_n}$ *and* d_n *tending to infinity, and assume the following.*

1. $\|X_n\|, \|Y_n\| \le C$ *for all* n*, with* C *a constant independent of* n.
2. *Every* X_n *is Hermitian and, for some radius* R *and for all functions* $p \in C_c(\mathbb{C})$ *coinciding over* $\overline{D(0,R)}$ *with a complex polynomial in* $\mathbb{C}[z]$*, there exists* $\lim_{n\to\infty} \frac{1}{d_n} \sum_{j=1}^{d_n} p(\lambda_j(X_n)) = \phi(p)$*, where* $\phi : C_c(\mathbb{C}) \to \mathbb{C}$ *satisfies* (2.16) *for some compact set* $S \subset \mathbb{R}$.
3. $|||Y_n|||_1 = o(d_n)$ *as* $n \to \infty$.

Then, setting $Z_n := X_n + Y_n$, for every $F \in C_c(\mathbb{C})$ there exists

$$\lim_{n\to\infty} \frac{1}{d_n} \sum_{j=1}^{d_n} F(\lambda_j(Z_n)) = \phi(F).$$

Proof. Let K be a compact subset of $\mathbb{R}$ containing both S and $[-2C, 2C]$. Note that K does not disconnect $\mathbb{C}$ and has empty interior. Moreover, ϕ satisfies (2.16) for the compact set S and hence also for the compact set K, by Remark 2.1. We show that the sequence $\{Z_n\}$ satisfies the assumptions of Theorem 2.5 with K in place of S, after which the proof is finished.

1. We have $Z_n = X_n + \Re(Y_n) + \mathrm{i}\Im(Y_n)$, where $\Re(Z_n) = X_n + \Re(Y_n)$ has all the eigenvalues in $[-2C, 2C]$, because $\|\Re(Z_n)\| \le \|X_n\| + \|\Re(Y_n)\| \le \|X_n\| + \|Y_n\| \le 2C$, while $|||\Im(Z_n)|||_1 = |||\Im(Y_n)|||_1 \le |||Y_n|||_1 = o(d_n)$. Hence, $\{Z_n\}$ is weakly clustered at $[-2C, 2C]$ by Lemma 2.7, and, a fortiori, is weakly clustered at $K \supseteq [-2C, 2C]$.

2. $\rho(Z_n) \le \|Z_n\| \le \|X_n\| + \|Y_n\| \le 2C$ for all n.

3. We show that $\lim_{n\to\infty} \frac{1}{d_n} \sum_{j=1}^{d_n} p(\lambda_j(Z_n)) = \phi(p)$ for all functions $p \in C_c(\mathbb{C})$ coinciding with a polynomial over $\overline{D(0, 2C + R)}$. Note first that, for a monomial z^k, $k \ge 0$, we have

$$Z_n^k = (X_n + Y_n)^k = X_n^k + R_{n,k},$$

where $R_{n,k} := (X_n + Y_n)^k - X_n^k$ satisfies $\left|\left|\left|R_{n,k}\right|\right|\right|_1 = o(d_n)$. This follows from the third assumption, from the fact that $\|X_n\|, \|Y_n\|$ are bounded from above by a constant C independent of n, and from the Hölder-type inequality $|||XY|||_1 \le \|X\|\,|||Y|||_1$ satisfied by the trace-norm (see [7, Problem III.6.2 and Corollary IV.2.6] for the Hölder-type inequalities satisfied by the Schatten p-norms and by the unitarily invariant norms in general). Therefore, for every polynomial $q(z) := q_0 + q_1 z + \ldots + q_m z^m \in \mathbb{C}[z]$ we have $q(Z_n) = q(X_n) + R_{n,q(z)}$, where $R_{n,q(z)} := \sum_{k=0}^{m} q_k R_{n,k}$ satisfies $\left|\left|\left|R_{n,q(z)}\right|\right|\right|_1 = o(d_n)$. By Lemma 2.6 we then obtain

$$|\mathrm{trace}(q(Z_n)) - \mathrm{trace}(q(X_n))| = |\mathrm{trace}(q(Z_n) - q(X_n))| = |\mathrm{trace}(R_{n,q(z)})| \le \left|\left|\left|R_{n,q(z)}\right|\right|\right|_1 = o(d_n),$$

implying that the sequence

$$\frac{1}{d_n}\mathrm{trace}(q(Z_n)) = \frac{1}{d_n} \sum_{j=1}^{d_n} q(\lambda_j(Z_n))$$

converges to the same limit of the sequence

$$\frac{1}{d_n}\mathrm{trace}(q(X_n)) = \frac{1}{d_n} \sum_{j=1}^{d_n} q(\lambda_j(X_n))$$

(provided the latter exists). To conclude, note that $\Lambda(Z_n), \Lambda(X_n) \subseteq \overline{D(0, 2C + R)}$ for all n. This implies that, for all $p \in C_c(\mathbb{C})$ coinciding over $\overline{D(0, 2C + R)}$ with a polynomial $q_p(z) \in \mathbb{C}[z]$, we have

$$\lim_{n\to\infty} \frac{1}{d_n} \sum_{j=1}^{d_n} p(\lambda_j(Z_n)) = \lim_{n\to\infty} \frac{1}{d_n} \sum_{j=1}^{d_n} q_p(\lambda_j(Z_n)) = \lim_{n\to\infty} \frac{1}{d_n} \sum_{j=1}^{d_n} q_p(\lambda_j(X_n)) = \lim_{n\to\infty} \frac{1}{d_n} \sum_{j=1}^{d_n} p(\lambda_j(X_n)) = \phi(p),$$

where the last equality follows from the second assumption. □

2.2.2 Some applications

In this subsection, we discuss some applications of Theorem 2.6: the case of matrix-sequences with asymptotic spectral distributions described by matrix-valued functions and the approximation by $\mathbb{Q}_p$ Finite Element Methods of convection-diffusion equations.

Matrix-sequences with asymptotic spectral distributions described by matrix-valued functions

From Theorem 2.6 we obtain the following generalization of Theorem 1.6. As we shall see, this generalization serves in particular to determine the asymptotic spectral distribution of matrix-sequences of the form $\{T_{\boldsymbol{n}(n)}(\mathbf{f}) + Y_n\}_n$, where $\mathbf{f}$ is some Hermitian matrix-valued function in L^∞, $\boldsymbol{n}(n) \to \infty$ as $n \to \infty$, and $\{Y_n\}$ satisfies the assumptions of Theorem 2.7.

Theorem 2.7. *Let $\{X_n\}$, $\{Y_n\}$ be sequences of matrices with $X_n, Y_n \in \mathbb{C}^{d_n \times d_n}$ and d_n tending to infinity, and assume the following.*

1. *$\|X_n\|, \|Y_n\| \le C$ for all n, with C a constant independent of n.*
2. *Every X_n is Hermitian and $\{X_n\} \sim_\lambda \mathbf{f}$, where $\mathbf{f} : D \to \mathbb{C}^{s\times s}$ is a measurable function defined on a measurable set $D \subset \mathbb{R}^d$ with $0 < m_d(D) < \infty$.*
3. *$|||Y_n|||_1 = o(d_n)$ as $n \to \infty$.*

Then $\overline{\bigcup_n \Lambda(X_n)} \subseteq [-C, C]$, $\lambda_1(\mathbf{f}), \ldots, \lambda_s(\mathbf{f}) \in \overline{\bigcup_n \Lambda(X_n)}$ a.e., and $\{Z_n\} \sim_\lambda \mathbf{f}$, where $Z_n := X_n + Y_n$.

Proof. Let $K := \overline{\bigcup_n \Lambda(X_n)}$. Since every X_n is Hermitian with $\|X_n\| \le C$, we have $K \subseteq [-C, C]$. We show that $\lambda_1(\mathbf{f}), \ldots, \lambda_s(\mathbf{f}) \in K$ a.e. Assume by contradiction that this is not the case. Then, we can find a disk $D(z, r)$ such that $\overline{D(z, r)} \cap K$ is empty and $m_d(\{\exists j : \lambda_j(\mathbf{f}) \in D(z, r)\}) > 0$, where $\{\exists j : \lambda_j(\mathbf{f}) \in D(z, r)\} := \{\mathbf{x} \in D : \exists j \in \{1, \ldots, s\} \text{ such that } \lambda_j(\mathbf{f}(\mathbf{x})) \in D(z, r)\}$. Choose a test function $F \in C_c(\mathbb{C})$ such that $F = 1$ over $\overline{D(z, r)}$, $F = 0$ over K, and $0 \le F \le 1$ over $\mathbb{C}$. For this test function the limit relation $\lim_{n\to\infty} \frac{1}{d_n} \sum_{j=1}^{d_n} F(\lambda_j(X_n)) = \frac{1}{m_d(D)} \int_D \frac{1}{s} \sum_{j=1}^{s} F(\lambda_j(\mathbf{f}(\mathbf{x}))) \, d\mathbf{x}$ cannot hold, because the first term is 0 while the second is positive. This is a contradiction to the second hypothesis. We conclude that $\lambda_1(\mathbf{f}), \ldots, \lambda_s(\mathbf{f}) \in K$ a.e.

Now, let $\phi : C_c(\mathbb{C}) \to \mathbb{C}$ be the functional defined as

$$\phi(F) := \frac{1}{m_d(D)} \int_D \frac{1}{s} \sum_{j=1}^{s} F(\lambda_j(\mathbf{f}(\mathbf{x}))) \, d\mathbf{x}.$$

This functional satisfies (2.16) with $S = [-C, C]$, because $\lambda_1(\mathbf{f}), \ldots, \lambda_s(\mathbf{f}) \in [-C, C]$ a.e. and hence, for all $F, G \in C_c(\mathbb{C})$,

$$|\phi(F) - \phi(G)| \le \frac{1}{m_d(D)} \int_D \frac{1}{s} \sum_{j=1}^{s} \left| F(\lambda_j(\mathbf{f}(\mathbf{x}))) - G(\lambda_j(\mathbf{f}(\mathbf{x}))) \right| d\mathbf{x} \le \|F - G\|_{\infty,[-C,C]}.$$

All the hypotheses of Theorem 2.6 are then satisfied and the thesis follows. □

Corollary 2.1. *Let $\mathbf{f} : [-\pi, \pi]^d \to \mathbb{C}^{s\times s}$ be a Hermitian matrix-valued function in $L^\infty([-\pi, \pi]^d)$, let $\{\boldsymbol{n}(n)\}_n \subseteq \mathbb{N}^d$ be a sequence of multi-indices such that $\boldsymbol{n}(n) \to \infty$ as $n \to \infty$, and let $\{Y_n\}$ be a sequence of matrices such that Y_n has size $N(\boldsymbol{n}(n))s$, $\|Y_n\|$ is uniformly bounded with respect to n, and $|||Y_n|||_1 = o(N(\boldsymbol{n}(n)))$. Then $\{T_{\boldsymbol{n}(n)}(\mathbf{f}) + Y_n\}_n \sim_\lambda \mathbf{f}$.*

Proof. Defining $X_n := T_{\boldsymbol{n}(n)}(\mathbf{f})$, all the assumptions of Theorem 2.7 are met, thanks to Theorem 1.8 and to the inequality (1.37), which ensures $\|T_{\boldsymbol{n}(n)}(\mathbf{f})\|$ to be uniformly bounded with respect to n. □

Approximation by $\mathbb{Q}_p$ Finite Element Methods of a model convection-diffusion equation

Let us consider the boundary value problem

$$\begin{cases} -\Delta u + \boldsymbol{\beta} \cdot \nabla u + \gamma u = f & \text{in } \Omega, \\ u = 0 & \text{on } \partial\Omega, \end{cases} \tag{2.20}$$

where $\Omega = (0,1)^d$, $f \in L^2(\Omega)$, $\boldsymbol{\beta} := (\beta_1, \ldots, \beta_d)$, and γ, β_j, $j = 1, \ldots, d$, are functions in $L^\infty(\Omega)$ with $\gamma \geq 0$. We approximate (2.20) by using the standard $\mathbb{Q}_p$ Lagrangian Finite Element Method on the uniform mesh determined by the hypercubes whose vertices are $(j_1/n, \ldots, j_d/n)$, $j_1, \ldots, j_d = 0, \ldots, n$. We refer the reader to Chapter 3 for the details on this method. Denote by A_n the stiffness matrix, of size $(np-1)^d$, resulting from this approximation technique. It can be proved that the $(np)^d \times (np)^d$ matrix $n^{d-2}A_n \oplus O_{(np)^d-(np-1)^d}$, obtained from A_n by adding zeros, is similar, through a permutation transformation, to $T_{\boldsymbol{n}(n)}(\mathbf{f}_p) + Y_n$, where:

- $\boldsymbol{n}(n) := (n, \ldots, n)$ (d components);
- $\mathbf{f}_p : [-\pi, \pi]^d \to \mathbb{C}^{p^d \times p^d}$ is a Hermitian matrix-valued function, continuous over $[-\pi, \pi]^d$, which is also positive semidefinite over $[-\pi, \pi]^d$, because $\mathbf{v}^* \mathbf{f}_p(\mathbf{x}) \mathbf{v} \geq 0$ for all $\mathbf{v} \in \mathbb{C}^{p^d \times p^d}$ and for all $\mathbf{x} \in [-\pi, \pi]^d$;
- the matrix-sequence $\{Y_n\}$ is real and non-symmetric (due to the presence of the convection term), but satisfies the assumptions that the trace-norm is $o((np)^d)$ when $n \to \infty$ and that the spectral norm $\|Y_n\|$ is uniformly bounded with respect to n.

Therefore, by Corollary 2.1 we have $\{n^{d-2}A_n \oplus O_{(np)^d-(np-1)^d}\} \sim_\lambda \mathbf{f}_p$. This implies that $\{n^{d-2}A_n\} \sim_\lambda \mathbf{f}_p$ by Definition 1.1, because the eigenvalues of $n^{d-2}A_n \oplus O_{(np)^d-(np-1)^d}$ are precisely those of $n^{d-2}A_n$, with only $(np)^d - (np-1)^d = o((np)^d)$ extra eigenvalues equal to 0.

A detailed spectral analysis of the stiffness matrices coming from the $\mathbb{Q}_p$ Lagrangian Finite Element approximation of classical convection-diffusion equations like (2.20), including the formal proof of the results cited above and the study of the properties of the matrix-valued function $\mathbf{f}_p$, will be the subject of Chapter 3. Here, we have just mentioned the application of the theoretical tools obtained in this section for determining the asymptotic spectral distribution of such matrices.

Chapter 3

Spectral analysis and spectral symbol of $\mathbb{Q}_{\boldsymbol{p}}$ Lagrangian FEM stiffness matrices

This chapter is devoted to the (asymptotic) spectral analysis of the stiffness matrices arising from the $\mathbb{Q}_{\boldsymbol{p}}$ Lagrangian Finite Element approximation of the following second-order d-dimensional elliptic differential problem:

$$\begin{cases} -\Delta u + \boldsymbol{\beta} \cdot \nabla u + \gamma u = f & \text{in } \Omega := (0,1)^d, \\ u = 0 & \text{on } \partial\Omega, \end{cases} \tag{3.1}$$

where $f \in L^2(\Omega)$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_d)$, and γ, β_j, $j = 1, \ldots, d$, are functions in $L^\infty(\Omega)$ with $\gamma \geq 0$ over Ω. The multi-index $\boldsymbol{p} := (p_1, \ldots, p_d) \in \mathbb{N}^d$, which appears as subscript of $\mathbb{Q}_{\boldsymbol{p}}$, is related to the Finite Element approximation order and, more specifically, p_j is the polynomial approximation degree in the j-th direction. We will provide in Section 3.1 all the necessary details for understanding the $\mathbb{Q}_{\boldsymbol{p}}$ Lagrangian Finite Element Method (FEM), but we also refer the reader to [47, 48, 49, 18, 55, 14, 15] for a wide account on this numerical technique and its evolution.

After presenting a construction of the $\mathbb{Q}_{\boldsymbol{p}}$ FEM stiffness matrices, we investigate the behavior of the extremal eigenvalues, the conditioning, and the asymptotic spectral distribution when the mesh is refined and the matrix size goes to infinity; in particular, we find out the associated spectral symbol (see Definition 1.1). We also study the properties of the symbol, which turns out to be a d-variate function taking values in the space of $N(\boldsymbol{p}) \times N(\boldsymbol{p})$ Hermitian matrices. Looking at Remark 1.2, this means that the spectrum of our FEM matrices is (asymptotically) described by $N(\boldsymbol{p})$ different functions, that is the $N(\boldsymbol{p})$ eigenvalues of the symbol, which give rise to $N(\boldsymbol{p})$ different 'spectral branches'. Unfortunately, as we shall see in Subsection 3.4.3, the eigenvalues of the symbol are well-separated, far away, and exponentially diverging with respect to $\boldsymbol{p}$ and d, implying that the eigenvalues of the FEM matrices behave in the same way. Even in the case $d = 1$ and $p = 3$, we see from Figure 3.2 that the maximum eigenvalue of the symbol is rather distant from the minimum eigenvalue. This very involved picture provides an explanation of:

(a) the difficulties encountered in designing robust solvers for the $\mathbb{Q}_{\boldsymbol{p}}$ FEM stiffness matrices, with convergence speed independent of the matrix size, of the approximation parameters $\boldsymbol{p}$, and of the dimensionality d;

(b) the possible convergence deterioration of known iterative methods, already for moderate $\boldsymbol{p}$ and d.

3.1 Galerkin method and $\mathbb{Q}_{\boldsymbol{p}}$ Lagrangian FEM

The weak (variational) form of the elliptic differential equation (3.1) can be stated as follows: find $u \in H_0^1(\Omega)$ such that

$$a(u, v) = \langle f, v \rangle, \quad \forall v \in H_0^1(\Omega), \tag{3.2}$$

where $a(u, v) := \int_\Omega (\nabla u \cdot \nabla v + \boldsymbol{\beta} \cdot \nabla u \, v + \gamma u v)$ and $\langle f, v \rangle := \int_\Omega f v$.

In the standard Galerkin approach, we find an approximation of u by choosing a finite dimensional subspace $W \subset H_0^1(\Omega)$, called the approximation space, and by solving the following (Galerkin) problem: find $u_W \in W$ such that

$$a(u_W, v) = \langle f, v \rangle, \quad \forall v \in W.^1 \tag{3.3}$$

If $\dim W = N$ and we fix a basis $\{\varphi_1, \dots, \varphi_N\}$ for W, then we can expand every function $v \in W$ as a linear combination of the form $v = \sum_{j=1}^N v_j \varphi_j$, and the computation of $u_W = \sum_{j=1}^N u_j \varphi_j$ is reduced to solving the linear system

$$A\mathbf{u} = \mathbf{f}, \tag{3.4}$$

where $A = [a(\varphi_j, \varphi_i)]_{i,j=1}^N$ is the stiffness matrix and $\mathbf{f} = [\langle f, \varphi_i \rangle]_{i=1}^N$. Once we find $\mathbf{u}$, we know $u_W = \sum_{j=1}^N u_j \varphi_j$.

In the context of $\mathbb{Q}_p$ Lagrangian FEM, W is chosen as a space of continuous piecewise polynomial functions vanishing on the boundary of Ω. More precisely, define for $p, n \geq 1$ the spaces

$$\begin{aligned} V_n^{(p)} &:= \left\{ s \in C([0,1]) : \ s_{|[\frac{i}{n}, \frac{i+1}{n})} \in \mathbb{P}_p \quad \forall i = 0, \dots, n-1 \right\}, \\ W_n^{(p)} &:= \{ s \in V_n^{(p)} : \ s(0) = s(1) = 0 \} \subset H_0^1(0,1). \end{aligned}$$

It is known that $\dim V_n^{(p)} = np + 1$ and $\dim W_n^{(p)} = np - 1$. Consider for $V_n^{(p)}$ the Lagrangian basis $\{\ell_{0,(p)}, \dots, \ell_{np,(p)}\}$ on the uniform knot sequence $\xi_i = \frac{i}{np}$, $i = 0, \dots, np$. This means that $\ell_{j,(p)}$ is the unique function in $V_n^{(p)}$ taking the value 1 at ξ_j and 0 at ξ_i for $i \neq j$:

$$\ell_{j,(p)}(\xi_i) = \delta_{ij}, \quad \forall i, j = 0, \dots, np.$$

Since $\ell_{1,(p)}, \dots, \ell_{np-1,(p)}$ vanish at the boundary of $[0,1]$, we infer that $\{\ell_{1,(p)}, \dots, \ell_{np-1,(p)}\}$ is a basis for $W_n^{(p)}$ (the Lagrangian basis of $W_n^{(p)}$). For later purposes, we report the explicit expressions of the basis functions $\ell_{1,(p)}, \dots, \ell_{np-1,(p)}$ and of their (Sobolev) derivatives in terms of the Lagrange polynomials $L_0, \dots, L_p$ associated with the knots $t_k = \frac{k}{p}$, $k = 0, \dots, p$, which are given by

$$L_h(t) = \prod_{\substack{k=0 \\ k \neq h}}^p \frac{t - t_k}{t_h - t_k} = \prod_{\substack{k=0 \\ k \neq h}}^p \frac{pt - k}{h - k}, \quad \forall h = 0, \dots, p, \qquad L_h(t_k) = \delta_{hk}, \quad \forall h, k = 0, \dots, p. \tag{3.5}$$

If j is a multiple of p, then the support of $\ell_{j,(p)}$ is $\operatorname{supp}(\ell_{j,(p)}) = [\xi_{j-p}, \xi_{j+p}]$,

$$\ell_{j,(p)}(x) = \begin{cases} L_p\left(\dfrac{x - \xi_{j-p}}{\xi_j - \xi_{j-p}}\right) & \xi_{j-p} \leq x \leq \xi_j, \\ L_0\left(\dfrac{x - \xi_j}{\xi_{j+p} - \xi_j}\right) & \xi_j \leq x \leq \xi_{j+p}, \\ 0 & \text{otherwise,} \end{cases} = \begin{cases} L_p(nx - n\xi_{j-p}) & \xi_{j-p} \leq x \leq \xi_j, \\ L_0(nx - n\xi_j) & \xi_j \leq x \leq \xi_{j+p}, \\ 0 & \text{otherwise,} \end{cases} \tag{3.6}$$

and the derivative of $\ell_{j,(p)}$ is

$$\ell'_{j,(p)}(x) = \begin{cases} nL'_p(nx - n\xi_{j-p}) & \xi_{j-p} < x < \xi_j, \\ nL'_0(nx - n\xi_j) & \xi_j < x < \xi_{j+p}, \\ 0 & \text{otherwise.} \end{cases} \tag{3.7}$$

[1] In the case where the bilinear form $a(u, v)$ is coercive, both the solution u_W of (3.3) and the solution u of (3.2) are unique; see [13]. In particular, they are unique if $\boldsymbol{\beta}$ is constant, because in this case $a(u, v)$ is coercive; see [47, Chapter 5, p. 140].

If j is not a multiple of p, let $j_p = j \bmod p \in \{1, \dots, p-1\}$. Then $\mathrm{supp}(\ell_{j,(p)}) = [\xi_{j-j_p}, \xi_{j-j_p+p}]$,

$$\ell_{j,(p)}(x) = \begin{cases} L_{j_p}\left(\dfrac{x - \xi_{j-j_p}}{\xi_{j-j_p+p} - \xi_{j-j_p}}\right) & \xi_{j-j_p} \le x \le \xi_{j-j_p+p}, \\ 0 & \text{otherwise,} \end{cases} = \begin{cases} L_{j_p}(nx - n\xi_{j-j_p}) & \xi_{j-j_p} \le x \le \xi_{j-j_p+p}, \\ 0 & \text{otherwise,} \end{cases} \tag{3.8}$$

and the derivative of $\ell_{j,(p)}$ is

$$\ell'_{j,(p)}(x) = \begin{cases} nL'_{j_p}(nx - n\xi_{j-j_p}) & \xi_{j-j_p} < x < \xi_{j-j_p+p}, \\ 0 & \text{otherwise.} \end{cases} \tag{3.9}$$

Figure 3.1 reports the graph of $\ell_{1,(p)}, \dots, \ell_{np-1,(p)}$ in the case $p = 2$, $n = 3$, together with the graph of the Lagrange polynomials $L_0, \dots, L_p$ in (3.5) for $p = 2$. Now, for any pair of multi-indices $\boldsymbol{p}, \boldsymbol{n} \in \mathbb{N}^d$, let

$$W_{\boldsymbol{n}}^{(\boldsymbol{p})} := W_{n_1}^{(p_1)} \otimes \cdots \otimes W_{n_d}^{(p_d)} := \mathrm{span}\left(\ell_{\boldsymbol{j},(\boldsymbol{p})} : \boldsymbol{j} = \boldsymbol{1}, \dots, \boldsymbol{np} - \boldsymbol{1}\right) \subset H_0^1(\Omega), \tag{3.10}$$

where $\ell_{\boldsymbol{j},(\boldsymbol{p})} := \ell_{j_1,(p_1)} \otimes \cdots \otimes \ell_{j_d,(p_d)}$.

In the framework of $\mathbb{Q}_{\boldsymbol{p}}$ Lagrangian FEM, the subspace W in the Galerkin problem (3.3) is chosen as $W_{\boldsymbol{n}}^{(\boldsymbol{p})}$ for some $\boldsymbol{p}, \boldsymbol{n} \in \mathbb{N}^d$ (usually $\boldsymbol{p} = (p, \dots, p)$ for some $p \ge 1$), and for $W_{\boldsymbol{n}}^{(\boldsymbol{p})}$ we choose the tensor Lagrangian basis in (3.10), ordered according to the standard lexicographic ordering (1.1) for the multi-index range $\boldsymbol{1}, \dots, \boldsymbol{np} - \boldsymbol{1}$. With these choices, we obtain in (3.4) a stiffness matrix A, which henceforth will be denoted by $A_{\boldsymbol{n}}^{(\boldsymbol{p})}$ in order to emphasize its dependence on $\boldsymbol{p}$ and $\boldsymbol{n}$:

$$A_{\boldsymbol{n}}^{(\boldsymbol{p})} := \left[a(\ell_{\boldsymbol{j},(\boldsymbol{p})}, \ell_{\boldsymbol{i},(\boldsymbol{p})})\right]_{\boldsymbol{i},\boldsymbol{j}=\boldsymbol{1}}^{\boldsymbol{np}-\boldsymbol{1}}. \tag{3.11}$$

Let us consider the following split of the matrix, according to the diffusion, advection and reaction terms, respectively:

$$A_{\boldsymbol{n}}^{(\boldsymbol{p})} = \left[\int_\Omega \nabla \ell_{\boldsymbol{j},(\boldsymbol{p})} \cdot \nabla \ell_{\boldsymbol{i},(\boldsymbol{p})}\right]_{\boldsymbol{i},\boldsymbol{j}=\boldsymbol{1}}^{\boldsymbol{np}-\boldsymbol{1}} + \left[\int_\Omega \boldsymbol{\beta} \cdot \nabla \ell_{\boldsymbol{j},(\boldsymbol{p})}\, \ell_{\boldsymbol{i},(\boldsymbol{p})}\right]_{\boldsymbol{i},\boldsymbol{j}=\boldsymbol{1}}^{\boldsymbol{np}-\boldsymbol{1}} + \left[\int_\Omega \gamma\, \ell_{\boldsymbol{j},(\boldsymbol{p})} \ell_{\boldsymbol{i},(\boldsymbol{p})}\right]_{\boldsymbol{i},\boldsymbol{j}=\boldsymbol{1}}^{\boldsymbol{np}-\boldsymbol{1}}. \tag{3.12}$$

For obvious reasons, the first matrix in the right-hand side of (3.12) is called diffusion matrix, the second advection matrix, and the third reaction matrix. With expressive notation, we denote these three matrices by $A_{\boldsymbol{n},D}^{(\boldsymbol{p})}, A_{\boldsymbol{n},A}^{(\boldsymbol{p})}, A_{\boldsymbol{n},R}^{(\boldsymbol{p})}$, respectively:

$$A_{\boldsymbol{n},D}^{(\boldsymbol{p})} := \left[\int_\Omega \nabla \ell_{\boldsymbol{j},(\boldsymbol{p})} \cdot \nabla \ell_{\boldsymbol{i},(\boldsymbol{p})}\right]_{\boldsymbol{i},\boldsymbol{j}=\boldsymbol{1}}^{\boldsymbol{np}-\boldsymbol{1}}, \quad A_{\boldsymbol{n},A}^{(\boldsymbol{p})} := \left[\int_\Omega \boldsymbol{\beta} \cdot \nabla \ell_{\boldsymbol{j},(\boldsymbol{p})}\, \ell_{\boldsymbol{i},(\boldsymbol{p})}\right]_{\boldsymbol{i},\boldsymbol{j}=\boldsymbol{1}}^{\boldsymbol{np}-\boldsymbol{1}}, \quad A_{\boldsymbol{n},R}^{(\boldsymbol{p})} := \left[\int_\Omega \gamma\, \ell_{\boldsymbol{j},(\boldsymbol{p})} \ell_{\boldsymbol{i},(\boldsymbol{p})}\right]_{\boldsymbol{i},\boldsymbol{j}=\boldsymbol{1}}^{\boldsymbol{np}-\boldsymbol{1}}. \tag{3.13}$$

The diffusion matrix is SPD, the reaction matrix is SPSD (SPD if $\gamma \neq 0$ a.e.), while the advection matrix is not symmetric and is responsible for the non-symmetry of $A_{\boldsymbol{n}}^{(\boldsymbol{p})}$. The following lemma provides an upper bound for the spectral norm $\|A_{\boldsymbol{n},A}^{(\boldsymbol{p})}\|$. In all this chapter, the symbol γ_* will denote a nonnegative constant such that $\gamma \ge \gamma_*$ a.e. on Ω. Moreover, $\|\boldsymbol{\beta}\|_{L^\infty(\Omega)} := \max_{j=1,\dots,d} \|\beta_j\|_{L^\infty(\Omega)}$.

Lemma 3.1. *Let $\boldsymbol{p} \in \mathbb{N}^d$, then there is a constant $B_{\boldsymbol{p}}$, depending only on $\boldsymbol{p}$, such that*

$$\|A_{\boldsymbol{n},A}^{(\boldsymbol{p})}\| \le B_{\boldsymbol{p}} \|\boldsymbol{\beta}\|_{L^\infty(\Omega)} \frac{\sum_{k=1}^d n_k}{n_1 \cdots n_d}, \quad \forall \boldsymbol{n} \in \mathbb{N}^d. \tag{3.14}$$

Proof. By (1.3) we have $\|A_{\boldsymbol{n},A}^{(\boldsymbol{p})}\| \le \sqrt{\|A_{\boldsymbol{n},A}^{(\boldsymbol{p})}\|_\infty \|A_{\boldsymbol{n},A}^{(\boldsymbol{p})}\|_1}$. Recalling that the ∞-norm of a matrix is the maximum 1-norm of its row vectors and that the 1-norm of a matrix is the maximum 1-norm of its column vectors, if we show that

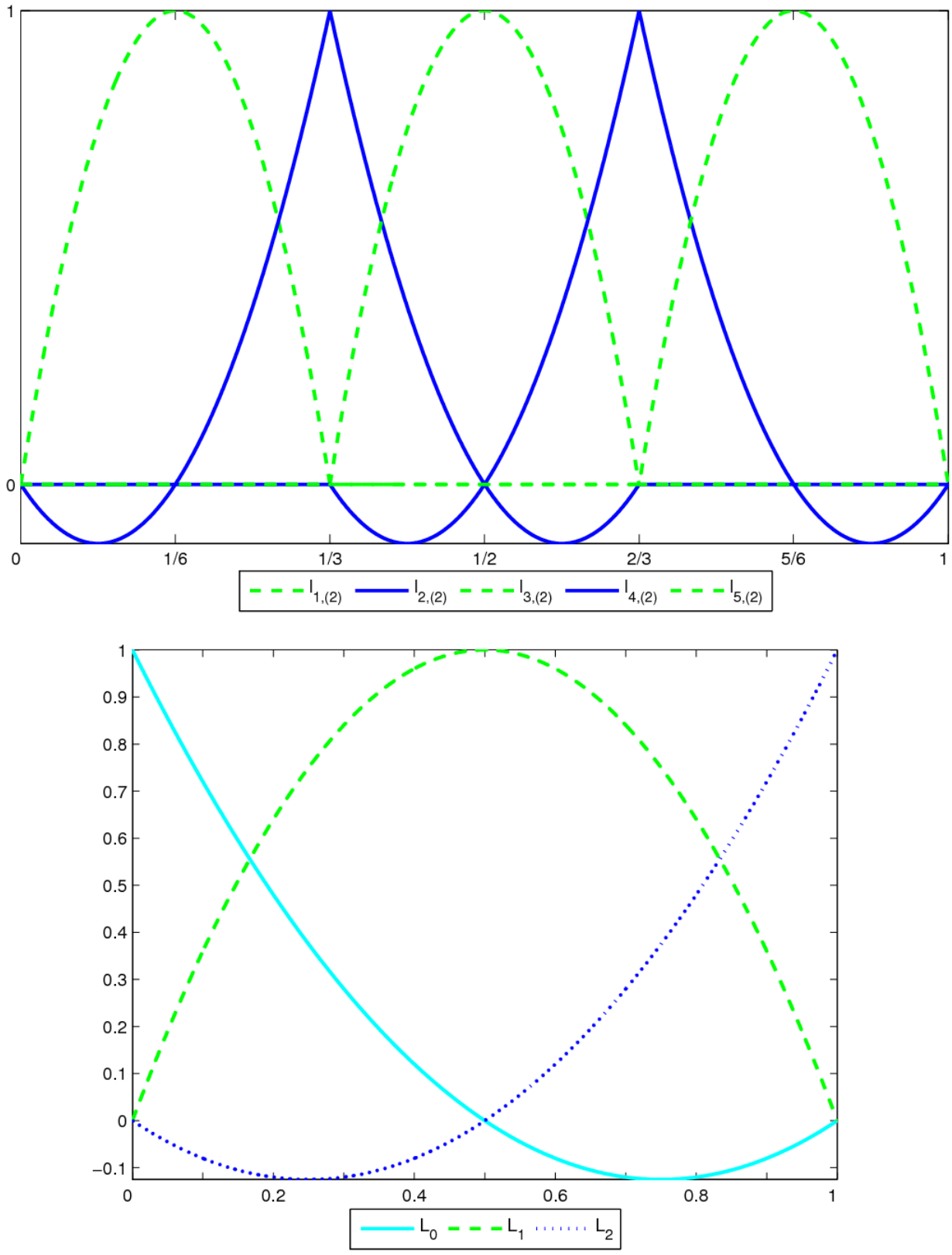


Figure 3.1: graph of the Lagrangian basis functions $\ell_{1,(p)}, \ldots, \ell_{np-1,(p)}$ in the case $p = 2$, $n = 3$, and of the Lagrange polynomials $L_0, \ldots, L_p$ in (3.5) for $p = 2$.

(a) each entry of $A^{(\boldsymbol{p})}_{\boldsymbol{n},A}$ is bounded from above by $\tilde{B}_{\boldsymbol{p}}\|\boldsymbol{\beta}\|_{L^\infty(\Omega)}\frac{\sum_{k=1}^d n_k}{n_1\cdots n_d}$ for some constant $\tilde{B}_{\boldsymbol{p}}$ depending only on $\boldsymbol{p}$,

(b) each row and column of $A^{(\boldsymbol{p})}_{\boldsymbol{n},A}$ contains a number of nonzero entries bounded from above by some constant $\hat{B}_{\boldsymbol{p}}$ depending only on $\boldsymbol{p}$,

then the thesis follows with $B_{\boldsymbol{p}} = \hat{B}_{\boldsymbol{p}}\tilde{B}_{\boldsymbol{p}}$.

For all $p \geq 1$, set $U_p := \max\{\|L_j\|_{L^\infty(0,1)}, \|L'_j\|_{L^\infty(0,1)} : j = 0,\dots,p\}$, where $L_0,\dots,L_p$ are the Lagrange polynomials (3.5). From the expressions of $\ell_{1,(p)},\dots,\ell_{np-1,(p)}$ and of their derivatives given in (3.6)–(3.9), and taking into account the supports of $\ell_{1,(p)},\dots,\ell_{np-1,(p)}$, for all $p, n \geq 1$ and for all $i, j = 1,\dots,np-1$ we have

$$\int_{(0,1)} |\ell_{j,(p)}||\ell_{i,(p)}| \leq \begin{cases} 2U_p^2/n & \text{if } |i-j| \leq p, \\ 0 & \text{otherwise,} \end{cases} \qquad \int_{(0,1)} |\ell'_{j,(p)}||\ell_{i,(p)}| \leq \begin{cases} 2U_p^2 & \text{if } |i-j| \leq p, \\ 0 & \text{otherwise.} \end{cases}$$

Now, for $\boldsymbol{p}, \boldsymbol{n} \in \mathbb{N}^d$, for $\boldsymbol{i}, \boldsymbol{j} = \boldsymbol{1},\dots,\boldsymbol{np}-\boldsymbol{1}$ and for $k = 1,\dots,d$, since $\ell_{\boldsymbol{j},(\boldsymbol{p})} = \ell_{j_1,(p_1)} \otimes \cdots \otimes \ell_{j_d,(p_d)}$ and $\Omega = (0,1)^d$ is rectangular, we have

$$\begin{aligned}
\frac{\partial \ell_{\boldsymbol{j},(\boldsymbol{p})}}{\partial x_k} &= \ell_{j_1,(p_1)} \otimes \cdots \otimes \ell_{j_{k-1},(p_{k-1})} \otimes \ell'_{j_k,(p_k)} \otimes \ell_{j_{k+1},(p_{k+1})} \otimes \cdots \otimes \ell_{j_d,(p_d)}, \\
\int_\Omega \left|\frac{\partial \ell_{\boldsymbol{j},(\boldsymbol{p})}}{\partial x_k}\right| |\ell_{\boldsymbol{i},(\boldsymbol{p})}| &= \int_{(0,1)} |\ell_{j_1,(p_1)}||\ell_{i_1,(p_1)}| \cdots \int_{(0,1)} |\ell_{j_{k-1},(p_{k-1})}||\ell_{i_{k-1},(p_{k-1})}| \int_{(0,1)} |\ell'_{j_k,(p_k)}||\ell_{i_k,(p_k)}| \\
&\quad \cdot \int_{(0,1)} |\ell_{j_{k+1},(p_{k+1})}||\ell_{i_{k+1},(p_{k+1})}| \cdots \int_{(0,1)} |\ell_{j_d,(p_d)}||\ell_{i_d,(p_d)}| \\
&\leq \begin{cases} \frac{2U_{p_1}^2}{n_1} \cdots \frac{2U_{p_{k-1}}^2}{n_{k-1}} \cdot 2U_{p_k}^2 \cdot \frac{2U_{p_{k+1}}^2}{n_{k+1}} \cdots \frac{2U_{p_d}^2}{n_d} & \text{if } |i_1 - j_1| \leq p_1, \dots, |i_d - j_d| \leq p_d \\ 0 & \text{otherwise} \end{cases} \\
&\leq \begin{cases} U_{\boldsymbol{p}} \frac{n_k}{n_1\cdots n_d} & \text{if } \|\boldsymbol{i}-\boldsymbol{j}\|_\infty \leq \|\boldsymbol{p}\|_\infty \\ 0 & \text{otherwise} \end{cases}
\end{aligned}$$

where $U_{\boldsymbol{p}} := 2^d \prod_{i=1}^d U_{p_i}^2$. Hence,

$$\begin{aligned}
\left|[A^{(\boldsymbol{p})}_{\boldsymbol{n},A}]_{\boldsymbol{i},\boldsymbol{j}}\right| &= \left|\int_\Omega \boldsymbol{\beta} \cdot \nabla \ell_{\boldsymbol{j},(\boldsymbol{p})}\, \ell_{\boldsymbol{i},(\boldsymbol{p})}\right| \leq \int_\Omega \left|\boldsymbol{\beta} \cdot \nabla \ell_{\boldsymbol{j},(\boldsymbol{p})}\, \ell_{\boldsymbol{i},(\boldsymbol{p})}\right| \leq \int_\Omega \|\boldsymbol{\beta}\|_{L^\infty(\Omega)} \sum_{k=1}^d \left|\frac{\partial \ell_{\boldsymbol{j},(\boldsymbol{p})}}{\partial x_k}\right| |\ell_{\boldsymbol{i},(\boldsymbol{p})}| \\
&= \|\boldsymbol{\beta}\|_{L^\infty(\Omega)} \sum_{k=1}^d \int_\Omega \left|\frac{\partial \ell_{\boldsymbol{j},(\boldsymbol{p})}}{\partial x_k}\right| |\ell_{\boldsymbol{i},(\boldsymbol{p})}| \leq \begin{cases} \tilde{B}_{\boldsymbol{p}} \|\boldsymbol{\beta}\|_{L^\infty(\Omega)} \frac{\sum_{k=1}^d n_k}{n_1\cdots n_d} & \text{if } \|\boldsymbol{i}-\boldsymbol{j}\|_\infty \leq \|\boldsymbol{p}\|_\infty \\ 0 & \text{otherwise} \end{cases}
\end{aligned}$$

where $\tilde{B}_{\boldsymbol{p}} := U_{\boldsymbol{p}}$. This implies that, for a fixed $\boldsymbol{i} \in \{\boldsymbol{1},\dots,\boldsymbol{np}-\boldsymbol{1}\}$, the $\boldsymbol{i}$-th row of the matrix $A^{(\boldsymbol{p})}_{\boldsymbol{n},A}$ contains at most $\hat{B}_{\boldsymbol{p}} := \prod_{i=1}^d (2p_i+1)$ nonzero entries (those corresponding to the column multi-indices $\boldsymbol{j}$ such that $\|\boldsymbol{i}-\boldsymbol{j}\|_\infty \leq \|\boldsymbol{p}\|_\infty$), and every nonzero entry is bounded from above by $\tilde{B}_{\boldsymbol{p}}\|\boldsymbol{\beta}\|_{L^\infty(\Omega)}\frac{\sum_{k=1}^d n_k}{n_1\cdots n_d}$. Similarly, for a fixed $\boldsymbol{j} \in \{\boldsymbol{1},\dots,\boldsymbol{np}-\boldsymbol{1}\}$, the $\boldsymbol{j}$-th column of $A^{(\boldsymbol{p})}_{\boldsymbol{n},A}$ contains at most $\hat{B}_{\boldsymbol{p}}$ nonzero entries (those corresponding to the row multi-indices $\boldsymbol{i}$ such that $\|\boldsymbol{i}-\boldsymbol{j}\|_\infty \leq \|\boldsymbol{p}\|_\infty$), and every nonzero entry is bounded from above by $\tilde{B}_{\boldsymbol{p}}\|\boldsymbol{\beta}\|_{L^\infty(\Omega)}\frac{\sum_{k=1}^d n_k}{n_1\cdots n_d}$. The conditions (a) and (b) are then satisfied and the thesis follows. □

Now we introduce the mass matrix

$$N^{(\boldsymbol{p})}_{\boldsymbol{n}} := \left[\int_\Omega \ell_{\boldsymbol{j},(\boldsymbol{p})} \ell_{\boldsymbol{i},(\boldsymbol{p})}\right]_{\boldsymbol{i},\boldsymbol{j}=\boldsymbol{1}}^{\boldsymbol{np}-\boldsymbol{1}}.$$

This matrix is of interest because

$$\gamma_* N_{\boldsymbol{n}}^{(\boldsymbol{p})} \leq A_{\boldsymbol{n},R}^{(\boldsymbol{p})} \leq \|\gamma\|_{L^\infty(\Omega)} N_{\boldsymbol{n}}^{(\boldsymbol{p})}. \tag{3.15}$$

Since all the matrices in (3.15) are SPSD, their spectral norm equals their maximal eigenvalue. Therefore

$$\gamma_* \|N_{\boldsymbol{n}}^{(\boldsymbol{p})}\| \leq \|A_{\boldsymbol{n},R}^{(\boldsymbol{p})}\| \leq \|\gamma\|_{L^\infty(\Omega)} \|N_{\boldsymbol{n}}^{(\boldsymbol{p})}\|. \tag{3.16}$$

3.2 Construction of the $\mathbb{Q}_{\boldsymbol{p}}$ Lagrangian FEM stiffness matrices $A_{\boldsymbol{n}}^{(\boldsymbol{p})}$

Taking into account the tensor structure of the $\mathbb{Q}_{\boldsymbol{p}}$ Lagrangian basis $\{\ell_{\boldsymbol{j},(\boldsymbol{p})} : \boldsymbol{j} = \boldsymbol{1}, \dots, \boldsymbol{np} - \boldsymbol{1}\}$ and the rectangularity of the domain Ω, we now prove the following result, which highlights the tensor structure of the $\mathbb{Q}_{\boldsymbol{p}}$ Lagrangian FEM diffusion and mass matrices.

Theorem 3.1. *Let $\boldsymbol{p}, \boldsymbol{n} \in \mathbb{N}^d$, then*

$$A_{\boldsymbol{n},D}^{(\boldsymbol{p})} = \sum_{k=1}^{d} \frac{1}{n_1} M_{n_1}^{(p_1)} \otimes \cdots \otimes \frac{1}{n_{k-1}} M_{n_{k-1}}^{(p_{k-1})} \otimes n_k K_{n_k}^{(p_k)} \otimes \frac{1}{n_{k+1}} M_{n_{k+1}}^{(p_{k+1})} \otimes \cdots \otimes \frac{1}{n_d} M_{n_d}^{(p_d)}, \tag{3.17}$$

$$N_{\boldsymbol{n}}^{(\boldsymbol{p})} = \frac{1}{n_1} M_{n_1}^{(p_1)} \otimes \cdots \otimes \frac{1}{n_d} M_{n_d}^{(p_d)}, \tag{3.18}$$

where, for $p, n \geq 1$, $K_n^{(p)}$ and $M_n^{(p)}$ are the SPD matrices given by

$$nK_n^{(p)} := \left[\int_{(0,1)} \ell'_{j,(p)} \ell'_{i,(p)} \right]_{i,j=1}^{np-1}, \qquad \frac{1}{n} M_n^{(p)} := \left[\int_{(0,1)} \ell_{j,(p)} \ell_{i,(p)} \right]_{i,j=1}^{np-1}. \tag{3.19}$$

Proof. The proof is very simple if we use the multi-index language and, especially, the fundamental property (1.12). We could say that this proof is an exemplification of the power of the multi-index notation over the conventional linear indexing whenever one has to deal with matrices formed by a sum of tensor products, like $A_{\boldsymbol{n},D}^{(\boldsymbol{p})}$ and $N_{\boldsymbol{n}}^{(\boldsymbol{p})}$. We only prove (3.17), because (3.18) is proved in the same way. For all $\boldsymbol{i}, \boldsymbol{j} = \boldsymbol{1}, \dots, \boldsymbol{np} - \boldsymbol{1}$, we have

$$\begin{aligned}
(A_{\boldsymbol{n},D}^{(\boldsymbol{p})})_{\boldsymbol{ij}} &= \int_\Omega \nabla \ell_{\boldsymbol{j},(\boldsymbol{p})} \cdot \nabla \ell_{\boldsymbol{i},(\boldsymbol{p})} \\
&= \int_{(0,1)^d} \sum_{k=1}^{d} \ell_{j_1,(p_1)}(x_1) \ell_{i_1,(p_1)}(x_1) \cdots \ell_{j_{k-1},(p_{k-1})}(x_{k-1}) \ell_{i_{k-1},(p_{k-1})}(x_{k-1}) \cdot \ell'_{j_k,(p_k)}(x_k) \ell'_{i_k,(p_k)}(x_k) \\
&\qquad \cdot \ell_{j_{k+1},(p_{k+1})}(x_{k+1}) \ell_{i_{k+1},(p_{k+1})}(x_{k+1}) \cdots \ell_{j_d,(p_d)}(x_d) \ell_{i_d,(p_d)}(x_d) dx_1 \cdots dx_d \\
&= \sum_{k=1}^{d} \int_{(0,1)} \ell_{j_1,(p_1)}(x_1) \ell_{i_1,(p_1)}(x_1) dx_1 \cdots \int_{(0,1)} \ell_{j_{k-1},(p_{k-1})}(x_{k-1}) \ell_{i_{k-1},(p_{k-1})}(x_{k-1}) dx_{k-1} \cdot \int_{(0,1)} \ell'_{j_k,(p_k)}(x_k) \ell'_{i_k,(p_k)}(x_k) dx_k \\
&\qquad \cdot \int_{(0,1)} \ell_{j_{k+1},(p_{k+1})}(x_{k+1}) \ell_{i_{k+1},(p_{k+1})}(x_{k+1}) dx_{k+1} \cdots \int_{(0,1)} \ell_{j_d,(p_d)}(x_d) \ell_{i_d,(p_d)}(x_d) dx_d \\
&= \sum_{k=1}^{d} \left[\frac{1}{n_1} M_{n_1}^{(p_1)} \otimes \cdots \otimes \frac{1}{n_{k-1}} M_{n_{k-1}}^{(p_{k-1})} \otimes n_k K_{n_k}^{(p_k)} \otimes \frac{1}{n_{k+1}} M_{n_{k+1}}^{(p_{k+1})} \otimes \cdots \otimes \frac{1}{n_d} M_{n_d}^{(p_d)} \right]_{\boldsymbol{ij}} \\
&= \left[\sum_{k=1}^{d} \frac{1}{n_1} M_{n_1}^{(p_1)} \otimes \cdots \otimes \frac{1}{n_{k-1}} M_{n_{k-1}}^{(p_{k-1})} \otimes n_k K_{n_k}^{(p_k)} \otimes \frac{1}{n_{k+1}} M_{n_{k+1}}^{(p_{k+1})} \otimes \cdots \otimes \frac{1}{n_d} M_{n_d}^{(p_d)} \right]_{\boldsymbol{ij}},
\end{aligned}$$

where the fourth equality holds by (1.12) and by definition of $nK_n^{(p)}$ and $\frac{1}{n} M_n^{(p)}$. □

3.2.1 Construction of $K_n^{(p)}$, $M_n^{(p)}$

This subsection is devoted to the proof of the following theorem. From now on, until the end of this chapter, the symbol $\langle\,,\rangle$ will be used to denote the scalar product in $L^2(0,1)$, i.e. $\langle\varphi,\psi\rangle := \int_{(0,1)}\varphi\psi$ for all $\varphi,\psi\in L^2(0,1)$.

Theorem 3.2. *Let $p,n\geq 1$. Then*

$$K_n^{(p)} = \begin{bmatrix} K_0 & K_1^T & & \\ K_1 & \ddots & \ddots & \\ & \ddots & \ddots & K_1^T \\ & & K_1 & K_0 \end{bmatrix}_-, \qquad M_n^{(p)} = \begin{bmatrix} M_0 & M_1^T & & \\ M_1 & \ddots & \ddots & \\ & \ddots & \ddots & M_1^T \\ & & M_1 & M_0 \end{bmatrix}_-, \tag{3.20}$$

where the subscripts '$-$' mean that the last row and column of the matrices in square brackets are deleted, while K_0, K_1, M_0, M_1 are $p\times p$ blocks given by

$$K_0 = \left[\begin{array}{ccc|c} \langle L_1',L_1'\rangle & \cdots & \langle L_{p-1}',L_1'\rangle & \langle L_p',L_1'\rangle \\ \vdots & & \vdots & \vdots \\ \langle L_1',L_{p-1}'\rangle & \cdots & \langle L_{p-1}',L_{p-1}'\rangle & \langle L_p',L_{p-1}'\rangle \\ \hline \langle L_1',L_p'\rangle & \cdots & \langle L_{p-1}',L_p'\rangle & \langle L_p',L_p'\rangle+\langle L_0',L_0'\rangle \end{array}\right], \qquad K_1 = \left[\begin{array}{ccc|c} 0 & \cdots & 0 & \langle L_0',L_1'\rangle \\ 0 & \cdots & 0 & \langle L_0',L_2'\rangle \\ \vdots & & \vdots & \vdots \\ 0 & \cdots & 0 & \langle L_0',L_p'\rangle \end{array}\right], \tag{3.21}$$

$$M_0 = \left[\begin{array}{ccc|c} \langle L_1,L_1\rangle & \cdots & \langle L_{p-1},L_1\rangle & \langle L_p,L_1\rangle \\ \vdots & & \vdots & \vdots \\ \langle L_1,L_{p-1}\rangle & \cdots & \langle L_{p-1},L_{p-1}\rangle & \langle L_p,L_{p-1}\rangle \\ \hline \langle L_1,L_p\rangle & \cdots & \langle L_{p-1},L_p\rangle & \langle L_p,L_p\rangle+\langle L_0,L_0\rangle \end{array}\right], \qquad M_1 = \left[\begin{array}{ccc|c} 0 & \cdots & 0 & \langle L_0,L_1\rangle \\ 0 & \cdots & 0 & \langle L_0,L_2\rangle \\ \vdots & & \vdots & \vdots \\ 0 & \cdots & 0 & \langle L_0,L_p\rangle \end{array}\right], \tag{3.22}$$

where $L_0,\dots,L_p$ are the Lagrange polynomials (3.5)*. In particular, $K_n^{(p)}, M_n^{(p)}$ are the leading principal submatrices of order $np-1$ of the block Toeplitz matrices $T_n(\mathbf{f}_p), T_n(\mathbf{h}_p)$, respectively, where $\mathbf{f}_p,\mathbf{h}_p:[-\pi,\pi]\to\mathbb{C}^{p\times p}$ are Hermitian matrix-valued functions given by*

$$\begin{aligned}\mathbf{f}_p(\theta) &:= K_0 + K_1\mathrm{e}^{\mathrm{i}\theta} + K_1^T\mathrm{e}^{-\mathrm{i}\theta} \\ &= \left[\begin{array}{ccc|c} \langle L_1',L_1'\rangle & \cdots & \langle L_{p-1}',L_1'\rangle & \langle L_p',L_1'\rangle+\langle L_0',L_1'\rangle\mathrm{e}^{\mathrm{i}\theta} \\ \vdots & & \vdots & \vdots \\ \langle L_1',L_{p-1}'\rangle & \cdots & \langle L_{p-1}',L_{p-1}'\rangle & \langle L_p',L_{p-1}'\rangle+\langle L_0',L_{p-1}'\rangle\mathrm{e}^{\mathrm{i}\theta} \\ \hline \langle L_1',L_p'\rangle+\langle L_0',L_1'\rangle\mathrm{e}^{-\mathrm{i}\theta} & \cdots & \langle L_{p-1}',L_p'\rangle+\langle L_0',L_{p-1}'\rangle\mathrm{e}^{-\mathrm{i}\theta} & \langle L_p',L_p'\rangle+\langle L_0',L_0'\rangle+2\langle L_0',L_p'\rangle\cos\theta \end{array}\right] \\ &= \left[\begin{array}{c|c} [\langle L_j',L_i'\rangle]_{i,j=1}^{p-1} & \left[\langle L_p',L_i'\rangle+\langle L_0',L_i'\rangle\mathrm{e}^{\mathrm{i}\theta}\right]_{i=1}^{p-1} \\ \hline \left[\langle L_p',L_i'\rangle+\langle L_0',L_i'\rangle\mathrm{e}^{-\mathrm{i}\theta}\right]_{i=1}^{p-1} & \langle L_p',L_p'\rangle+\langle L_0',L_0'\rangle+2\langle L_0',L_p'\rangle\cos\theta \end{array}\right],\end{aligned} \tag{3.23}$$

$$\begin{aligned}\mathbf{h}_p(\theta) &:= M_0 + M_1\mathrm{e}^{\mathrm{i}\theta} + M_1^T\mathrm{e}^{-\mathrm{i}\theta} \\ &= \left[\begin{array}{ccc|c} \langle L_1,L_1\rangle & \cdots & \langle L_{p-1},L_1\rangle & \langle L_p,L_1\rangle+\langle L_0,L_1\rangle\mathrm{e}^{\mathrm{i}\theta} \\ \vdots & & \vdots & \vdots \\ \langle L_1,L_{p-1}\rangle & \cdots & \langle L_{p-1},L_{p-1}\rangle & \langle L_p,L_{p-1}\rangle+\langle L_0,L_{p-1}\rangle\mathrm{e}^{\mathrm{i}\theta} \\ \hline \langle L_1,L_p\rangle+\langle L_0,L_1\rangle\mathrm{e}^{-\mathrm{i}\theta} & \cdots & \langle L_{p-1},L_p\rangle+\langle L_0,L_{p-1}\rangle\mathrm{e}^{-\mathrm{i}\theta} & \langle L_p,L_p\rangle+\langle L_0,L_0\rangle+2\langle L_0,L_p\rangle\cos\theta \end{array}\right] \\ &= \left[\begin{array}{c|c} [\langle L_j,L_i\rangle]_{i,j=1}^{p-1} & \left[\langle L_p,L_i\rangle+\langle L_0,L_i\rangle\mathrm{e}^{\mathrm{i}\theta}\right]_{i=1}^{p-1} \\ \hline \left[\langle L_p,L_i\rangle+\langle L_0,L_i\rangle\mathrm{e}^{-\mathrm{i}\theta}\right]_{i=1}^{p-1} & \langle L_p,L_p\rangle+\langle L_0,L_0\rangle+2\langle L_0,L_p\rangle\cos\theta \end{array}\right].\end{aligned} \tag{3.24}$$

Proof. We only give the construction of $K_n^{(p)}$, since the construction of $M_n^{(p)}$ is similar. For convenience, denote by K the matrix in the right-hand side of the first equality in (3.20): we have to show that $K_n^{(p)} = K$. Since both $K_n^{(p)}$ and K are symmetric, it suffices to show that

$$(K_n^{(p)})_{ij} = K_{ij} \qquad \forall i, j = 1, \dots, np-1 \ \text{ with } \ i \geq j. \tag{3.25}$$

As in Section 3.1, set $\xi_i = \frac{i}{np}$ for $i = 0, \dots, np$ and let $\{\ell_{1,(p)}, \dots, \ell_{np-1,(p)}\}$ be the Lagrangian basis for $W_n^{(p)}$. For all integers j, let $j_p = j \bmod p \in \{0, \dots, p-1\}$. To prove (3.25), we first notice that, for all $i, j = 1, \dots, np-1$ with $i \geq j$,

$$K_{ij} = \begin{cases} \langle L_p', L_p' \rangle + \langle L_0', L_0' \rangle & \text{if } j \text{ is a multiple of } p \text{ and } i = j \\ \langle L_0', L_{i_p}' \rangle & \text{if } j \text{ is a multiple of } p \text{ and } j < i < j + p \\ \langle L_0', L_p' \rangle & \text{if } j \text{ is a multiple of } p \text{ and } i = j + p \\ 0 & \text{if } j \text{ is a multiple of } p \text{ and } i > j + p \\ \langle L_{j_p}', L_{i_p}' \rangle & \text{if } j \text{ is not a multiple of } p \text{ and } j \leq i < j - j_p + p \\ \langle L_{j_p}', L_p' \rangle & \text{if } j \text{ is not a multiple of } p \text{ and } i = j - j_p + p \\ 0 & \text{if } j \text{ is not a multiple of } p \text{ and } i > j - j_p + p \end{cases} \tag{3.26}$$

We verify that (3.25) holds by considering the seven cases in (3.26). The verification is plain: it suffices to use the expressions of $\ell_{1,(p)}, \dots, \ell_{np-1,(p)}$ and of their derivatives given in (3.6)–(3.9). For completeness, we include this verification.

(i) If j is a multiple of p and $i = j$, then

$$\begin{aligned} (K_n^{(p)})_{ij} &= \frac{1}{n} \int_0^1 \ell_{j,(p)}'(x)^2 dx = n \int_{\xi_{j-p}}^{\xi_j} L_p'(nx - n\xi_{j-p})^2 dx + n \int_{\xi_j}^{\xi_{j+p}} L_0'(nx - n\xi_j)^2 dx \quad \text{(by (3.7))} \\ &= \int_0^1 L_p'(t)^2 dt + \int_0^1 L_0'(t)^2 dt = \langle L_p', L_p' \rangle + \langle L_0', L_0' \rangle = K_{ij}. \end{aligned}$$

(ii) If j is a multiple of p and $j < i < j + p$, then i is not a multiple of p, $i - i_p = j$, $\operatorname{supp}(\ell_{i,(p)}) = [\xi_j, \xi_{j+p}]$ and

$$\begin{aligned} (K_n^{(p)})_{ij} &= \frac{1}{n} \int_0^1 \ell_{j,(p)}'(x) \ell_{i,(p)}'(x) dx = n \int_{\xi_j}^{\xi_{j+p}} L_0'(nx - n\xi_j) L_{i_p}'(nx - n\xi_j) dx \quad \text{(by (3.7) and (3.9))} \\ &= \int_0^1 L_0'(t) L_{i_p}'(t) dt = \langle L_0', L_{i_p}' \rangle = K_{ij}. \end{aligned}$$

(iii) If j is a multiple of p and $i = j + p$, then i is a multiple of p, $\operatorname{supp}(\ell_{i,(p)}) \cap \operatorname{supp}(\ell_{j,(p)}) = [\xi_{i-p}, \xi_{i+p}] \cap [\xi_{j-p}, \xi_{j+p}] = [\xi_j, \xi_{j+2p}] \cap [\xi_{j-p}, \xi_{j+p}] = [\xi_j, \xi_{j+p}]$ and

$$\begin{aligned} (K_n^{(p)})_{ij} &= \frac{1}{n} \int_0^1 \ell_{j,(p)}'(x) \ell_{i,(p)}'(x) dx = n \int_{\xi_j}^{\xi_{j+p}} L_0'(nx - n\xi_j) L_p'(nx - n\xi_j) dx \quad \text{(by (3.7))} \\ &= \int_0^1 L_0'(t) L_p'(t) dt = \langle L_0', L_p' \rangle = K_{ij}. \end{aligned}$$

(iv) If j is a multiple of p and $i > j + p$, then $\operatorname{supp}(\ell_{i,(p)}) \subseteq [\xi_{j+p}, 1]$ and $\operatorname{supp}(\ell_{j,(p)}) = [\xi_{j-p}, \xi_{j+p}]$, and so

$$(K_n^{(p)})_{ij} = \frac{1}{n} \int_0^1 \ell_{j,(p)}'(x) \ell_{i,(p)}'(x) dx = 0 = K_{ij}.$$

(v) If j is not a multiple of p and $j \le i < j - j_p + p$, then i is not a multiple of p, $i - i_p = j - j_p$, $\mathrm{supp}(\ell_{i,(p)}) = [\xi_{i-i_p}, \xi_{i-i_p+p}] = [\xi_{j-j_p}, \xi_{j-j_p+p}] = \mathrm{supp}(\ell_{j,(p)})$ and

$$(K_n^{(p)})_{ij} = \frac{1}{n}\int_0^1 \ell'_{j,(p)}(x)\ell'_{i,(p)}(x)dx = n\int_{\xi_{j-j_p}}^{\xi_{j-j_p+p}} L'_{j_p}(nx - n\xi_{j-j_p})L'_{i_p}(nx - n\xi_{j-j_p})dx \quad \text{(by (3.9))}$$
$$= \int_0^1 L'_{j_p}(t)L'_{i_p}(t)dt = \langle L'_{j_p}, L'_{i_p}\rangle = K_{ij}.$$

(vi) If j is not a multiple of p and $i = j - j_p + p$, then i is a multiple of p, $i - p = j - j_p$, $\mathrm{supp}(\ell_{i,(p)}) \cap \mathrm{supp}(\ell_{j,(p)}) = [\xi_{i-p}, \xi_{i+p}] \cap [\xi_{j-j_p}, \xi_{j-j_p+p}] = [\xi_{j-j_p}, \xi_{j-j_p+2p}] \cap [\xi_{j-j_p}, \xi_{j-j_p+p}] = [\xi_{j-j_p}, \xi_{j-j_p+p}]$ and

$$(K_n^{(p)})_{ij} = \frac{1}{n}\int_0^1 \ell'_{j,(p)}(x)\ell'_{i,(p)}(x)dx = n\int_{\xi_{j-j_p}}^{\xi_{j-j_p+p}} L'_{j_p}(nx - n\xi_{j-j_p})L'_{p}(nx - n\xi_{j-j_p})dx \quad \text{(by (3.9))}$$
$$= \int_0^1 L'_{j_p}(t)L'_{p}(t)dt = \langle L'_{j_p}, L'_{p}\rangle = K_{ij}.$$

(vii) If j is not a multiple of p and $i > j - j_p + p$, then $\mathrm{supp}(\ell_{i,(p)}) \subseteq [\xi_{j-j_p+p}, 1]$ and $\mathrm{supp}(\ell_{j,(p)}) = [\xi_{j-j_p}, \xi_{j-j_p+p}]$, and so

$$(K_n^{(p)})_{ij} = \frac{1}{n}\int_0^1 \ell'_{j,(p)}(x)\ell'_{i,(p)}(x)dx = 0 = K_{ij}.$$

□

We note that, if $L_0, \dots, L_p$ are the Lagrange polynomials (3.5), then, for every $h = 0, \dots, p$ and every $t \in \mathbb{R}$, a direct verification shows that $L_h(1-t) = L_{p-h}(t)$. As a consequence, the equalities $\langle L_i, L_j\rangle = \langle L_{p-i}, L_{p-j}\rangle$ and $\langle L'_i, L'_j\rangle = \langle L'_{p-i}, L'_{p-j}\rangle$ hold for all $i, j = 0, \dots, p$. These relations may be used to give alternative expressions for the entries of the blocks K_0, K_1, M_0, M_1 in (3.21)–(3.22).

3.3 Properties of $\mathbf{f}_p(\theta)$ and $\mathbf{h}_p(\theta)$

In this section we derive some properties of the Hermitian matrix-valued functions $\mathbf{f}_p(\theta)$, $\mathbf{h}_p(\theta)$ defined in (3.23)–(3.24). We need some results concerning the Lagrange polynomials.

Lemma 3.2. *Let $p \ge 1$ and let $L_0, \dots, L_p$ be the Lagrange polynomials* (3.5)*. Then*

$$\sum_{j=1}^{p} jL'_j = p \quad \textit{identically,} \tag{3.27}$$

$$\sum_{j=0}^{p} L'_j = 0 \quad \textit{identically,} \tag{3.28}$$

while every proper subset of $\{L'_0, \dots, L'_p\}$ is linearly independent.

Proof. (3.27) holds because $\sum_{j=1}^{p} jL_j = \sum_{j=0}^{p} jL_j$ is the interpolating polynomial which takes the value j over the knot $t_j = \frac{j}{p}$ for $j = 0, \dots, p$, and hence $\sum_{j=1}^{p} jL_j(t) = pt$ identically. (3.28) holds because $\sum_{j=0}^{p} L_j$ is the interpolating polynomial which takes the value 1 over the uniform knots $t_k = \frac{k}{p}$, $k = 0, \dots, p$, and hence $\sum_{j=0}^{p} L_j = 1$ identically.

We prove that every proper subset of $\{L'_0, \dots, L'_p\}$ is linearly independent. To this end, it suffices to prove that every proper subset of $\{L'_0, \dots, L'_p\}$ with cardinality p is linearly independent. Actually, we will only

prove that $\{L_1', \ldots, L_p'\}$ is linearly independent, since the proof for the other subsets is similar. Let $\alpha_1, \ldots, \alpha_p$ be numbers such that $\sum_{i=1}^p \alpha_i L_i' = \left(\sum_{i=1}^p \alpha_i L_i\right)' = 0$ identically. Then there exists a constant C such that

$$\sum_{i=1}^{p} \alpha_i L_i = C \quad \text{identically.} \tag{3.29}$$

By evaluating (3.29) in $t_k = \frac{k}{p}$, $k = 0, \ldots, p$, and by remembering (3.5), we find that $C = 0$ and $\alpha_1 = \ldots = \alpha_p = C$, which yields $\alpha_1 = \ldots = \alpha_p = 0$. Thus $L_1', \ldots, L_p'$ are linearly independent. □

Lemma 3.3. *Let $p \geq 1$ and set $d_p := \det([\langle L_j', L_i' \rangle]_{i,j=1}^p)$, where $L_0, \ldots, L_p$ are the Lagrange polynomials* (3.5)*. Then $d_p > 0$ and $d_p = \det([\langle L_j', L_i' \rangle]_{i,j=1}^{p-1})$.*[2]

Proof. The lemma is true if $p = 1$, because $L_1(t) = t$, $L_1'(t) = 1$, and $d_1 = \langle L_1', L_1' \rangle = 1$. In the following we assume $p \geq 2$. We have $d_p > 0$ because the matrix $[\langle L_j', L_i' \rangle]_{i,j=1}^p$ is SPD, due to the fact that $L_1', \ldots, L_p'$ are linearly independent (Lemma 3.2).

We want to show that $d_p = \det(\mathscr{L})$, where $\mathscr{L} := [\langle L_j', L_i' \rangle]_{i,j=1}^{p-1}$. To this end, we perform the block Gauss transformation that creates zeros in the first $p-1$ components of the last row and column of $[\langle L_j', L_i' \rangle]_{i,j=1}^p$. Setting

$$G := \left[\begin{array}{c|c} I_{p-1} & \mathbf{0} \\ \hline -[\langle L_1', L_p' \rangle \ \cdots \ \langle L_{p-1}', L_p' \rangle]\mathscr{L}^{-1} & 1 \end{array}\right],$$

we have

$$G[\langle L_j', L_i' \rangle]_{i,j=1}^p G^T = G \left[\begin{array}{ccc|c} & \mathscr{L} & & \begin{array}{c}\langle L_p', L_1' \rangle \\ \vdots \\ \langle L_p', L_{p-1}' \rangle\end{array} \\ \hline \langle L_1', L_p' \rangle & \cdots & \langle L_{p-1}', L_p' \rangle & \langle L_p', L_p' \rangle \end{array}\right] G^T = \left[\begin{array}{c|c} \mathscr{L} & \mathbf{0} \\ \hline \mathbf{0}^T & s \end{array}\right] =: Z,$$

where $s := \langle L_p', L_p' \rangle - [\langle L_1', L_p' \rangle \ \cdots \ \langle L_{p-1}', L_p' \rangle]\mathscr{L}^{-1} \begin{bmatrix} \langle L_p', L_1' \rangle \\ \vdots \\ \langle L_p', L_{p-1}' \rangle \end{bmatrix}$ is the Schur complement of $\mathscr{L}$. Since $\det(G) = \det(G^T) = 1$, we have $d_p = \det(Z) = \det(\mathscr{L})s$, and so the lemma is proved if we show that $s = 1$. To prove this, we note that $\mathscr{L}^{-1} \begin{bmatrix} \langle L_p', L_1' \rangle \\ \vdots \\ \langle L_p', L_{p-1}' \rangle \end{bmatrix}$ is the solution of the linear system $\mathscr{L}\mathbf{u} = \begin{bmatrix} \langle L_p', L_1' \rangle \\ \vdots \\ \langle L_p', L_{p-1}' \rangle \end{bmatrix}$, which is easily seen to be $\mathbf{u} := \left[-\frac{1}{p}, -\frac{2}{p}, \ldots, -\frac{p-1}{p}\right]^T$. Indeed, by Lemma 3.2, for all $i = 1, \ldots, p-1$ we have

$$\begin{aligned}(\mathscr{L}\mathbf{u})_i &= \sum_{j=1}^{p-1} \langle L_j', L_i' \rangle u_j = -\frac{1}{p} \sum_{j=1}^{p-1} j \langle L_j', L_i' \rangle = -\frac{1}{p} \left\langle \sum_{j=1}^{p-1} j L_j', L_i' \right\rangle = -\frac{1}{p} \langle p - p L_p', L_i' \rangle = -\langle 1, L_i' \rangle + \langle L_p', L_i' \rangle \\ &= -\int_0^1 L_i'(t) dt + \langle L_p', L_i' \rangle = -L_i(1) + L_i(0) + \langle L_p', L_i' \rangle = \langle L_p', L_i' \rangle,\end{aligned}$$

[2] We use the (standard) convention that the determinant of the empty matrix is 1, so that the latter formula gives $d_1 = 1$.

where the last equality is due to the fact that $L_i(0) = L_i(1) = 0$ for $i = 1, \dots, p-1$. Using again Lemma 3.2, we obtain

$$s = \langle L'_p, L'_p \rangle - [\langle L'_1, L'_p \rangle \ \cdots \ \langle L'_{p-1}, L'_p \rangle] \mathscr{L}^{-1} \begin{bmatrix} \langle L'_p, L'_1 \rangle \\ \vdots \\ \langle L'_p, L'_{p-1} \rangle \end{bmatrix} = \langle L'_p, L'_p \rangle - [\langle L'_1, L'_p \rangle \ \cdots \ \langle L'_{p-1}, L'_p \rangle] \mathbf{u}$$

$$= \langle L'_p, L'_p \rangle - \sum_{j=1}^{p-1} \langle L'_j, L'_p \rangle u_j = \langle L'_p, L'_p \rangle + \left\langle \sum_{j=1}^{p-1} \frac{j}{p} L'_j, L'_p \right\rangle = \left\langle \sum_{j=1}^{p} \frac{j}{p} L'_j, L'_p \right\rangle = \langle 1, L'_p \rangle$$

$$= \int_0^1 L'_p(t) dt = L_p(1) - L_p(0) = 1,$$

which concludes the proof. □

Theorem 3.3. *Let $p \geq 1$, then*

$$\det(\mathbf{f}_p(\theta)) = d_p(2 - 2\cos\theta), \tag{3.30}$$

where d_p is defined in Lemma 3.3.

Proof. The theorem is true if $p = 1$, because $d_1 = 1$ and $\mathbf{f}_1(\theta) = 2 - 2\cos\theta$. In the following we assume $p \geq 2$. By (3.23) and by the linearity of the determinant with respect to each row and column, we have

$$\begin{aligned}
\det(\mathbf{f}_p(\theta)) &= \left| \begin{array}{c|c} [\langle L'_j, L'_i \rangle]_{i,j=1}^{p-1} & [\langle L'_p, L'_i \rangle + \langle L'_0, L'_i \rangle \mathrm{e}^{\mathrm{i}\theta}]_{i=1}^{p-1} \\ \hline \left[\langle L'_p, L'_i \rangle + \langle L'_0, L'_i \rangle \mathrm{e}^{-\mathrm{i}\theta}\right]_{i=1}^{p-1} & \langle L'_p, L'_p \rangle + \langle L'_0, L'_0 \rangle + 2\langle L'_0, L'_p \rangle \cos\theta \end{array} \right| \\
&= \left| \begin{array}{c|c} [\langle L'_j, L'_i \rangle]_{i,j=1}^{p-1} & [\langle L'_p, L'_i \rangle]_{i=1}^{p-1} \\ \hline \left[\langle L'_p, L'_i \rangle + \langle L'_0, L'_i \rangle \mathrm{e}^{-\mathrm{i}\theta}\right]_{i=1}^{p-1} & \langle L'_p, L'_p \rangle + \langle L'_0, L'_p \rangle \mathrm{e}^{-\mathrm{i}\theta} \end{array} \right| \\
&\quad + \left| \begin{array}{c|c} [\langle L'_j, L'_i \rangle]_{i,j=1}^{p-1} & [\langle L'_0, L'_i \rangle \mathrm{e}^{\mathrm{i}\theta}]_{i=1}^{p-1} \\ \hline \left[\langle L'_p, L'_i \rangle + \langle L'_0, L'_i \rangle \mathrm{e}^{-\mathrm{i}\theta}\right]_{i=1}^{p-1} & \langle L'_0, L'_0 \rangle + \langle L'_0, L'_p \rangle \mathrm{e}^{\mathrm{i}\theta} \end{array} \right| \\
&= \left| \begin{array}{c|c} [\langle L'_j, L'_i \rangle]_{i,j=1}^{p-1} & [\langle L'_p, L'_i \rangle]_{i=1}^{p-1} \\ \hline \left[\langle L'_p, L'_i \rangle\right]_{i=1}^{p-1} & \langle L'_p, L'_p \rangle \end{array} \right| + \left| \begin{array}{c|c} [\langle L'_j, L'_i \rangle]_{i,j=1}^{p-1} & [\langle L'_p, L'_i \rangle]_{i=1}^{p-1} \\ \hline \left[\langle L'_0, L'_i \rangle \mathrm{e}^{-\mathrm{i}\theta}\right]_{i=1}^{p-1} & \langle L'_0, L'_p \rangle \mathrm{e}^{-\mathrm{i}\theta} \end{array} \right| \\
&\quad + \left| \begin{array}{c|c} [\langle L'_j, L'_i \rangle]_{i,j=1}^{p-1} & [\langle L'_0, L'_i \rangle \mathrm{e}^{\mathrm{i}\theta}]_{i=1}^{p-1} \\ \hline \left[\langle L'_p, L'_i \rangle\right]_{i=1}^{p-1} & \langle L'_0, L'_p \rangle \mathrm{e}^{\mathrm{i}\theta} \end{array} \right| + \left| \begin{array}{c|c} [\langle L'_j, L'_i \rangle]_{i,j=1}^{p-1} & [\langle L'_0, L'_i \rangle \mathrm{e}^{\mathrm{i}\theta}]_{i=1}^{p-1} \\ \hline \left[\langle L'_0, L'_i \rangle \mathrm{e}^{-\mathrm{i}\theta}\right]_{i=1}^{p-1} & \langle L'_0, L'_0 \rangle \end{array} \right| \\
&= d_p + \mathrm{e}^{-\mathrm{i}\theta} \left| \begin{array}{c|c} [\langle L'_j, L'_i \rangle]_{i,j=1}^{p-1} & [\langle L'_p, L'_i \rangle]_{i=1}^{p-1} \\ \hline \left[\langle L'_0, L'_i \rangle\right]_{i=1}^{p-1} & \langle L'_0, L'_p \rangle \end{array} \right| + \mathrm{e}^{\mathrm{i}\theta} \left| \begin{array}{c|c} [\langle L'_j, L'_i \rangle]_{i,j=1}^{p-1} & [\langle L'_0, L'_i \rangle]_{i=1}^{p-1} \\ \hline \left[\langle L'_p, L'_i \rangle\right]_{i=1}^{p-1} & \langle L'_0, L'_p \rangle \end{array} \right| \\
&\quad + \left| \begin{array}{c|c} [\langle L'_j, L'_i \rangle]_{i,j=1}^{p-1} & [\langle L'_0, L'_i \rangle]_{i=1}^{p-1} \\ \hline \left[\langle L'_0, L'_i \rangle\right]_{i=1}^{p-1} & \langle L'_0, L'_0 \rangle \end{array} \right| =: d_p + \mathrm{e}^{-\mathrm{i}\theta} d'_p + \mathrm{e}^{\mathrm{i}\theta} d'_p + d''_p = d_p + 2d'_p \cos\theta + d''_p.
\end{aligned} \tag{3.31}$$

We prove that

$$\det(\mathbf{f}_p(0)) = d_p + 2d'_p + d''_p = 0, \tag{3.32}$$

$$d_p + d'_p = 0, \tag{3.33}$$

after which (3.30) follows from (3.31). By (3.23) we have

$$\mathbf{f}_p(0) = \left[\begin{array}{c|c} [\langle L'_j, L'_i \rangle]_{i,j=1}^{p-1} & [\langle L'_0 + L'_p, L'_i \rangle]_{i=1}^{p-1} \\ \hline [\langle L'_0 + L'_p, L'_i \rangle]_{i=1}^{p-1} & \langle L'_0 + L'_p, L'_0 + L'_p \rangle \end{array} \right] =: [\langle N'_j, N'_i \rangle]_{i,j=1}^{p},$$

where $N_i := L_i$ for $i = 1, \dots, p-1$ and $N_p := L_0 + L_p$. Since $\sum_{i=1}^{p} N_i' = \sum_{i=0}^{p} L_i' = 0$ identically (Lemma 3.2), it follows that $N_1', \dots, N_p'$ are linearly dependent, $\mathbf{f}_p(0)$ is singular, and (3.32) holds. To prove (3.33) we simply note that

$$d_p + d_p' = \left| \begin{array}{c|c} [\langle L_j', L_i' \rangle]_{i,j=1}^{p-1} & [\langle L_p', L_i' \rangle]_{i=1}^{p-1} \\ \hline [\langle L_p', L_i' \rangle]_{i=1}^{p-1} & \langle L_p', L_p' \rangle \end{array} \right| + \left| \begin{array}{c|c} [\langle L_j', L_i' \rangle]_{i,j=1}^{p-1} & [\langle L_0', L_i' \rangle]_{i=1}^{p-1} \\ \hline [\langle L_p', L_i' \rangle]_{i=1}^{p-1} & \langle L_0', L_p' \rangle \end{array} \right| = \left| \begin{array}{c|c} [\langle L_j', L_i' \rangle]_{i,j=1}^{p-1} & [\langle L_p' + L_0', L_i' \rangle]_{i=1}^{p-1} \\ \hline [\langle L_p', L_i' \rangle]_{i=1}^{p-1} & \langle L_p' + L_0', L_p' \rangle \end{array} \right| = 0,$$

where the latter is a consequence of the fact that, by Lemma 3.2, $L_p' + L_0'$ is a linear combination of $L_1', \dots, L_{p-1}'$, which implies that the last column of $d_p + d_p'$ is a linear combination of the others. □

Theorem 3.4. *Let $p \in \{1, \dots, 15\}$, then*

$$\det(\mathbf{h}_p(\theta)) = a_p \left(1 + \frac{(-1)^{p+1}}{p+1} \cos\theta\right), \tag{3.34}$$

where $a_p = \det(\mathbf{h}_p(\frac{\pi}{2})) > 0$.

Proof. The result has been verified by direct computation using MAPLE. □

Although the result of Theorem 3.4 has not been proved for all $p \geq 1$, we can certainly formulate the following conjecture.

Conjecture 3.1. *Theorem 3.4 holds for all $p \geq 1$.*

Remark 3.1. Using the same computations as in the proof of Theorem 3.3, it is not difficult to see that $\det(\mathbf{h}_p(\theta)) = a_p + b_p \cos\theta$ for some constant a_p, b_p independent of θ. Thus, Eq. (3.34) is proved if we are able to show that $b_p = a_p \frac{(-1)^{p+1}}{p+1}$. Once we have proved this, we do not need to prove also that $a_p = \det(\mathbf{h}_p(\frac{\pi}{2})) > 0$. Indeed, if (3.34) holds with some constant a_p, then, by evaluating both sides at $\theta = \frac{\pi}{2}$, we immediately get $a_p = \det(\mathbf{h}_p(\frac{\pi}{2}))$; moreover, $a_p > 0$. To see this, note that $\mathbf{h}_p(0) = [\langle N_j, N_i \rangle]_{i,j=1}^{p}$ with $N_i := L_i$ for $i = 1, \dots, p-1$ and $N_p := L_0 + L_p$. Since $N_1, \dots, N_p$ are linearly independent, due to the linear independence of $L_0, \dots, L_p$, it follows that $\mathbf{h}_p(0) > O$. Hence, $\det(\mathbf{h}_p(0)) > 0$ and

$$a_p = \frac{\det(\mathbf{h}_p(0))}{\left(1 + \frac{(-1)^{p+1}}{p+1}\right)} > 0.$$

From now on, we will assume that Conjecture 3.1 holds. The results relying on this conjecture are certainly true for $p = 1, \dots, 15$.

In the following, for $p \geq 2$ we denote by $\mu_1^{(p)} \geq \dots \geq \mu_{p-1}^{(p)} > 0$ and $\eta_1^{(p)} \geq \dots \geq \eta_{p-1}^{(p)} > 0$ the eigenvalues of the SPD matrices $[\langle L_j', L_i' \rangle]_{i,j=1}^{p-1}$ and $[\langle L_j, L_i \rangle]_{i,j=1}^{p-1}$, respectively, where $L_0, \dots, L_p$ are the Lagrange polynomials (3.5). Moreover, we define

$$m_{\mathbf{f}_p} := \min_{\theta \in [-\pi, \pi]} \lambda_{\min}(\mathbf{f}_p(\theta)), \qquad M_{\mathbf{f}_p} := \max_{\theta \in [-\pi, \pi]} \lambda_{\max}(\mathbf{f}_p(\theta)),$$

$$m_{\mathbf{h}_p} := \min_{\theta \in [-\pi, \pi]} \lambda_{\min}(\mathbf{h}_p(\theta)), \qquad M_{\mathbf{h}_p} := \max_{\theta \in [-\pi, \pi]} \lambda_{\max}(\mathbf{h}_p(\theta)).$$

Corollary 3.1. *The following properties hold.*

1. *Let $p \geq 2$, then $\lambda_1(\mathbf{f}_p(\theta)) \geq \mu_1^{(p)} \geq \lambda_2(\mathbf{f}_p(\theta)) \geq \mu_2^{(p)} \geq \dots \geq \lambda_{p-1}(\mathbf{f}_p(\theta)) \geq \mu_{p-1}^{(p)} \geq \lambda_p(\mathbf{f}_p(\theta))$ for all θ.*

2. *Let $p \geq 1$, then there exists a constant $c_p > 0$ such that, for all θ,*

$$c_p(2 - 2\cos\theta) \leq \lambda_{\min}(\mathbf{f}_p(\theta)) \leq 2 - 2\cos\theta. \tag{3.35}$$

In (3.35) we can take $c_1 = 1$ and $c_p = \dfrac{\mu_{p-1}^{(p)}}{M_{\mathbf{f}_p}}$ for $p \geq 2$. In particular, $m_{\mathbf{f}_p} = 0$, $\mathbf{f}_p(\theta) \geq O$ for all $\theta \in [-\pi, \pi]$, and $\mathbf{f}_p(\theta) > O$ for all nonzero $\theta \in [-\pi, \pi]$.

Proof. For $p = 1$ the corollary can be directly verified, because $\mathbf{f}_1(\theta) = 2 - 2\cos\theta$. Assume $p \geq 2$. Item 1 follows from the Cauchy interlacing theorem and from the fact that $[\langle L_j', L_i' \rangle]_{i,j=1}^{p-1}$ is the leading principal submatrix of $\mathbf{f}_p(\theta)$ for all θ. To prove item 2, observe that, by Theorem 3.3,

$$\lambda_1(\mathbf{f}_p(\theta)) \cdots \lambda_p(\mathbf{f}_p(\theta)) = \det(\mathbf{f}_p(\theta)) = d_p(2 - 2\cos\theta) \;\Rightarrow\; \lambda_{\min}(\mathbf{f}_p(\theta)) = \frac{d_p}{\lambda_1(\mathbf{f}_p(\theta)) \cdots \lambda_{p-1}(\mathbf{f}_p(\theta))}(2 - 2\cos\theta).$$

Furthermore, by item 1 and Lemma 3.3, for all θ we have

$$\begin{aligned}
\lambda_1(\mathbf{f}_p(\theta)) \cdots \lambda_{p-1}(\mathbf{f}_p(\theta)) &\geq \mu_1^{(p)} \cdots \mu_{p-1}^{(p)} = \det([\langle L_j', L_i' \rangle]_{i,j=1}^{p-1}) = d_p, \\
\lambda_1(\mathbf{f}_p(\theta)) \cdots \lambda_{p-1}(\mathbf{f}_p(\theta)) &\leq M_{\mathbf{f}_p} \mu_1^{(p)} \cdots \mu_{p-2}^{(p)} = \frac{M_{\mathbf{f}_p} \mu_1^{(p)} \cdots \mu_{p-1}^{(p)}}{\mu_{p-1}^{(p)}} = \frac{M_{\mathbf{f}_p} d_p}{\mu_{p-1}^{(p)}},
\end{aligned}$$

and item 2 follows. □

Corollary 3.2. *The following properties hold.*

1. *Let $p \geq 2$, then $\lambda_1(\mathbf{h}_p(\theta)) \geq \eta_1^{(p)} \geq \lambda_2(\mathbf{h}_p(\theta)) \geq \eta_2^{(p)} \geq \ldots \geq \lambda_{p-1}(\mathbf{h}_p(\theta)) \geq \eta_{p-1}^{(p)} \geq \lambda_p(\mathbf{h}_p(\theta))$ for all θ.*

2. *Let $p \geq 1$, then $m_{\mathbf{h}_p} > 0$. In particular, $\mathbf{h}_p(\theta) > O$ for all θ. In addition, $m_{\mathbf{h}_1} = \frac{1}{3}$, while for $p \geq 2$ we have $m_{\mathbf{h}_p} \geq \dfrac{p a_p \eta_{p-1}^{(p)}}{(p+1)\eta_1^{(p)} \cdots \eta_{p-1}^{(p)}}$, where $a_p = \det(\mathbf{h}_p(\frac{\pi}{2})) > 0$.*

Proof. For $p = 1$ the corollary can be directly verified, because $\mathbf{h}_1(\theta) = \frac{2}{3} + \frac{1}{3}\cos\theta$. Assume $p \geq 2$. Item 1 follows from the Cauchy interlacing theorem and from the fact that $[\langle L_j, L_i \rangle]_{i,j=1}^{p-1}$ is the leading principal submatrix of $\mathbf{h}_p(\theta)$ for all θ. To prove item 2, we simply note that, by item 1 and Conjecture 3.1,

$$\begin{aligned}
&\lambda_1(\mathbf{h}_p(\theta)) \cdots \lambda_p(\mathbf{h}_p(\theta)) = \det(\mathbf{h}_p(\theta)) = a_p\left(1 + \frac{(-1)^{p+1}}{p+1}\cos\theta\right) \\
&\Rightarrow\; \lambda_{\min}(\mathbf{h}_p(\theta)) = \frac{a_p}{\lambda_1(\mathbf{h}_p(\theta)) \cdots \lambda_{p-1}(\mathbf{h}_p(\theta))}\left(1 + \frac{(-1)^{p+1}}{p+1}\cos\theta\right) \geq \frac{a_p}{M_{\mathbf{h}_p}\eta_1^{(p)} \cdots \eta_{p-2}^{(p)}}\left(1 - \frac{1}{p+1}\right).
\end{aligned}$$

□

Item 1 in Corollary 3.1 has the following geometric interpretation: the $p-1$ horizontal lines in the plane with ordinates $\mu_j^{(p)}$, $j = 1, \ldots, p-1$, are 'separating lines' for the eigenvalues of $\mathbf{f}_p(\theta)$. This is illustrated in Figure 3.2 for the cases $p = 2, 3$. Item 1 in Corollary 3.2 has an analogous geometric interpretation.

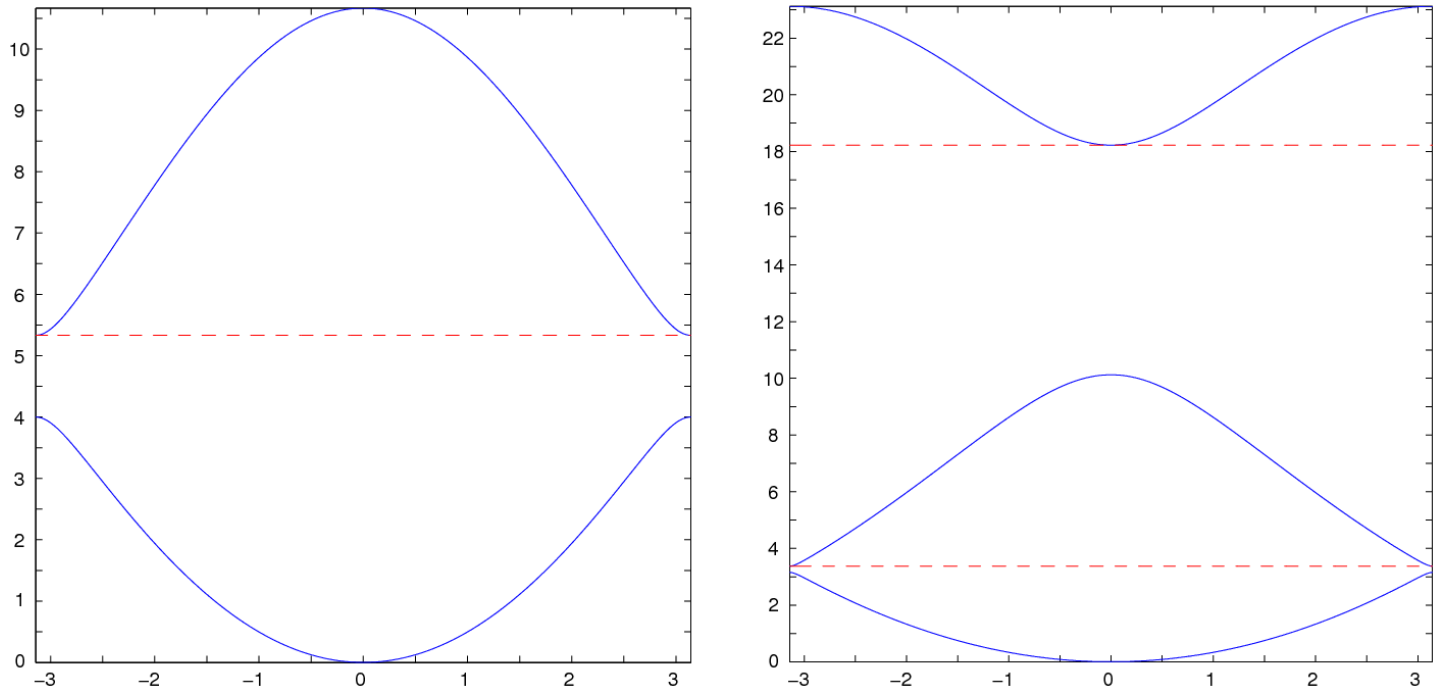


Figure 3.2: left: graph of the eigenvalue functions $\theta \mapsto \lambda_j(\mathbf{f}_2(\theta))$, $j = 1, 2$ (solid lines), and of the separating line with ordinate $\frac{16}{3}$ (dashed line); right: graph of the eigenvalue functions $\theta \mapsto \lambda_j(\mathbf{f}_3(\theta))$, $j = 1, 2, 3$ (solid lines), and of the separating lines with ordinates $\frac{729}{40}$ and $\frac{27}{8}$ (dashed lines).

3.4 Spectral analysis and spectral symbol

In this section we study the spectral properties of the stiffness matrix $A_{\boldsymbol{n}}^{(\boldsymbol{p})}$ in (3.11), focusing on the asymptotic behavior as the fineness parameters $\boldsymbol{n} \to \infty$. In particular, we give estimates for the eigenvalues and for the spectral condition number $\kappa(A_{\boldsymbol{n}}^{(\boldsymbol{p})})$. Moreover, assuming $\boldsymbol{n} = \boldsymbol{\nu} n = (\nu_1 n, \dots, \nu_d n) \in \mathbb{N}^d$ for a fixed $\boldsymbol{\nu} \in \mathbb{Q}_+^d$, we prove that the sequence $\{n^{d-2} A_{\boldsymbol{n}}^{(\boldsymbol{p})}\}_n$ has an asymptotic spectral distribution characterized by the Hermitian matrix-valued function

$$\mathbf{f}_{\boldsymbol{p}}^{(\boldsymbol{\nu})}(\boldsymbol{\theta}) : [-\pi, \pi]^d \to \mathbb{C}^{N(\boldsymbol{p}) \times N(\boldsymbol{p})}$$

$$\mathbf{f}_{\boldsymbol{p}}^{(\boldsymbol{\nu})}(\boldsymbol{\theta}) := \sum_{k=1}^{d} c_k(\boldsymbol{\nu})\, \mathbf{h}_{p_1}(\theta_1) \otimes \cdots \otimes \mathbf{h}_{p_{k-1}}(\theta_{k-1}) \otimes \mathbf{f}_{p_k}(\theta_k) \otimes \mathbf{h}_{p_{k+1}}(\theta_{k+1}) \otimes \cdots \otimes \mathbf{h}_{p_d}(\theta_d), \tag{3.36}$$

where $\mathbf{f}_p$ and $\mathbf{h}_p$ are given in (3.23)–(3.24), and

$$c_k(\boldsymbol{\nu}) := \frac{\nu_k}{\nu_1 \cdots \nu_{k-1} \nu_{k+1} \cdots \nu_d}, \quad k = 1, \dots, d. \tag{3.37}$$

Unfortunately, it turns out that the spectrum of $\mathbf{f}_{\boldsymbol{p}}^{(\boldsymbol{\nu})}$ presents an exponential scattering with respect to $\boldsymbol{p}$ and d, and this implies a substantial numerical difficulty in treating the linear systems associated with the matrix $n^{d-2} A_{\boldsymbol{n}}^{(\boldsymbol{p})}$, already for moderate $\boldsymbol{p}$ and d. In the last subsection, still assuming that $\boldsymbol{n} = \boldsymbol{\nu} n$ for some $\boldsymbol{\nu} \in \mathbb{Q}_+^d$, we investigate the clustering properties of the sequence $\{n^{d-2} A_{\boldsymbol{n}}^{(\boldsymbol{p})}\}_n$ and we show that $\{n^{d-2} A_{\boldsymbol{n}}^{(\boldsymbol{p})}\}_n$ is strongly clustered at $[0, M_{\mathbf{f}_{\boldsymbol{p}}^{(\boldsymbol{\nu})}}]$, where $M_{\mathbf{f}_{\boldsymbol{p}}^{(\boldsymbol{\nu})}} := \max_{\boldsymbol{\theta} \in [-\pi,\pi]^d} \lambda_{\max}(\mathbf{f}_{\boldsymbol{p}}^{(\boldsymbol{\nu})}(\boldsymbol{\theta}))$.

3.4.1 Estimates for the eigenvalues, localization of the spectrum and conditioning of $A_{\boldsymbol{n}}^{(\boldsymbol{p})}$

We first provide estimates for the eigenvalues of $K_n^{(p)}, M_n^{(p)}$. This is fundamental for estimating the condition number $\kappa(A_{\boldsymbol{n}}^{(\boldsymbol{p})})$. By Theorem 3.2, the matrices $K_n^{(p)}, M_n^{(p)}$ are the leading principal submatrices of order $np - 1$

of the Hermitian block Toeplitz matrices $T_n(\mathbf{f}_p), T_n(\mathbf{h}_p)$, respectively. Moreover, $T_{n-1}(\mathbf{f}_p), T_{n-1}(\mathbf{h}_p)$ are the leading principal submatrices of order $np-p$ of $K_n^{(p)}, M_n^{(p)}$, respectively. Hence, by Theorem 1.3 we have, for all j,

$$\lambda_j(T_n(\mathbf{f}_p)) \geq \lambda_j(K_n^{(p)}) \geq \lambda_{j+1}(T_n(\mathbf{f}_p)), \quad \lambda_j(K_n^{(p)}) \geq \lambda_j(T_{n-1}(\mathbf{f}_p)) \geq \lambda_{j+p-1}(K_n^{(p)}), \tag{3.38}$$

$$\lambda_j(T_n(\mathbf{h}_p)) \geq \lambda_j(M_n^{(p)}) \geq \lambda_{j+1}(T_n(\mathbf{h}_p)), \quad \lambda_j(M_n^{(p)}) \geq \lambda_j(T_{n-1}(\mathbf{h}_p)) \geq \lambda_{j+p-1}(M_n^{(p)}). \tag{3.39}$$

By (3.38)–(3.39), by Theorem 1.8, and by recalling that $m_{\mathbf{f}_p} = 0$ (Corollary 3.1), we have

$$\Lambda(K_n^{(p)}) \subset (0, M_{\mathbf{f}_p}], \qquad \Lambda(M_n^{(p)}) \subset [m_{\mathbf{h}_p}, M_{\mathbf{h}_p}]. \tag{3.40}$$

Note that the point 0 is excluded from $\Lambda(K_n^{(p)})$ either because $K_n^{(p)}$ is positive definite (see Theorem 3.1), or because, by Corollary 3.1, $\lambda_{\min}(\mathbf{f}_p(\theta))$ is not constant and so Theorem 1.8 excludes 0 from $\Lambda(T_n(\mathbf{f}_p))$. Furthermore, Theorem 1.8 and (3.38)–(3.39) imply that, for each fixed $j \geq 1$, when $n \to \infty$ we have

$$\begin{aligned} &\lambda_j(K_n^{(p)}) \nearrow M_{\mathbf{f}_p}, \qquad \lambda_j(M_n^{(p)}) \nearrow M_{\mathbf{h}_p}, \\ &\lambda_{np-j}(K_n^{(p)}) \searrow 0, \qquad \lambda_{np-j}(M_n^{(p)}) \searrow m_{\mathbf{h}_p}, \end{aligned} \tag{3.41}$$

where the convergence is monotone by Theorem 1.3 and by the fact that $K_n^{(p)}$ (resp. $M_n^{(p)}$) is a leading principal submatrix of $K_{n+1}^{(p)}$ (resp. $M_{n+1}^{(p)}$) for every n. Relation (3.41) says that, for fixed p, the matrix $K_n^{(p)}$ is ill-conditioned for large n, while $M_n^{(p)}$ is not (recall that $m_{\mathbf{h}_p} > 0$ by Corollary 3.2). Theorem 3.5 allows us to understand 'how much' $K_n^{(p)}$ is ill-conditioned. Before proving it, we provide two useful lemmas.

Lemma 3.4 (**Poincaré's inequality**). *For all $v \in H_0^1(0,1)$,*

$$\|v\|_{L^2(0,1)} \leq \frac{1}{\pi}\|v'\|_{L^2(0,1)}. \tag{3.42}$$

In [12] we find that $\dfrac{1}{\pi} = \sqrt{\dfrac{1}{c_{1,1}}}$ is the best constant such that (3.42) is satisfied for all $v \in H_0^1(0,1)$. Here, $c_{1,1}$ is the number appearing in (1.38) for $s = j = 1$; see also Remarks 1.3–1.5.

Lemma 3.5. *For all $p, n \geq 1$,*

$$K_n^{(p)} \geq \frac{\pi^2}{n^2} M_n^{(p)}. \tag{3.43}$$

Proof. By using the definition of $K_n^{(p)}$, see (3.19), for all $\mathbf{y} \in \mathbb{R}^{np-1}$ we have

$$\mathbf{y}^T \left(nK_n^{(p)}\right)\mathbf{y} = \sum_{i,j=1}^{np-1} y_i y_j \int_{(0,1)} \ell'_{j,(p)}\ell'_{i,(p)} = \int_{(0,1)} \sum_{i=1}^{np-1} y_i \ell'_{i,(p)} \sum_{j=1}^{np-1} y_j \ell'_{j,(p)} = \left\| \sum_{i=1}^{np-1} y_i \ell'_{i,(p)} \right\|^2_{L^2(0,1)} = \|v'_{\mathbf{y}}\|^2_{L^2(0,1)},$$

where $v_{\mathbf{y}} := \sum_{i=1}^{np-1} y_i \ell_{i,(p)} \in W_n^{(p)}$; see Section 3.1 for the definition of $W_n^{(p)}$. Similarly,

$$\mathbf{y}^T \left(\frac{1}{n} M_n^{(p)}\right)\mathbf{y} = \|v_{\mathbf{y}}\|^2_{L^2(0,1)}.$$

By the Poincaré inequality (3.42), we have $\|v'_{\mathbf{y}}\|^2_{L^2(0,1)} \geq \pi^2 \|v_{\mathbf{y}}\|^2_{L^2(0,1)}$. It follows that

$$\mathbf{y}^T \left(nK_n^{(p)}\right)\mathbf{y} \geq \pi^2 \mathbf{y}^T \left(\frac{1}{n} M_n^{(p)}\right)\mathbf{y},$$

i.e., the matrix inequality (3.43). □

Note that the argument shown in the proof of Lemma 3.5 is quite general and, in particular, it does not depend on the specific basis $\{\ell_{i,(p)} : i = 1, \dots, np-1\}$. In fact, a version of Lemma 3.5 holds in a more general setting; see [31, Proposition 1].

Theorem 3.5. *Let $p, n \geq 1$ and let $c_p > 0$ be a constant satisfying* (3.35)*. Then the following properties hold.*

1. *We have*

$$\lambda_j(K_n^{(p)}) \geq \max\left(\frac{\pi^2}{n^2}\lambda_j(M_n^{(p)}), c_p\lambda_{j+1}(T_n(2-2\cos\theta)\otimes I_p)\right) \qquad \forall\, j = 1, \dots, np-1, \tag{3.44}$$

$$\lambda_{\min}(K_n^{(p)}) \geq \max\left(\frac{\pi^2}{n^2}m_{\mathbf{h}_p}, 4c_p\sin^2\left(\frac{\pi}{2n+2}\right)\right) \overset{n\to\infty}{\sim} \frac{\pi^2\max(m_{\mathbf{h}_p}, c_p)}{n^2}. \tag{3.45}$$

2. *If $n \geq 3$, we have*

$$\lambda_{j+2}(C_n^{(p)}) \leq \lambda_j(K_n^{(p)}) \leq \lambda_{j-1}(C_n^{(p)}) \qquad \forall\, j = 1, \dots, np-1, \tag{3.46}$$

$$\lambda_{\min}(K_n^{(p)}) \leq 4\sin^2\left(\frac{\pi}{n}\right) \overset{n\to\infty}{\sim} \frac{4\pi^2}{n^2}, \tag{3.47}$$

where $C_n^{(p)}$ is the Hermitian block circulant matrix of order np defined in (3.48)*.*

Proof. 1. By Lemma 3.5 and by (1.10) it holds that $\lambda_j(K_n^{(p)}) \geq \frac{\pi^2}{n^2}\lambda_j(M_n^{(p)})$ for all $j = 1, \dots, np-1$. Moreover, by (3.35), for all θ we have

$$\mathbf{f}_p(\theta) \geq c_p(2-2\cos\theta)I_p.$$

By Proposition 1.1, this implies that

$$T_n(\mathbf{f}_p) \geq T_n(c_p(2-2\cos\theta)I_p) = c_pT_n(2-2\cos\theta)\otimes I_p,$$

where the last equality follows from the definitions of tensor product and $T_n(c_p(2-2\cos\theta)I_p)$. By (1.10) we deduce that

$$\lambda_j(T_n(\mathbf{f}_p)) \geq c_p\lambda_j(T_n(2-2\cos\theta)\otimes I_p) \qquad \forall\, j = 1, \dots, np,$$

and consequently, by (3.38),

$$\lambda_j(K_n^{(p)}) \geq c_p\lambda_{j+1}(T_n(2-2\cos\theta)\otimes I_p) \qquad \forall\, j = 1, \dots, np-1.$$

This completes the proof of (3.44). Relation (3.45) is obtained from (3.44) by setting $j = np-1$. To see this, note that $\lambda_{\min}(M_n^{(p)}) \geq m_{\mathbf{h}_p}$ by (3.40); moreover,

$$\lambda_{\min}(T_n(2-2\cos\theta)\otimes I_p) = \lambda_{\min}(T_n(2-2\cos\theta)) = 4\sin^2\left(\frac{\pi}{2n+2}\right),$$

where the last equality holds because the eigenvalues of $T_n(2-2\cos\theta)$ are known and, in particular, the minimal eigenvalue equals $2-2\cos\frac{\pi}{n+1} = 4\sin^2\left(\frac{\pi}{2n+2}\right)$; see Theorem 1.9.

2. Let $n \geq 3$. With the notation of Theorem 3.2, we have

$$T_n(\mathbf{f}_p) = \begin{bmatrix} K_0 & K_1^T & & \\ K_1 & \ddots & \ddots & \\ & \ddots & \ddots & K_1^T \\ & & K_1 & K_0 \end{bmatrix} = \begin{bmatrix} K_0 & K_1^T & & K_1 \\ K_1 & \ddots & \ddots & \\ & \ddots & \ddots & K_1^T \\ K_1^T & & K_1 & K_0 \end{bmatrix} - \begin{bmatrix} & & & K_1 \\ & & & \\ & & & \\ K_1^T & & & \end{bmatrix} =: C_n^{(p)} - E_n^{(p)}, \tag{3.48}$$

where $C_n^{(p)}$ is a block circulant matrix, while $E_n^{(p)}$ is Hermitian with $\operatorname{rank}(E_n^{(p)}) = 2$. The latter is true because $\operatorname{rank}(K_1) = 1$.[3] Therefore, $E_n^{(p)}$ has exactly two nonzero eigenvalues λ, μ, which are one the opposite of the other because $\lambda + \mu = \operatorname{trace}(E_n^{(p)}) = 0$. Thus, we can apply Theorem 1.4 with $k^+ = k^- = 1$ and we obtain

$$\lambda_{j-1}(C_n^{(p)}) \geq \lambda_j(T_n(\mathbf{f}_p)) \geq \lambda_{j+1}(C_n^{(p)}) \qquad \forall\, j = 1, \ldots, np. \tag{3.49}$$

The inequalities (3.46) follow from (3.49),(3.38). To obtain (3.47), note that the spectral decomposition of $C_n^{(p)}$ is known (Theorem 1.11) and, when applying Theorem 1.11 to $C_n^{(p)}$, the function $\mathbf{g}$ in (1.44) satisfies $\mathbf{g}\left(\frac{2\pi j}{n}\right) = \mathbf{f}_p\left(\frac{2\pi j}{n}\right)$ for all $j = 0, \ldots, n-1$. Moreover, by Corollary 3.1, $\lambda_{\min}(\mathbf{f}_p(\theta)) \leq 2 - 2\cos\theta$ for all θ, and $\lambda_{\min}(\mathbf{f}_p(\theta))$ is 'well-separated' from the other eigenvalue functions $\lambda_j(\mathbf{f}_p(\theta))$, $j = 1, \ldots, p-1$, by the separating line $\mu_{p-1}^{(p)}$. Hence, for $j = np - 1$, from (3.46) we obtain

$$\begin{aligned}
\lambda_{\min}(K_n^{(p)}) &\leq \lambda_{np-2}(C_n^{(p)}) = \text{the third smallest number in the set } \left\{\lambda_{\min}\left(\mathbf{f}_p\left(\tfrac{2\pi j}{n}\right)\right)\right\}_{j=0,\ldots,n-1} \\
&\leq \text{the third smallest number in the set } \left\{2 - 2\cos\tfrac{2\pi j}{n}\right\}_{j=0,\ldots,n-1} \\
&= 2 - 2\cos\tfrac{2\pi}{n} = 4\sin^2(\tfrac{\pi}{n}).
\end{aligned}$$

□

Remark 3.2. The argument used for proving (3.49) can be generalized to the case where $\mathbf{f}_p$ is replaced by any Hermitian matrix-valued trigonometric polynomial. To be precise, let $\mathbf{q}(\theta) = \sum_{k=-m}^{m} \mathbf{q}_k \mathrm{e}^{\mathrm{i}k\theta} : [-\pi, \pi] \to \mathbb{C}^{p\times p}$ be a Hermitian matrix-valued trigonometric polynomial. Then $\mathbf{q}_{-j} = \mathbf{q}_j^*$ for every $j = 0, \ldots, m$ and $T_n(\mathbf{q})$ is Hermitian for all $n \geq 1$ (see Subsection 1.4.1). For every $n \geq 2m+1$ we can write $T_n(\mathbf{q}) = C_n - E_n$, where $C_n := T_n(\mathbf{q}) + E_n$ is a block circulant matrix and the matrix E_n, given by

$$E_n := \begin{bmatrix} O & O & B \\ O & O & O \\ B^* & O & O \end{bmatrix}, \qquad B := \begin{bmatrix} \mathbf{q}_m & \cdots & \mathbf{q}_1 \\ & \ddots & \vdots \\ & & \mathbf{q}_m \end{bmatrix},$$

is Hermitian with $\operatorname{rank}(E_n) \leq 2mp$. It can be shown that the nonzero eigenvalues of E_n coincide with the nonzero singular values of B together with their negatives; see [7, p. 35]. Hence, E_n has the same number mp of positive and negative eigenvalues and so, by Theorem 1.4, we get

$$\lambda_{j-mp}(C_n) \geq \lambda_j(T_n(\mathbf{q})) \geq \lambda_{j+mp}(C_n), \qquad \forall\, j = 1, \ldots, np.$$

Notice also that the spectral decomposition of C_n for $n \geq 2m+1$ is given by (1.44) with

$$\mathbf{g}(\theta) = \sum_{k=0}^{m} \mathbf{q}_k \mathrm{e}^{\mathrm{i}k\theta} + \sum_{k=n-m}^{n-1} \mathbf{q}_{k-n} \mathrm{e}^{\mathrm{i}k\theta} = \sum_{k=0}^{m} \mathbf{q}_k \mathrm{e}^{\mathrm{i}k\theta} + \sum_{\ell=-m}^{-1} \mathbf{q}_\ell \mathrm{e}^{\mathrm{i}(\ell+n)\theta} = \sum_{k=0}^{m} \mathbf{q}_k \mathrm{e}^{\mathrm{i}k\theta} + \mathrm{e}^{\mathrm{i}n\theta} \sum_{\ell=-m}^{-1} \mathbf{q}_\ell \mathrm{e}^{\mathrm{i}\ell\theta},$$

hence $\mathbf{g}\left(\frac{2\pi j}{n}\right) = \mathbf{q}\left(\frac{2\pi j}{n}\right)$ for every $j = 0, \ldots, n-1$.

Table 3.1 shows the results of some numerical experiments. They confirm that $\lambda_{\min}(K_n^{(p)})$ goes to 0 as $1/n^2$ when $n \to \infty$, in accordance with Theorem 3.5, and they also allow us to formulate the following conjecture.

Conjecture 3.2. *For every $p, j \geq 1$ we have*

$$\lim_{n\to\infty} pn^2 \lambda_{np-j}(K_n^{(p)}) = j^2\pi^2, \tag{3.50}$$

where we recall that the eigenvalues of any Hermitian matrix like $K_n^{(p)}$ are arranged in non-increasing order, so that $\lambda_{np-j}(K_n^{(p)})$ is the j-th smallest eigenvalue of $K_n^{(p)}$.

[3] Note that $K_1 \neq O$, otherwise we would have $\langle L_0', L_1'\rangle = \ldots = \langle L_0', L_p'\rangle = 0$, implying $\langle L_0', L_1' + \ldots + L_p'\rangle = 0$ and, by Lemma 3.2, $-\langle L_0', L_0'\rangle = 0$: this is impossible, because L_0' is not identically 0.

n	$2n^2\lambda_{2n-1}(K_n^{(2)})$	$3n^2\lambda_{3n-1}(K_n^{(3)})$	$2n^2\lambda_{2n-2}(K_n^{(2)})$	$3n^2\lambda_{3n-2}(K_n^{(3)})$	$2n^2\lambda_{2n-3}(K_n^{(2)})$	$3n^2\lambda_{3n-3}(K_n^{(3)})$
20	9.8683332	9.8693541	39.4579402	39.4744220	88.7216045	88.8062922
40	9.8692871	9.8695418	39.4733327	39.4774163	88.8006247	88.8213755
80	9.8695251	9.8695887	39.4771485	39.4781671	88.8200104	88.8251718
160	9.8695846	9.8696005	39.4781005	39.4783550	88.8248338	88.8261226
320	9.8695994	9.8696034	39.4783383	39.4784019	88.8260383	88.8263603
640	9.8696032	9.8696042	39.4783978	39.4784137	88.8263393	88.8264198

Table 3.1: computation of $pn^2\lambda_{np-j}(K_n^{(p)})$ for $p = 2, 3$, $j = 1, 2, 3$, and for increasing values of n.

The limit relation (3.50) is verified for $p = 2, 3$ and $j = 1, 2, 3$ in Table 3.1. Moreover, it certainly holds for $p = 1$ and $j \geq 1$, since $K_n^{(1)} = T_{n-1}(2 - 2\cos\theta)$ and it is known that $\lambda_{n-j}(K_n^{(1)}) = 2 - 2\cos\frac{j\pi}{n}$, $j = 1, \ldots, n-1$; see Theorem 1.9.

Conjecture 3.2 can be motivated as follows. The matrix $K_n^{(p)}$ is associated with the Finite Element discretization of the 1D boundary value problem (1.40), because $nK_n^{(p)}$ coincides with the univariate $\mathbb{Q}_p$ Lagrangian FEM stiffness matrix $A_n^{(p)} = A_{n,D}^{(p)}$ in the case $\beta = \gamma = 0$; see (3.12)–(3.13) and the definition of $K_n^{(p)}$ in (3.19). The numbers $j^2\pi^2$, $j = 1, 2, \ldots$, are precisely the eigenvalues of (1.40); see Remark 1.4. The matrices $T_m(2 - 2\cos\theta)$, $m = 1, 2, \ldots$, are also associated with the (Finite Difference) discretization of (1.40) and for these matrices Theorem 1.10 estabilishes the analogous limit relation $\lim_{m\to\infty}\left(m^2\,\lambda_{m-j+1}(T_m(2-2\cos\theta))\right) = j^2\pi^2$ for each fixed $j \geq 1$; see Remark 1.5.

We now provide a localization of the spectrum of $A_{\boldsymbol{n}}^{(\boldsymbol{p})}$ and an estimate of its condition number under the assumption that $\boldsymbol{\beta} \in \mathbb{R}^d$ is constant. In this case, the advection matrix $A_{\boldsymbol{n},A}^{(\boldsymbol{p})}$ in (3.13) is skew-symmetric and, consequently, the real and imaginary parts of $A_{\boldsymbol{n}}^{(\boldsymbol{p})}$ are explicitly given by

$$\Re(A_{\boldsymbol{n}}^{(\boldsymbol{p})}) = A_{\boldsymbol{n},D}^{(\boldsymbol{p})} + A_{\boldsymbol{n},R}^{(\boldsymbol{p})}, \tag{3.51}$$

$$\Im(A_{\boldsymbol{n}}^{(\boldsymbol{p})}) = -\mathrm{i}A_{\boldsymbol{n},A}^{(\boldsymbol{p})}. \tag{3.52}$$

Note that, from (3.51), (3.15), (3.17)–(3.18), we obtain

$$\Re(A_{\boldsymbol{n}}^{(\boldsymbol{p})}) \geq \sum_{k=1}^{d} \frac{1}{n_1}M_{n_1}^{(p_1)} \otimes \cdots \otimes \frac{1}{n_{k-1}}M_{n_{k-1}}^{(p_{k-1})} \otimes n_k K_{n_k}^{(p_k)} \otimes \frac{1}{n_{k+1}}M_{n_{k+1}}^{(p_{k+1})} \otimes \cdots \otimes \frac{1}{n_d}M_{n_d}^{(p_d)} + \gamma_* \frac{1}{n_1}M_{n_1} \otimes \cdots \otimes \frac{1}{n_d}M_{n_d}, \tag{3.53}$$

$$\Re(A_{\boldsymbol{n}}^{(\boldsymbol{p})}) \leq \sum_{k=1}^{d} \frac{1}{n_1}M_{n_1}^{(p_1)} \otimes \cdots \otimes \frac{1}{n_{k-1}}M_{n_{k-1}}^{(p_{k-1})} \otimes n_k K_{n_k}^{(p_k)} \otimes \frac{1}{n_{k+1}}M_{n_{k+1}}^{(p_{k+1})} \otimes \cdots \otimes \frac{1}{n_d}M_{n_d}^{(p_d)} + \|\gamma\|_{L^\infty(\Omega)} \frac{1}{n_1}M_{n_1} \otimes \cdots \otimes \frac{1}{n_d}M_{n_d}. \tag{3.54}$$

In particular, (1.17) combined with Lemma 3.5 yields $\Re(A_{\boldsymbol{n}}^{(\boldsymbol{p})}) \geq \dfrac{\pi^2 d + \gamma_*}{n_1 \cdots n_d} M_{n_1}^{(p_1)} \otimes \cdots \otimes M_{n_d}^{(p_d)}$.

Lemma 3.6 (localization of the spectrum of $\Re(A_{\boldsymbol{n}}^{(\boldsymbol{p})})$). *Assume that $\boldsymbol{\beta} \in \mathbb{R}^d$ is constant and, for $p, n \geq 1$, define $\zeta_{n,p} := \max\left(\pi^2, \dfrac{4c_p}{m_{\mathbf{h}_p}} n^2 \sin\left(\dfrac{\pi}{2n+2}\right)\right)$, where $c_p > 0$ is a constant satisfying* (3.35)*. Then, for every $\boldsymbol{n}, \boldsymbol{p} \in \mathbb{N}^d$,*

$$\lambda_{\min}\left(\Re(A_{\boldsymbol{n}}^{(\boldsymbol{p})})\right) \geq \frac{\sum_{k=1}^{d} \zeta_{n_k,p_k} + \gamma_*}{n_1 \cdots n_d} G_{\boldsymbol{p}} \geq \frac{\pi^2 d + \gamma_*}{n_1 \cdots n_d} G_{\boldsymbol{p}}, \tag{3.55}$$

$$\lambda_{\max}(\Re(A_{\boldsymbol{n}}^{(\boldsymbol{p})})) \leq \frac{\sum_{k=1}^{d} n_k^2 (M_{\mathbf{f}_{p_k}}/M_{\mathbf{h}_{p_k}}) + \|\gamma\|_{L^\infty(\Omega)}}{n_1 \cdots n_d} S_{\boldsymbol{p}}, \tag{3.56}$$

where $G_{\boldsymbol{p}} := m_{\mathbf{h}_{p_1}} \cdots m_{\mathbf{h}_{p_d}}$ and $S_{\boldsymbol{p}} := M_{\mathbf{h}_{p_1}} \cdots M_{\mathbf{h}_{p_d}}$.

Proof. Apply (1.8),(1.16),(3.40),(3.45) in (3.53) to obtain (3.55). Then, apply (1.9),(1.16),(3.40) in (3.54) to obtain (3.56). □

Theorem 3.6 (**localization of the spectrum of** $A_{\boldsymbol{n}}^{(\boldsymbol{p})}$). *Assume that* $\boldsymbol{\beta} \in \mathbb{R}^d$ *is constant and, for* $p, n \geq 1$, *let* $\zeta_{n,p}$ *be as in Lemma 3.6. Then, for every* $\boldsymbol{n}, \boldsymbol{p} \in \mathbb{N}^d$,

$$\begin{aligned}\Lambda(A_{\boldsymbol{n}}^{(\boldsymbol{p})}) &\subseteq [\lambda_{\min}(\Re(A_{\boldsymbol{n}}^{(\boldsymbol{p})})), \lambda_{\max}(\Re(A_{\boldsymbol{n}}^{(\boldsymbol{p})}))] \times [\lambda_{\min}(\Im(A_{\boldsymbol{n}}^{(\boldsymbol{p})})), \lambda_{\max}(\Im(A_{\boldsymbol{n}}^{(\boldsymbol{p})}))] \\ &\subseteq \left[\frac{\sum_{k=1}^d \zeta_{n_k,p_k} + \gamma_*}{n_1 \cdots n_d} G_{\boldsymbol{p}}, \frac{\sum_{k=1}^d n_k^2 (M_{\mathbf{f}_{p_k}}/M_{\mathbf{h}_{p_k}}) + \|\gamma\|_{L^\infty(\Omega)}}{n_1 \cdots n_d} S_{\boldsymbol{p}}\right] \times \left[-B_{\boldsymbol{p}}\|\boldsymbol{\beta}\|_\infty \frac{\sum_{k=1}^d n_k}{n_1 \cdots n_d}, B_{\boldsymbol{p}}\|\boldsymbol{\beta}\|_\infty \frac{\sum_{k=1}^d n_k}{n_1 \cdots n_d}\right],\end{aligned}$$

where $G_{\boldsymbol{p}} := m_{\mathbf{h}_{p_1}} \cdots m_{\mathbf{h}_{p_d}}$, $S_{\boldsymbol{p}} := M_{\mathbf{h}_{p_1}} \cdots M_{\mathbf{h}_{p_d}}$, *and* $B_{\boldsymbol{p}}$ *is a constant satisfying* (3.14).

Proof. From Lemma 3.6 we have

$$\frac{\sum_{k=1}^d \zeta_{n_k,p_k} + \gamma_*}{n_1 \cdots n_d} G_{\boldsymbol{p}} \leq \lambda_{\min}(\Re(A_{\boldsymbol{n}}^{(\boldsymbol{p})})) \leq \lambda_{\max}(\Re(A_{\boldsymbol{n}}^{(\boldsymbol{p})})) \leq \frac{\sum_{k=1}^d n_k^2 (M_{\mathbf{f}_{p_k}}/M_{\mathbf{h}_{p_k}}) + \|\gamma\|_{L^\infty(\Omega)}}{n_1 \cdots n_d} S_{\boldsymbol{p}},$$

and from Lemma 3.1, combined with (3.52) and with the fact that $\boldsymbol{\beta}$ is constant, we have

$$-B_{\boldsymbol{p}}\|\boldsymbol{\beta}\|_\infty \frac{\sum_{k=1}^d n_k}{n_1 \cdots n_d} \leq -\|\Im(A_{\boldsymbol{n}}^{(\boldsymbol{p})})\| \leq \lambda_{\min}(\Im(A_{\boldsymbol{n}}^{(\boldsymbol{p})})) \leq \lambda_{\max}(\Im(A_{\boldsymbol{n}}^{(\boldsymbol{p})})) \leq \|\Im(A_{\boldsymbol{n}}^{(\boldsymbol{p})})\| \leq B_{\boldsymbol{p}}\|\boldsymbol{\beta}\|_\infty \frac{\sum_{k=1}^d n_k}{n_1 \cdots n_d}.$$

The thesis follows from (1.7). □

Theorem 3.7 (**conditioning**). *Assume that* $\boldsymbol{\beta}$ *is constant. Then, for every* $\boldsymbol{p} \in \mathbb{N}^d$ *there exists a constant* $\alpha_{\boldsymbol{p}}$ *such that, for all* $\boldsymbol{n} \in \mathbb{N}^d$,

$$\kappa(A_{\boldsymbol{n}}^{(\boldsymbol{p})}) \leq \alpha_{\boldsymbol{p}} \sum_{k=1}^d n_k^2. \tag{3.57}$$

Proof. From $A_{\boldsymbol{n}}^{(\boldsymbol{p})} = \Re(A_{\boldsymbol{n}}^{(\boldsymbol{p})}) + \mathrm{i}\Im(A_{\boldsymbol{n}}^{(\boldsymbol{p})})$ and from the fact that $\Re(A_{\boldsymbol{n}}^{(\boldsymbol{p})})$, $\Im(A_{\boldsymbol{n}}^{(\boldsymbol{p})})$ are Hermitian, we have

$$\sigma_{\max}(A_{\boldsymbol{n}}^{(\boldsymbol{p})}) = \|A_{\boldsymbol{n}}^{(\boldsymbol{p})}\| \leq \|\Re(A_{\boldsymbol{n}}^{(\boldsymbol{p})})\| + \|\Im(A_{\boldsymbol{n}}^{(\boldsymbol{p})})\| = \rho(\Re(A_{\boldsymbol{n}}^{(\boldsymbol{p})})) + \rho(\Im(A_{\boldsymbol{n}}^{(\boldsymbol{p})})).$$

Hence, by Theorem 3.6 we see that

$$\|A_{\boldsymbol{n}}^{(\boldsymbol{p})}\| \leq \hat{\alpha}_{\boldsymbol{p}} \frac{\sum_{k=1}^d n_k^2}{n_1 \cdots n_d},$$

for some constant $\hat{\alpha}_{\boldsymbol{p}}$ independent of $\boldsymbol{n}$. Furthermore, by Lemma 3.6 and by the Fan-Hoffman theorem,

$$\sigma_{\min}(A_{\boldsymbol{n}}^{(\boldsymbol{p})}) \geq \lambda_{\min}(\Re(A_{\boldsymbol{n}}^{(\boldsymbol{p})})) \geq \frac{\tilde{\alpha}_{\boldsymbol{p}}}{n_1 \cdots n_d},$$

for some constant $\tilde{\alpha}_{\boldsymbol{p}} > 0$ independent of $\boldsymbol{n}$. Thus, $\kappa(A_{\boldsymbol{n}}^{(\boldsymbol{p})}) = \frac{\sigma_{\max}(A_{\boldsymbol{n}}^{(\boldsymbol{p})})}{\sigma_{\min}(A_{\boldsymbol{n}}^{(\boldsymbol{p})})} \leq \alpha_{\boldsymbol{p}} \sum_{k=1}^d n_k^2$, with $\alpha_{\boldsymbol{p}} = \hat{\alpha}_{\boldsymbol{p}}/\tilde{\alpha}_{\boldsymbol{p}}$. □

(3.57) says that $\kappa(A_{\boldsymbol{n}}^{(\boldsymbol{p})})$ is bounded from above by $\max(\boldsymbol{n}^2) = \max(n_1^2, \ldots, n_d^2)$ multiplied by some constant independent of $\boldsymbol{n}$ (for instance $\alpha_{\boldsymbol{p}} d$). This upper bound is the sharpest possible, as shown by the numerical experiments in Table 3.2, where we fixed $d = 2$, $\boldsymbol{\beta} = \mathbf{0}$, $\gamma = 0$, $\boldsymbol{p} = (2, 2)$, and we computed $\kappa(A_{\boldsymbol{n}}^{(\boldsymbol{p})}) = \kappa(A_{\boldsymbol{n},D}^{(\boldsymbol{p})})$ (normalized by n^2) for $\boldsymbol{n} = (n, \log_2 n)$ and for increasing values of n. For a nice comparison with Finite Differences (FD), in the third column of Table 3.2 we reported the values of $\kappa(A_{\boldsymbol{n}})/n^2$ for $d = 2$ and for $\boldsymbol{n} = (n, \log_2 n)$, where

$$A_{\boldsymbol{n}} := \sum_{k=1}^d n_k^2 \, I_{n_1-1} \otimes \cdots \otimes I_{n_{k-1}-1} \otimes T_{n_k-1}(2 - 2\cos\theta) \otimes I_{n_{k+1}-1} \otimes \cdots \otimes I_{n_d-1} = T_{\boldsymbol{n}-1}\left(\sum_{k=1}^d n_k^2 (2 - 2\cos\theta_k)\right) \tag{3.58}$$

is the (diffusion) matrix coming from the standard centered FD approximation of (3.1) on the mesh $\boldsymbol{j}/\boldsymbol{n}$, $\boldsymbol{j} = \mathbf{0}, \ldots, \boldsymbol{n}$, in the case $\boldsymbol{\beta} = \mathbf{0}$, $\gamma = 0$.

n	$\kappa(A_{\boldsymbol{n}}^{(\boldsymbol{p})})/n^2$	$\kappa(A_{\boldsymbol{n}})/n^2$
8	1.2597	0.2278
16	1.2573	0.2173
32	1.2553	0.2101
64	1.2543	0.2065
128	1.2539	0.2049
256	1.2538	0.2041

Table 3.2: computation of $\kappa(A_{\boldsymbol{n}}^{(\boldsymbol{p})})/n^2$ and $\kappa(A_{\boldsymbol{n}})/n^2$ in the case $d = 2$, $\boldsymbol{\beta} = \mathbf{0}$, $\gamma = 0$, $\boldsymbol{p} = (2,2)$, $\boldsymbol{n} = (n, \log_2 n)$, for increasing values of n. Note that we are in the presence of a non-uniform mesh refinement.

3.4.2 Spectral distribution and symbol of the normalized sequence $\{n^{d-2}A_{\boldsymbol{n}}^{(\boldsymbol{p})}\}_n$

In this subsection we assume that $n_j = \nu_j n$ for all $j = 1, \dots, d$, i.e. $\boldsymbol{n} = \boldsymbol{\nu} n = (\nu_1 n, \dots, \nu_d n) \in \mathbb{N}^d$, where $\boldsymbol{\nu} \in \mathbb{Q}_+^d$ is fixed and n varies in the set of natural numbers such that $\boldsymbol{n} \in \mathbb{N}^d$. Under this assumption, from (3.12)–(3.13) and (3.17) we have

$$\begin{aligned} n^{d-2}A_{\boldsymbol{n}}^{(\boldsymbol{p})} &= n^{d-2}A_{\boldsymbol{n},D}^{(\boldsymbol{p})} + n^{d-2}A_{\boldsymbol{n},A}^{(\boldsymbol{p})} + n^{d-2}A_{\boldsymbol{n},R}^{(\boldsymbol{p})} \\ &= \sum_{k=1}^{d} c_k(\boldsymbol{\nu})\, M_{n_1}^{(p_1)} \otimes \cdots \otimes M_{n_{k-1}}^{(p_{k-1})} \otimes K_{n_k}^{(p_k)} \otimes M_{n_{k+1}}^{(p_{k+1})} \otimes \cdots \otimes M_{n_d}^{(p_d)} + n^{d-2}A_{\boldsymbol{n},A}^{(\boldsymbol{p})} + n^{d-2}A_{\boldsymbol{n},R}^{(\boldsymbol{p})}, \end{aligned} \tag{3.59}$$

where the values $c_k(\boldsymbol{\nu})$, $k = 1, \dots, d$, are given in (3.37). Recall from (3.11) that $A_{\boldsymbol{n}}^{(\boldsymbol{p})}$ is of size $N(\boldsymbol{np} - \mathbf{1}) = (n_1 p_1 - 1) \cdots (n_d p_d - 1)$.

In Theorem 3.8 we prove that the sequence of matrices $\{n^{d-2}A_{\boldsymbol{n}}^{(\boldsymbol{p})}\}_n$ in (3.59) is distributed, in the sense of the eigenvalues, like the Hermitian matrix-valued function $\mathbf{f}_{\boldsymbol{p}}^{(\boldsymbol{\nu})}$ in (3.36), which is therefore the symbol of the sequence $\{n^{d-2}A_{\boldsymbol{n}}^{(\boldsymbol{p})}\}_n$. Note that $\{n^{d-2}A_{\boldsymbol{n}}^{(\boldsymbol{p})}\}_n$ is really a sequence of matrices, due to the assumption $\boldsymbol{n} = \boldsymbol{\nu} n$. This assumption must be kept in mind while reading this subsection.

Before stating and proving Theorem 3.8, let us observe that, by the properties of $\mathbf{f}_p(\theta)$ and $\mathbf{h}_p(\theta)$, see Corollaries 3.1–3.2, and by the properties of tensor products, see Subsection 1.2.1, $\mathbf{f}_{\boldsymbol{p}}^{(\boldsymbol{\nu})}(\boldsymbol{\theta}) \geq O$ for all $\boldsymbol{\theta} \in [-\pi, \pi]^d$ and $\mathbf{f}_{\boldsymbol{p}}^{(\boldsymbol{\nu})}(\boldsymbol{\theta}) > O$ for all $\boldsymbol{\theta} \in [-\pi, \pi]^d \backslash \{\mathbf{0}\}$.

Theorem 3.8. *Let $\boldsymbol{p} \in \mathbb{N}^d$, $\boldsymbol{\nu} \in \mathbb{Q}_+^d$ and $\boldsymbol{n} = \boldsymbol{\nu} n$, then $\{n^{d-2}A_{\boldsymbol{n}}^{(\boldsymbol{p})}\}_n \sim_\lambda \mathbf{f}_{\boldsymbol{p}}^{(\boldsymbol{\nu})}$. In particular, $\{n^{d-2}A_{\boldsymbol{n}}^{(\boldsymbol{p})}\}_n$ is weakly clustered at the essential range $\mathcal{ER}(\mathbf{f}_{\boldsymbol{p}}^{(\boldsymbol{\nu})})$ and every point $z \in \mathcal{ER}(\mathbf{f}_{\boldsymbol{p}}^{(\boldsymbol{\nu})})$ strongly attracts $\Lambda(n^{d-2}A_{\boldsymbol{n}}^{(\boldsymbol{p})})$ with infinite order (see Theorem 1.5).*

Proof. For all $p, n \geq 1$, define the following matrices, of size np:

$$\tilde{K}_n^{(p)} := K_n^{(p)} \oplus [0], \qquad \tilde{M}_n^{(p)} := M_n^{(p)} \oplus [0].$$

Let $n^{d-2}\tilde{A}_{\boldsymbol{n},D}^{(\boldsymbol{p})}$ be the matrix of size $N(\boldsymbol{np}) = n_1 p_1 \cdots n_d p_d = (\nu_1 p_1 \cdots \nu_d p_d) n^d = N(\boldsymbol{\nu p}) n^d$ obtained from $n^{d-2}A_{\boldsymbol{n},D}^{(\boldsymbol{p})}$ by replacing the symbols K, M appearing in its expression (3.59) with $\tilde{K}, \tilde{M}$:

$$n^{d-2}\tilde{A}_{\boldsymbol{n},D}^{(\boldsymbol{p})} = \sum_{k=1}^{d} c_k(\boldsymbol{\nu})\, \tilde{M}_{n_1}^{(p_1)} \otimes \cdots \otimes \tilde{M}_{n_{k-1}}^{(p_{k-1})} \otimes \tilde{K}_{n_k}^{(p_k)} \otimes \tilde{M}_{n_{k+1}}^{(p_{k+1})} \otimes \cdots \otimes \tilde{M}_{n_d}^{(p_d)}.$$

By Lemma 1.5, there exists the permutation matrix $P_{\boldsymbol{n},\boldsymbol{p}} := P_{n_1 p_1 - 1, 1, n_2 p_2 - 1, 1, \dots, n_d p_d - 1, 1}$, depending only on $\boldsymbol{n}, \boldsymbol{p}$, such that

$$n^{d-2}\tilde{A}_{\boldsymbol{n},D}^{(\boldsymbol{p})} = P_{\boldsymbol{n},\boldsymbol{p}}[(n^{d-2}A_{\boldsymbol{n},D}^{(\boldsymbol{p})}) \oplus O]P_{\boldsymbol{n},\boldsymbol{p}}^T,$$

where O is the zero matrix of order $n_1 p_1 \cdots n_d p_d - (n_1 p_1 - 1) \cdots (n_d p_d - 1) = o(n^d)$. Hence,

$$n^{d-2}\tilde{A}_{\boldsymbol{n}}^{(\boldsymbol{p})} := P_{\boldsymbol{n},\boldsymbol{p}} \left[(n^{d-2}A_{\boldsymbol{n}}^{(\boldsymbol{p})}) \oplus O \right] P_{\boldsymbol{n},\boldsymbol{p}}^T = P_{\boldsymbol{n},\boldsymbol{p}}[n^{d-2}A_{\boldsymbol{n},D}^{(\boldsymbol{p})} \oplus O + n^{d-2}A_{\boldsymbol{n},A}^{(\boldsymbol{p})} \oplus O + n^{d-2}A_{\boldsymbol{n},R}^{(\boldsymbol{p})} \oplus O]P_{\boldsymbol{n},\boldsymbol{p}}^T$$

$$= n^{d-2}\tilde{A}^{(\boldsymbol{p})}_{\boldsymbol{n},D} + n^{d-2}\tilde{A}^{(\boldsymbol{p})}_{\boldsymbol{n},A} + n^{d-2}\tilde{A}^{(\boldsymbol{p})}_{\boldsymbol{n},R},$$

where $n^{d-2}\tilde{A}^{(\boldsymbol{p})}_{\boldsymbol{n},A} := P_{\boldsymbol{n},\boldsymbol{p}}[(n^{d-2}A^{(\boldsymbol{p})}_{\boldsymbol{n},A}) \oplus O]P^T_{\boldsymbol{n},\boldsymbol{p}}$ and $n^{d-2}\tilde{A}^{(\boldsymbol{p})}_{\boldsymbol{n},R} := P_{\boldsymbol{n},\boldsymbol{p}}[(n^{d-2}A^{(\boldsymbol{p})}_{\boldsymbol{n},R}) \oplus O]P^T_{\boldsymbol{n},\boldsymbol{p}}$. The eigenvalues of $n^{d-2}\tilde{A}^{(\boldsymbol{p})}_{\boldsymbol{n}}$ are those of $n^{d-2}A^{(\boldsymbol{p})}_{\boldsymbol{n}}$ with only $o(n^d)$ extra eigenvalues equal to 0. Consequently, by Definition 1.1, if we prove that $\{n^{d-2}\tilde{A}^{(\boldsymbol{p})}_{\boldsymbol{n}}\}_n \sim_\lambda \mathbf{f}^{(\boldsymbol{\nu})}_{\boldsymbol{p}}$ then $\{n^{d-2}A^{(\boldsymbol{p})}_{\boldsymbol{n}}\}_n \sim_\lambda \mathbf{f}^{(\boldsymbol{\nu})}_{\boldsymbol{p}}$.

Now, let

$$T^{(\boldsymbol{p})}_{\boldsymbol{n}} := \sum_{k=1}^{d} c_k(\boldsymbol{\nu})\, T_{n_1}(\mathbf{h}_{p_1}) \otimes \cdots \otimes T_{n_{k-1}}(\mathbf{h}_{p_{k-1}}) \otimes T_{n_k}(\mathbf{f}_{p_k}) \otimes T_{n_{k+1}}(\mathbf{h}_{p_{k+1}}) \otimes \cdots \otimes T_{n_d}(\mathbf{h}_{p_d}). \tag{3.60}$$

To show that $\{n^{d-2}\tilde{A}^{(\boldsymbol{p})}_{\boldsymbol{n}}\}_n \sim_\lambda \mathbf{f}^{(\boldsymbol{\nu})}_{\boldsymbol{p}}$, we prove that the hypotheses of Theorem 2.7 are satisfied with $X_n := T^{(\boldsymbol{p})}_{\boldsymbol{n}}$, $Y_n := n^{d-2}\tilde{A}^{(\boldsymbol{p})}_{\boldsymbol{n}} - T^{(\boldsymbol{p})}_{\boldsymbol{n}}$ and $\mathbf{f} = \mathbf{f}^{(\boldsymbol{\nu})}_{\boldsymbol{p}}$.

Note that each $T^{(\boldsymbol{p})}_{\boldsymbol{n}}$ is Hermitian because $\mathbf{f}_p$, $\mathbf{h}_p$ are Hermitian matrix-valued functions for all $p \geq 1$. By Lemma 1.9, $T^{(\boldsymbol{p})}_{\boldsymbol{n}}$ is also similar to $T_{\boldsymbol{n}}(\mathbf{f}^{(\boldsymbol{\nu})}_{\boldsymbol{p}})$, and, by Theorem 1.8, $\{T_{\boldsymbol{n}}(\mathbf{f}^{(\boldsymbol{\nu})}_{\boldsymbol{p}})\}_n \sim_\lambda \mathbf{f}^{(\boldsymbol{\nu})}_{\boldsymbol{p}}$, implying $\{T^{(\boldsymbol{p})}_{\boldsymbol{n}}\}_n \sim_\lambda \mathbf{f}^{(\boldsymbol{\nu})}_{\boldsymbol{p}}$. Now observe that, since $K^{(p)}_n$, $M^{(p)}_n$, $\tilde{K}^{(p)}_n$, $\tilde{M}^{(p)}_n$, $T_n(\mathbf{f}_p)$, $T_n(\mathbf{h}_p)$ are normal for all $p, n \geq 1$, we have

$$\|\tilde{K}^{(p)}_n\| = \rho(\tilde{K}^{(p)}_n) = \rho(K^{(p)}_n) = \|K^{(p)}_n\| \leq M_{\mathbf{f}_p}, \qquad \|T_n(\mathbf{f}_p)\| = \rho(T_n(\mathbf{f}_p)) \leq M_{\mathbf{f}_p}, \tag{3.61}$$

$$\|\tilde{M}^{(p)}_n\| = \rho(\tilde{M}^{(p)}_n) = \rho(M^{(p)}_n) = \|M^{(p)}_n\| \leq M_{\mathbf{h}_p}, \qquad \|T_n(\mathbf{h}_p)\| = \rho(T_n(\mathbf{h}_p)) \leq M_{\mathbf{h}_p}. \tag{3.62}$$

From (3.61)–(3.62), from the triangle inequality, and from (1.13), it follows that the norms $\|T^{(\boldsymbol{p})}_{\boldsymbol{n}}\|$, $\|n^{d-2}\tilde{A}^{(\boldsymbol{p})}_{\boldsymbol{n},D}\| = \|n^{d-2}A^{(\boldsymbol{p})}_{\boldsymbol{n},D}\|$ are bounded from above by some constant independent of n. Moreover, from Lemma 3.1 and (3.16), (3.18), (3.62), (1.13), we have

$$\|n^{d-2}\tilde{A}^{(\boldsymbol{p})}_{\boldsymbol{n},A}\| = \|n^{d-2}A^{(\boldsymbol{p})}_{\boldsymbol{n},A}\| \leq \frac{n^{d-2}B_{\boldsymbol{p}}\|\boldsymbol{\beta}\|_{L^\infty(\Omega)}\sum_{k=1}^{d} n_k}{n_1 \cdots n_d} = \frac{B_{\boldsymbol{p}}\|\boldsymbol{\beta}\|_{L^\infty(\Omega)}\sum_{k=1}^{d} \nu_k}{\nu_1 \cdots \nu_d n}, \tag{3.63}$$

$$\|n^{d-2}\tilde{A}^{(\boldsymbol{p})}_{\boldsymbol{n},R}\| = \|n^{d-2}A^{(\boldsymbol{p})}_{\boldsymbol{n},R}\| \leq \frac{n^{d-2}\|\gamma\|_{L^\infty(\Omega)}S_{\boldsymbol{p}}}{n_1 \cdots n_d} = \frac{\|\gamma\|_{L^\infty(\Omega)}S_{\boldsymbol{p}}}{\nu_1 \cdots \nu_d n^2}, \tag{3.64}$$

where $S_{\boldsymbol{p}} := M_{\mathbf{h}_{p_1}} \cdots M_{\mathbf{h}_{p_d}}$. Therefore, taking into account the triangle inequality

$$\|n^{d-2}\tilde{A}^{(\boldsymbol{p})}_{\boldsymbol{n}}\| \leq \|n^{d-2}\tilde{A}^{(\boldsymbol{p})}_{\boldsymbol{n},D}\| + \|n^{d-2}\tilde{A}^{(\boldsymbol{p})}_{\boldsymbol{n},A}\| + \|n^{d-2}\tilde{A}^{(\boldsymbol{p})}_{\boldsymbol{n},R}\|,$$

we conclude that $\|n^{d-2}\tilde{A}^{(\boldsymbol{p})}_{\boldsymbol{n}}\|$ is bounded from above by some constant independent of n. Hence,

$$\|T^{(\boldsymbol{p})}_{\boldsymbol{n}}\|,\ \|n^{d-2}\tilde{A}^{(\boldsymbol{p})}_{\boldsymbol{n},D}\|,\ \|n^{d-2}\tilde{A}^{(\boldsymbol{p})}_{\boldsymbol{n},A}\|,\ \|n^{d-2}\tilde{A}^{(\boldsymbol{p})}_{\boldsymbol{n},R}\|,\ \|n^{d-2}\tilde{A}^{(\boldsymbol{p})}_{\boldsymbol{n}}\|,\ \|n^{d-2}\tilde{A}^{(\boldsymbol{p})}_{\boldsymbol{n}} - T^{(\boldsymbol{p})}_{\boldsymbol{n}}\| \leq C, \tag{3.65}$$

for some C independent of n. To finish the proof, we have to show that $\left|\!\left|\!\left| n^{d-2}\tilde{A}^{(\boldsymbol{p})}_{\boldsymbol{n}} - T^{(\boldsymbol{p})}_{\boldsymbol{n}} \right|\!\right|\!\right|_1 = o(n^d)$ as $n \to \infty$. Note that, for all $p, n \geq 1$,

$$\operatorname{rank}(\tilde{K}^{(p)}_n - T_n(\mathbf{f}_p)) \leq 2, \qquad \operatorname{rank}(\tilde{M}^{(p)}_n - T_n(\mathbf{h}_p)) \leq 2.$$

Therefore, by (1.5) and by the property (1.18) of tensor products we infer

$$\begin{aligned}
\left|\!\left|\!\left| n^{d-2}\tilde{A}^{(\boldsymbol{p})}_{\boldsymbol{n}} - T^{(\boldsymbol{p})}_{\boldsymbol{n}} \right|\!\right|\!\right|_1 &\leq \left|\!\left|\!\left| n^{d-2}\tilde{A}^{(\boldsymbol{p})}_{\boldsymbol{n},D} - T^{(\boldsymbol{p})}_{\boldsymbol{n}} \right|\!\right|\!\right|_1 + \left|\!\left|\!\left| n^{d-2}\tilde{A}^{(\boldsymbol{p})}_{\boldsymbol{n},A} \right|\!\right|\!\right|_1 + \left|\!\left|\!\left| n^{d-2}\tilde{A}^{(\boldsymbol{p})}_{\boldsymbol{n},R} \right|\!\right|\!\right|_1 \\
&\leq \left(d \sum_{i=1}^{d} 2n_1p_1 \cdots n_{i-1}p_{i-1}n_{i+1}p_{i+1} \cdots n_d p_d \right) \left\| n^{d-2}\tilde{A}^{(\boldsymbol{p})}_{\boldsymbol{n},D} - T^{(\boldsymbol{p})}_{\boldsymbol{n}} \right\| \\
&\qquad + n_1p_1 \cdots n_d p_d \left\| n^{d-2}\tilde{A}^{(\boldsymbol{p})}_{\boldsymbol{n},A} \right\| + n_1p_1 \cdots n_d p_d \left\| n^{d-2}\tilde{A}^{(\boldsymbol{p})}_{\boldsymbol{n},R} \right\|,
\end{aligned}$$

and the latter is $o(n^d)$, thanks to (3.63)–(3.65). □

p	2	3	4	5	6	7	8	9	10
$\phi^{(1)}_{p,1}$	1.33	5.78	$1.84 \cdot 10$	$5.45 \cdot 10$	$1.59 \cdot 10^2$	$4.84 \cdot 10^2$	$1.54 \cdot 10^3$	$5.12 \cdot 10^3$	$1.77 \cdot 10^4$
$[\phi^{(1)}_{p,1}]^{1/p}$	1.15	1.79	2.07	2.22	2.33	2.42	2.50	2.58	2.66

Table 3.3: computation of $\phi^{(\boldsymbol{\nu})}_{\boldsymbol{p},d}$ and $[\phi^{(\boldsymbol{\nu})}_{\boldsymbol{p},d}]^{1/N(\boldsymbol{p})}$ in the case $d = 1$, $\boldsymbol{p} = p$, $\boldsymbol{\nu} = 1$, for $p = 2, \ldots, 10$. Note that in this case $\mathbf{f}^{(\boldsymbol{\nu})}_{\boldsymbol{p}}(\boldsymbol{\theta})$ is nothing else than $\mathbf{f}_p(\theta)$.

p	2	3	4	5	6	7	8	9	10
$\phi^{(1,1)}_{(p,p),2}$	2.13	9.72	$3.44 \cdot 10$	$1.54 \cdot 10^2$	$7.47 \cdot 10^2$	$4.39 \cdot 10^3$	$3.01 \cdot 10^4$	$2.42 \cdot 10^5$	$2.17 \cdot 10^6$
$[\phi^{(1,1)}_{(p,p),2}]^{1/p^2}$	1.21	1.29	1.25	1.22	1.20	1.19	1.17	1.17	1.16

Table 3.4: computation of $\phi^{(\boldsymbol{\nu})}_{\boldsymbol{p},d}$ and $[\phi^{(\boldsymbol{\nu})}_{\boldsymbol{p},d}]^{1/N(\boldsymbol{p})}$ in the case $d = 2$, $\boldsymbol{p} = (p, p)$, $\boldsymbol{\nu} = (1, 1)$, for $p = 2, \ldots, 10$.

3.4.3 Exponential scattering and ill-conditioning of the symbol

The discussion on the exponential ill-conditioning of the symbol contained in this subsection is based on the informal meaning behind the definition of spectral distribution. According to Remark 1.2, the spectral information contained in the symbol $\mathbf{f}^{(\boldsymbol{\nu})}_{\boldsymbol{p}}$ can be summarized as follows: the eigenvalues of $n^{d-2}A^{(\boldsymbol{p})}_{\boldsymbol{n}}$ are approximately given by a uniform sampling of the eigenvalue functions $\lambda_i(\mathbf{f}^{(\boldsymbol{\nu})}_{\boldsymbol{p}})$ over an equispaced grid in the domain $[-\pi, \pi]^d$. To fix the ideas, assume that the equispaced grid is

$$-\boldsymbol{\pi} + \frac{2\boldsymbol{j}\boldsymbol{\pi}}{\boldsymbol{n}} = \left(-\pi + \frac{2j_1\pi}{n_1}, \ldots, -\pi + \frac{2j_d\pi}{n_d}\right), \quad \boldsymbol{j} = \mathbf{0}, \ldots, \boldsymbol{n} - \mathbf{1},$$

where $\boldsymbol{\pi} := (\pi, \ldots, \pi)$. Then, the eigenvalues of $n^{d-2}A^{(\boldsymbol{p})}_{\boldsymbol{n}}$ are approximately given by[4]

$$\lambda_i\left(\mathbf{f}^{(\boldsymbol{\nu})}_{\boldsymbol{p}}\left(-\boldsymbol{\pi} + \frac{2\boldsymbol{j}\boldsymbol{\pi}}{\boldsymbol{n}}\right)\right), \quad \boldsymbol{j} = \mathbf{0}, \ldots, \boldsymbol{n} - \mathbf{1}, \quad i = 1, \ldots, N(\boldsymbol{p}). \tag{3.66}$$

From (3.66) we infer that the ratio

$$\phi^{(\boldsymbol{\nu})}_{\boldsymbol{p},d} := \frac{\min_{\boldsymbol{\theta} \in [-\pi,\pi]^d} \lambda_{\max}(\mathbf{f}^{(\boldsymbol{\nu})}_{\boldsymbol{p}}(\boldsymbol{\theta}))}{\max_{\boldsymbol{\theta} \in [-\pi,\pi]^d} \lambda_{\min}(\mathbf{f}^{(\boldsymbol{\nu})}_{\boldsymbol{p}}(\boldsymbol{\theta}))}$$

is an index of the scattering of the eigenvalues of $n^{d-2}A^{(\boldsymbol{p})}_{\boldsymbol{n}}$. Indeed, if $\phi^{(\boldsymbol{\nu})}_{\boldsymbol{p},d}$ is large (resp. small), then the eigenvalues of $n^{d-2}A^{(\boldsymbol{p})}_{\boldsymbol{n}}$ obtained from (3.66) for $i = 1$, which correspond to the maximal eigenvalue function of the symbol, are far away from (resp. very close to) the eigenvalues obtained for $i = N(\boldsymbol{p})$, which correspond to the minimal eigenvalue function of the symbol. Furthermore, in the case where $\phi^{(\boldsymbol{\nu})}_{\boldsymbol{p},d}$ is large, the 'ill-conditioned subspace', that is the subspace corresponding to the largest eigenvalues of $n^{d-2}A^{(\boldsymbol{p})}_{\boldsymbol{n}}$ obtained by setting $i = 1$ in (3.66), is very large: its dimension is about

$$\#\left\{\lambda_{\max}\left(\mathbf{f}^{(\boldsymbol{\nu})}_{\boldsymbol{p}}\left(-\boldsymbol{\pi} + \frac{2\boldsymbol{j}\boldsymbol{\pi}}{\boldsymbol{n}}\right)\right) : \ \boldsymbol{j} = \mathbf{0}, \ldots, \boldsymbol{n} - \mathbf{1}\right\} = \frac{N(\boldsymbol{n})}{N(\boldsymbol{p})} = \frac{n_1 \cdots n_d}{p_1 \cdots p_d}. \tag{3.67}$$

Tables 3.3–3.4 shows, for $d = 1, 2$, the behavior of $\phi^{(\boldsymbol{\nu})}_{\boldsymbol{p},d}$ in the case $\boldsymbol{p} = (p, \ldots, p)$, $\boldsymbol{\nu} = (1, \ldots, 1)$, for different values of p. Not only we observe an exponential ill-conditioning with p and d, as already proved in [44], but we can also predict, on the base of (3.67), that the subspace where this exponential ill-conditioning occurs is very large: for the case displayed in Tables 3.3–3.4, the size of such subspace is approximately

[4] Ignore the mismatch with the size of $n^{d-2}A^{(\boldsymbol{p})}_{\boldsymbol{n}}$: the reasoning that we are following in this subsection is heuristic. Think of $n^{d-2}A^{(\boldsymbol{p})}_{\boldsymbol{n}}$ as if it were exactly the Toeplitz matrix $T_{\boldsymbol{n}}(\mathbf{f}^{(\boldsymbol{\nu})}_{\boldsymbol{p}})$ generated by the symbol $\mathbf{f}^{(\boldsymbol{\nu})}_{\boldsymbol{p}}$.

n^d/p^d, $d = 1, 2$. This involved picture shows that the numerical solution of the linear systems associated with the matrix $n^{d-2}A_{\boldsymbol{n}}^{(\boldsymbol{p})}$ is a really hard problem for large $\boldsymbol{p}$ and d, not only because of the exponential ill-conditioning, but also for the large size of the subspace where this ill-conditioning is attained.

3.4.4 Clustering of the normalized sequence $\{n^{d-2}A_{\boldsymbol{n}}^{(\boldsymbol{p})}\}_n$

In this subsection, we still assume that $\boldsymbol{n} = \boldsymbol{\nu}n$, where $\boldsymbol{\nu} \in \mathbb{Q}_+^d$ is fixed and n varies in the set of natural numbers such that $\boldsymbol{n} \in \mathbb{N}^d$. In this situation, we have seen in Theorem 3.8 that $\{n^{d-2}A_{\boldsymbol{n}}^{(\boldsymbol{p})}\}_n \sim_\lambda \mathbf{f}_{\boldsymbol{p}}^{(\boldsymbol{\nu})}$ and, consequently, $\{n^{d-2}A_{\boldsymbol{n}}^{(\boldsymbol{p})}\}_n$ is weakly clustered at the essential range of $\mathbf{f}_{\boldsymbol{p}}^{(\boldsymbol{\nu})}$, given by the union of the essential ranges of the eigenvalue functions $\lambda_i(\mathbf{f}_{\boldsymbol{p}}^{(\boldsymbol{\nu})})$, $i = 1, \ldots, N(\boldsymbol{p})$, that is $\mathcal{ER}(\mathbf{f}_{\boldsymbol{p}}^{(\boldsymbol{\nu})}) = \bigcup_{i=1}^{N(\boldsymbol{p})} \mathcal{ER}(\lambda_i(\mathbf{f}_{\boldsymbol{p}}^{(\boldsymbol{\nu})}))$; see Theorem 1.5. Note that $\mathbf{f}_{\boldsymbol{p}}^{(\boldsymbol{\nu})}$ is continuous over $[-\pi, \pi]^d$, hence the eigenvalue functions are continuous over $[-\pi, \pi]^d$, which means that their essential ranges coincide exactly with their images. Being weakly clustered at $\mathcal{ER}(\mathbf{f}_{\boldsymbol{p}}^{(\boldsymbol{\nu})})$, the sequence $\{n^{d-2}A_{\boldsymbol{n}}^{(\boldsymbol{p})}\}_n$ is a fortiori weakly clustered at the convex hull of $\mathcal{ER}(\mathbf{f}_{\boldsymbol{p}}^{(\boldsymbol{\nu})})$, which is given by $[0, M_{\mathbf{f}_{\boldsymbol{p}}^{(\boldsymbol{\nu})}}]$, $M_{\mathbf{f}_{\boldsymbol{p}}^{(\boldsymbol{\nu})}} := \max_{\boldsymbol{\theta} \in [-\pi,\pi]^d} \mathbf{f}_{\boldsymbol{p}}^{(\boldsymbol{\nu})}(\boldsymbol{\theta})$. We are going to see that actually $\{n^{d-2}A_{\boldsymbol{n}}^{(\boldsymbol{p})}\}_n$ is strongly clustered at $[0, M_{\mathbf{f}_{\boldsymbol{p}}^{(\boldsymbol{\nu})}}]$, in the case where $\boldsymbol{\beta}$ is constant.

Theorem 3.9. *We have $\Lambda(n^{d-2}A_{\boldsymbol{n},D}^{(\boldsymbol{p})}) \subset (0, M_{\mathbf{f}_{\boldsymbol{p}}^{(\boldsymbol{\nu})}}]$, and, moreover, for each fixed j and for $n \to \infty$ we have*

$$\lambda_{N(\boldsymbol{np}-\mathbf{1})-j}(n^{d-2}A_{\boldsymbol{n},D}^{(\boldsymbol{p})}) \to 0, \qquad \lambda_j(n^{d-2}A_{\boldsymbol{n},D}^{(\boldsymbol{p})}) \to M_{\mathbf{f}_{\boldsymbol{p}}^{(\boldsymbol{\nu})}}. \tag{3.68}$$

Proof. Since $A_{\boldsymbol{n},D}^{(\boldsymbol{p})}$ is SPD, $\lambda_{\min}(n^{d-2}A_{\boldsymbol{n},D}^{(\boldsymbol{p})}) > 0$. To prove the inclusion $\Lambda(n^{d-2}A_{\boldsymbol{n},D}^{(\boldsymbol{p})}) \subset (0, M_{\mathbf{f}_{\boldsymbol{p}}^{(\boldsymbol{\nu})}}]$, recall that in the proof of Theorem 3.8 we have defined the Hermitian matrix $T_{\boldsymbol{n}}^{(\boldsymbol{p})}$, see Eq. (3.60), and we have noticed that $T_{\boldsymbol{n}}^{(\boldsymbol{p})}$ is similar to $T_{\boldsymbol{n}}(\mathbf{f}_{\boldsymbol{p}}^{(\boldsymbol{\nu})})$. We show that for every $\mathbf{x} \in \mathbb{C}^{N(\boldsymbol{np}-\mathbf{1})}$ there exists $\mathbf{y} \in \mathbb{C}^{N(\boldsymbol{np})}$ with $\|\mathbf{y}\| = \|\mathbf{x}\|$ such that

$$\mathbf{x}^*(n^{d-2}A_{\boldsymbol{n},D}^{(\boldsymbol{p})})\mathbf{x} = \mathbf{y}^* T_{\boldsymbol{n}}^{(\boldsymbol{p})} \mathbf{y}, \tag{3.69}$$

which implies, by the minimax principle,

$$\lambda_{\max}(n^{d-2}A_{\boldsymbol{n},D}^{(\boldsymbol{p})}) = \max_{\|\mathbf{x}\|=1}(\mathbf{x}^*(n^{d-2}A_{\boldsymbol{n},D}^{(\boldsymbol{p})})\mathbf{x}) \le \max_{\|\mathbf{y}\|=1}(\mathbf{y}^* T_{\boldsymbol{n}}^{(\boldsymbol{p})} \mathbf{y}) = \lambda_{\max}(T_{\boldsymbol{n}}^{(\boldsymbol{p})}) = \lambda_{\max}(T_{\boldsymbol{n}}(\mathbf{f}_{\boldsymbol{p}}^{(\boldsymbol{\nu})})) \le M_{\mathbf{f}_{\boldsymbol{p}}^{(\boldsymbol{\nu})}},$$

the last inequality being justified by Theorem 1.8.

In order to prove (3.69), it is convenient to index vectors and matrices using multi-indices in $\mathbb{N}^d$, with the standard lexicographic ordering on them; see Subsection 1.1.1. For every $\mathbf{x} \in \mathbb{C}^{N(\boldsymbol{np}-\mathbf{1})}$ we have

$$\mathbf{x}^*(n^{d-2}A_{\boldsymbol{n},D}^{(\boldsymbol{p})})\mathbf{x} = \sum_{\boldsymbol{i},\boldsymbol{j}=\mathbf{1}}^{\boldsymbol{np}-\mathbf{1}} \overline{x}_{\boldsymbol{i}}(n^{d-2}A_{\boldsymbol{n},D}^{(\boldsymbol{p})})_{\boldsymbol{ij}} x_{\boldsymbol{j}} = \sum_{\boldsymbol{i},\boldsymbol{j} \in \{\mathbf{1},\ldots,\boldsymbol{np}-\mathbf{1}\}} \overline{x}_{\boldsymbol{i}}(n^{d-2}A_{\boldsymbol{n},D}^{(\boldsymbol{p})})_{\boldsymbol{ij}} x_{\boldsymbol{j}}.$$

Define $\mathbf{y} \in \mathbb{C}^{N(\boldsymbol{np})}$ in the following way:

$$y_{\boldsymbol{i}} = x_{\boldsymbol{i}} \quad \text{if } \boldsymbol{i} \in \{\mathbf{1}, \ldots, \boldsymbol{np}-\mathbf{1}\}, \qquad y_{\boldsymbol{i}} = 0 \quad \text{if } \boldsymbol{i} \in \{\mathbf{1}, \ldots, \boldsymbol{np}\} \backslash \{\mathbf{1}, \ldots, \boldsymbol{np}-\mathbf{1}\}.$$

Then $\|\mathbf{y}\| = \|\mathbf{x}\|$ and, moreover,

$$\mathbf{x}^*(n^{d-2}A_{\boldsymbol{n},D}^{(\boldsymbol{p})})\mathbf{x} = \sum_{\boldsymbol{i},\boldsymbol{j} \in \{\mathbf{1},\ldots,\boldsymbol{np}-\mathbf{1}\}} \overline{x}_{\boldsymbol{i}}(n^{d-2}A_{\boldsymbol{n},D}^{(\boldsymbol{p})})_{\boldsymbol{ij}} x_{\boldsymbol{j}} = \sum_{\boldsymbol{i},\boldsymbol{j} \in \{\mathbf{1},\ldots,\boldsymbol{np}\}} \overline{y}_{\boldsymbol{i}}(T_{\boldsymbol{n}}^{(\boldsymbol{p})})_{\boldsymbol{ij}} y_{\boldsymbol{j}} = \mathbf{y}^* T_{\boldsymbol{n}}^{(\boldsymbol{p})} \mathbf{y}. \tag{3.70}$$

This concludes the proof of the inclusion $\Lambda(n^{d-2}A_{\boldsymbol{n},D}^{(\boldsymbol{p})}) \subset (0, M_{\mathbf{f}_{\boldsymbol{p}}^{(\boldsymbol{\nu})}}]$, but we wish to prove in some more detail the central equality in (3.70).

- If $\boldsymbol{i} \in \{\mathbf{1}, \dots, \boldsymbol{np}\} \backslash \{\mathbf{1}, \dots, \boldsymbol{np}-\mathbf{1}\}$ or $\boldsymbol{j} \in \{\mathbf{1}, \dots, \boldsymbol{np}\} \backslash \{\mathbf{1}, \dots, \boldsymbol{np}-\mathbf{1}\}$, the $(\boldsymbol{i}, \boldsymbol{j})$ term in the right-hand side of the central equality is 0, due to the definition of $\boldsymbol{y}$.

- If $\boldsymbol{i} \in \{\mathbf{1}, \dots, \boldsymbol{np}-\mathbf{1}\}$ and $\boldsymbol{j} \in \{\mathbf{1}, \dots, \boldsymbol{np}-\mathbf{1}\}$, the $(\boldsymbol{i}, \boldsymbol{j})$ term in the right-hand side of the central equality is $\overline{y}_{\boldsymbol{i}}(T_{\boldsymbol{n}}^{(\boldsymbol{p})})_{\boldsymbol{ij}} y_{\boldsymbol{j}} = \overline{x}_{\boldsymbol{i}}(n^{d-2} A_{\boldsymbol{n},D}^{(\boldsymbol{p})})_{\boldsymbol{ij}} x_{\boldsymbol{j}}$, because $y_{\boldsymbol{i}} = x_{\boldsymbol{i}}$, $y_{\boldsymbol{j}} = x_{\boldsymbol{j}}$, and, recalling the fundamental equality (1.12) and the fact that $K_n^{(p)}$ and $M_n^{(p)}$ are the leading principal submatrices of order $np-1$ of $T_n(\mathbf{f}_p)$ and $T_n(\mathbf{h}_p)$, respectively, we have

$$
\begin{aligned}
(T_{\boldsymbol{n}}^{(\boldsymbol{p})})_{\boldsymbol{ij}} &= \sum_{k=1}^{d} c_k(\boldsymbol{\nu}) \, [T_{n_1}(\mathbf{h}_{p_1}) \otimes \cdots \otimes T_{n_{k-1}}(\mathbf{h}_{p_{k-1}}) \otimes T_{n_k}(\mathbf{f}_{p_k}) \otimes T_{n_{k+1}}(\mathbf{h}_{p_{k+1}}) \otimes \cdots \otimes T_{n_d}(\mathbf{h}_{p_d})]_{\boldsymbol{ij}} \\
&= \sum_{k=1}^{d} c_k(\boldsymbol{\nu}) \, [T_{n_1}(\mathbf{h}_{p_1})]_{i_1 j_1} \cdots [T_{n_{k-1}}(\mathbf{h}_{p_{k-1}})]_{i_{k-1} j_{k-1}} [T_{n_k}(\mathbf{f}_{p_k})]_{i_k j_k} [T_{n_{k+1}}(\mathbf{h}_{p_{k+1}})]_{i_{k+1} j_{k+1}} \cdots [T_{n_d}(\mathbf{h}_{p_d})]_{i_d j_d} \\
&= \sum_{k=1}^{d} c_k(\boldsymbol{\nu}) \, [M_{n_1}^{(p_1)}]_{i_1 j_1} \cdots [M_{n_{k-1}}^{(p_{k-1})}]_{i_{k-1} j_{k-1}} [K_{n_k}^{(p_k)}]_{i_k j_k} [M_{n_{k+1}}^{(p_{k+1})}]_{i_{k+1} j_{k+1}} \cdots [M_{n_d}^{(p_d)}]_{i_d j_d} \\
&= \sum_{k=1}^{d} c_k(\boldsymbol{\nu}) \, [M_{n_1}^{(p_1)} \otimes \cdots \otimes M_{n_{k-1}}^{(p_{k-1})} \otimes K_{n_k}^{(p_k)} \otimes M_{n_{k+1}}^{(p_{k+1})} \otimes \cdots \otimes M_{n_d}^{(p_d)}]_{\boldsymbol{ij}} = (n^{d-2} A_{\boldsymbol{n},D}^{(\boldsymbol{p})})_{\boldsymbol{ij}}.
\end{aligned}
$$

This concludes the proof of the central equality in (3.70) and the proof of the inclusion $\Lambda(n^{d-2} A_{\boldsymbol{n},D}^{(\boldsymbol{p})}) \subset (0, M_{\mathbf{f}_{\boldsymbol{p}}^{(\boldsymbol{\nu})}}]$. Relation (3.68) follows from this inclusion and from the fact that $\{n^{d-2} A_{\boldsymbol{n},D}^{(\boldsymbol{p})}\}_n \sim_\lambda \mathbf{f}_{\boldsymbol{p}}^{(\boldsymbol{\nu})}$ (by Theorem 3.8 applied with $\boldsymbol{\beta} = \mathbf{0}$ and $\gamma = 0$). We omit the formal proof of (3.68), because it is based on the same argument used for proving that items 1 and 3 in Theorem 1.8 imply item 4. □

Theorem 3.10. *Assume that $\boldsymbol{\beta}$ is constant. Then*

$$
\Lambda(n^{d-2} A_{\boldsymbol{n}}^{(\boldsymbol{p})}) \subset \left[\frac{\sum_{k=1}^{d} \zeta_{n_k, p_k} + \gamma_*}{n^2} G_{\boldsymbol{p}}, M_{\mathbf{f}_{\boldsymbol{p}}^{(\boldsymbol{\nu})}} + \frac{\|\gamma\|_{L^\infty(\Omega)}}{\nu_1 \cdots \nu_d n^2} S_{\boldsymbol{p}} \right] \times \left[-\frac{B_{\boldsymbol{p}} \|\boldsymbol{\beta}\|_\infty \sum_{k=1}^{d} \nu_k}{\nu_1 \cdots \nu_d n}, \frac{B_{\boldsymbol{p}} \|\boldsymbol{\beta}\|_\infty \sum_{k=1}^{d} \nu_k}{\nu_1 \cdots \nu_d n} \right],
$$

with $\zeta_{n,p}$, $G_{\boldsymbol{p}}$, $S_{\boldsymbol{p}}$, $B_{\boldsymbol{p}}$ as in Theorem 3.6. In particular, $\{n^{d-2} A_{\boldsymbol{n}}^{(\boldsymbol{p})}\}_n$ is strongly clustered at $[0, M_{\mathbf{f}_{\boldsymbol{p}}^{(\boldsymbol{\nu})}}]$.

Proof. The real and imaginary parts of $n^{d-2} A_{\boldsymbol{n}}^{(\boldsymbol{p})}$ are

$$
\Re(n^{d-2} A_{\boldsymbol{n}}^{(\boldsymbol{p})}) = n^{d-2} A_{\boldsymbol{n},D}^{(\boldsymbol{p})} + n^{d-2} A_{\boldsymbol{n},R}^{(\boldsymbol{p})}, \qquad \Im(n^{d-2} A_{\boldsymbol{n}}^{(\boldsymbol{p})}) = -\mathrm{i}\, n^{d-2} A_{\boldsymbol{n},A}^{(\boldsymbol{p})};
$$

cf. (3.51)–(3.52). By Theorem 3.6, Theorem 3.9, (1.9) and (3.64) we have

$$
\frac{\sum_{k=1}^{d} \zeta_{n_k, p_k} + \gamma_*}{\nu_1 \cdots \nu_d n^2} G_{\boldsymbol{p}} \le \lambda_{\min}(\Re(n^{d-2} A_{\boldsymbol{n}}^{(\boldsymbol{p})})) \le \lambda_{\max}(\Re(n^{d-2} A_{\boldsymbol{n}}^{(\boldsymbol{p})})) \le \lambda_{\max}(n^{d-2} A_{\boldsymbol{n},D}^{(\boldsymbol{p})}) + \lambda_{\max}(n^{d-2} A_{\boldsymbol{n},R}^{(\boldsymbol{p})}) \le M_{\mathbf{f}_{\boldsymbol{p}}^{(\boldsymbol{\nu})}} + \frac{\|\gamma\|_{L^\infty(\Omega)} S_{\boldsymbol{p}}}{\nu_1 \cdots \nu_d n^2}.
$$

By (3.63) we have

$$
-\frac{B_{\boldsymbol{p}} \|\boldsymbol{\beta}\|_\infty \sum_{k=1}^{d} \nu_k}{\nu_1 \cdots \nu_d n} \le \lambda_{\min}(\Im(n^{d-2} A_{\boldsymbol{n}}^{(\boldsymbol{p})})) \le \lambda_{\max}(\Im(n^{d-2} A_{\boldsymbol{n}}^{(\boldsymbol{p})})) \le \frac{B_{\boldsymbol{p}} \|\boldsymbol{\beta}\|_\infty \sum_{k=1}^{d} \nu_k}{\nu_1 \cdots \nu_d n}.
$$

The thesis follows from (1.7). □

Chapter 4

Spectral analysis and spectral symbol of Galerkin B-spline IgA stiffness matrices

In this chapter, we perform a spectral analysis completely analogous to the one carried out in Chapter 3: we choose again a model problem like (3.1), we introduce a numerical method for approximating its solution, and we study the spectral properties of the discretization matrices associated with this numerical method, with particular attention to the conditioning, the behavior of the extremal eigenvalues, the asymptotic spectral distribution when the matrix size goes to infinity, and the properties of the spectral symbol. The only significant difference with respect to Chapter 3 is that the approximation technique investigated in this chapter is the so-called Galerkin B-spline Isogeometric Analysis (IgA). We refer the reader to [33, Section 1.2] for a quick overview of the IgA paradigm and to [19, 41] for a more detailed introduction to this fascinating subject. Here, we limit to say that the goal of IgA is to improve the connection between numerical simulation of PDE and Computer Aided Design (CAD) systems, the latter being widely employed in Engineering.

As already pointed out in the Introduction of this thesis, we emphasize once again that the (asymptotic) spectral analysis in this chapter is a preliminary step for designing efficient preconditioners and iterative solvers for the Galerkin B-spline IgA stiffness matrices. In particular, the knowledge of the symbol and of its properties is fundamental to this purpose. The design of fast iterative solvers for the Galerkin B-spline IgA stiffness matrices will be the subject of Chapter 6, where we will use the specific features of the symbol studied in this chapter to obtain a robust and optimal multi-iterative multigrid method, whose convergence rate will be substantially independent not only of the matrix size and the fineness parameters, but also of the spline approximation degrees and the dimensionality d of the considered model problem.

4.1 Problem setting and Galerkin B-spline IgA

Let us consider as our model problem the following second-order elliptic differential equation with homogeneous Dirichlet boundary conditions:

$$\begin{cases} -\Delta u + \boldsymbol{\beta} \cdot \nabla u + \gamma u = f & \text{in } \Omega := (0,1)^d, \\ u = 0 & \text{on } \partial\Omega, \end{cases} \tag{4.1}$$

where $f \in L^2(\Omega)$, $\boldsymbol{\beta} = (\beta_1, \dots, \beta_d) \in \mathbb{R}^d$ and $\gamma \geq 0$. The only difference with respect to the model problem (3.1) considered in Chapter 3 is that now we assume $\boldsymbol{\beta}$ and γ constant. This assumption is made only to simplify the presentation because, in fact, nothing significant would change if $\beta_1, \dots, \beta_d, \gamma$ were only assumed to be in $L^\infty(\Omega)$, as in Chapter 3. The weak form of (4.1) and the Galerkin method for approximating its solution u have already been described in Section 3.1, see (3.2)–(3.4), and so we do not repeat them here. We just point out that, since we have assumed $\boldsymbol{\beta}$ constant, the bilinear form $a(\cdot,\cdot)$ in (3.2)–(3.3) is coercive; see the footnote in correspondence of Eq. (3.3). Therefore, the matrix A in (3.4) is positive definite, in the sense that

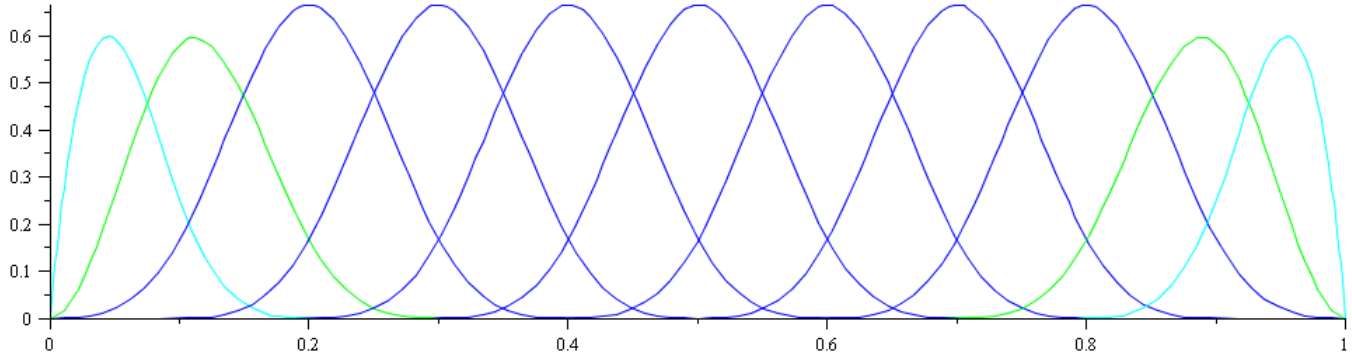


Figure 4.1: graph of the basis functions $N_{i,[p]}(x)$, $i = 2, \ldots, n+p-1$, for $p = 3$ and $n = 10$. The blue functions $N_{i,[p]}$, $i = p+1, \ldots, n$, are the so-called 'central basis functions'; see Subsection 4.3.1 and, especially, Eq. (4.42).

$\mathbf{v}^T A\mathbf{v} > 0$ for all $\mathbf{v} \in \mathbb{R}^N \backslash \{\mathbf{0}\}$; see [48, Theorem 4.1]. If moreover $\boldsymbol{\beta} = \mathbf{0}$, then the bilinear form $a(\cdot,\cdot)$ is also symmetric and A is SPD.

In the context of IgA based on B-splines of degree p, the approximation space W in the Galerkin method is chosen as a space of C^{p-1}-continuous piecewise polynomial functions vanishing on the boundary of Ω. More precisely, define for $p, n \geq 1$ the spline spaces

$$V_n^{[p]} := \left\{ s \in C^{p-1}([0,1]) : \; s_{|[\frac{i}{n}, \frac{i+1}{n})} \in \mathbb{P}_p \quad \forall i = 0, \ldots, n-1 \right\},$$
$$W_n^{[p]} := \left\{ s \in V_n^{[p]} : \; s(0) = s(1) = 0 \right\} \subset H_0^1(0,1).$$

It is known that $\dim V_n^{[p]} = n + p$ and $\dim W_n^{[p]} = n + p - 2$. We consider for $V_n^{[p]}$ the B-spline basis $\{N_{1,[p]}, \ldots, N_{n+p,[p]}\}$, which is defined recursively as follows; see also [21].

Definition 4.1 (**B-spline basis**). Consider the knot sequence

$$t_1 = \ldots = t_{p+1} = 0 < t_{p+2} < \ldots < t_{p+n} < 1 = t_{p+n+1} = \ldots = t_{2p+n+1}, \tag{4.2}$$

where

$$t_{p+i+1} := \frac{i}{n}, \qquad i = 0, \ldots, n. \tag{4.3}$$

Using the convention that a fraction with zero denominator is zero, for every (k, i) with $0 \leq k \leq p$ and $1 \leq i \leq (n+p) + p - k$, define the function $N_{i,[k]} : [0,1] \to \mathbb{R}$ as follows:

$$N_{i,[0]}(x) := \begin{cases} 1 & \text{if } x \in [t_i, t_{i+1}), \\ 0 & \text{elsewhere,} \end{cases}$$

and, if $k > 0$,

$$N_{i,[k]}(x) := \frac{x - t_i}{t_{i+k} - t_i} N_{i,[k-1]}(x) + \frac{t_{i+k+1} - x}{t_{i+k+1} - t_{i+1}} N_{i+1,[k-1]}(x). \tag{4.4}$$

The functions $N_{1,[p]}, \ldots, N_{n+p,[p]}$ constructed in this way form a basis for $V_n^{[p]}$ (the B-spline basis of $V_n^{[p]}$); see [21]. Moreover, since we have [21]

$$N_{i,[p]}(0) = N_{i,[p]}(1) = 0, \qquad \forall i = 2, \ldots, n + p - 1,$$

$\{N_{2,[p]}, \ldots, N_{n+p-1,[p]}\}$ is a basis for $W_n^{[p]}$ (the B-spline basis of $W_n^{[p]}$). Figure 4.1 shows the graph of the basis functions $N_{2,[p]}, \ldots, N_{n+p-1,[p]}$ in the case $p = 3$ and $n = 10$.

Now, for any pair of multi-indices $\boldsymbol{p}, \boldsymbol{n} \in \mathbb{N}^d$, we define

$$W_{\boldsymbol{n}}^{[\boldsymbol{p}]} := W_{n_1}^{[p_1]} \otimes \cdots \otimes W_{n_d}^{[p_d]} := \operatorname{span}(N_{\boldsymbol{i},[\boldsymbol{p}]} : \ \boldsymbol{i} = \boldsymbol{2}, \ldots, \boldsymbol{n} + \boldsymbol{p} - \boldsymbol{1}) \subset H_0^1(\Omega), \tag{4.5}$$

where $N_{\boldsymbol{i},[\boldsymbol{p}]} := N_{i_1,[p_1]} \otimes \cdots \otimes N_{i_d,[p_d]}$.

In the framework of Galerkin B-spline IgA, the model problem (4.1) is approximated by the standard Galerkin method, the approximation space W in the Galerkin problem (3.3) is chosen as $W_{\boldsymbol{n}}^{[\boldsymbol{p}]}$ for some $\boldsymbol{n}, \boldsymbol{p} \in \mathbb{N}^d$ (usually $\boldsymbol{p} = (p, \ldots, p)$ for some $p \geq 1$), and the basis for $W_{\boldsymbol{n}}^{[\boldsymbol{p}]}$ is chosen as the tensor-product B-spline basis in (4.5), ordered according to the standard lexicographic ordering (1.1) for the multi-index range $\boldsymbol{2}, \ldots, \boldsymbol{n} + \boldsymbol{p} - \boldsymbol{1}$. With these choices, we obtain in (3.4) a stiffness matrix A, which henceforth will be denoted by $A_{\boldsymbol{n}}^{[\boldsymbol{p}]}$ in order to emphasize its dependence on $\boldsymbol{p}$ and $\boldsymbol{n}$:

$$A_{\boldsymbol{n}}^{[\boldsymbol{p}]} := \left[a(N_{\boldsymbol{j},[\boldsymbol{p}]}, N_{\boldsymbol{i},[\boldsymbol{p}]})\right]_{\boldsymbol{i},\boldsymbol{j}=\boldsymbol{2}}^{\boldsymbol{n}+\boldsymbol{p}-\boldsymbol{1}} = \left[a(N_{\boldsymbol{j}+\boldsymbol{1},[\boldsymbol{p}]}, N_{\boldsymbol{i}+\boldsymbol{1},[\boldsymbol{p}]})\right]_{\boldsymbol{i},\boldsymbol{j}=\boldsymbol{1}}^{\boldsymbol{n}+\boldsymbol{p}-\boldsymbol{2}}. \tag{4.6}$$

Let us consider the following split of the matrix $A_{\boldsymbol{n}}^{[\boldsymbol{p}]}$, according to the diffusion, advection and reaction terms, respectively:

$$A_{\boldsymbol{n}}^{[\boldsymbol{p}]} = \left[\int_\Omega \nabla N_{\boldsymbol{j}+\boldsymbol{1},[\boldsymbol{p}]} \cdot \nabla N_{\boldsymbol{i}+\boldsymbol{1},[\boldsymbol{p}]}\right]_{\boldsymbol{i},\boldsymbol{j}=\boldsymbol{1}}^{\boldsymbol{n}+\boldsymbol{p}-\boldsymbol{2}} + \left[\int_\Omega \boldsymbol{\beta} \cdot \nabla N_{\boldsymbol{j}+\boldsymbol{1},[\boldsymbol{p}]}\, N_{\boldsymbol{i}+\boldsymbol{1},[\boldsymbol{p}]}\right]_{\boldsymbol{i},\boldsymbol{j}=\boldsymbol{1}}^{\boldsymbol{n}+\boldsymbol{p}-\boldsymbol{2}} + \left[\int_\Omega \gamma\, N_{\boldsymbol{j}+\boldsymbol{1},[\boldsymbol{p}]} N_{\boldsymbol{i}+\boldsymbol{1},[\boldsymbol{p}]}\right]_{\boldsymbol{i},\boldsymbol{j}=\boldsymbol{1}}^{\boldsymbol{n}+\boldsymbol{p}-\boldsymbol{2}}. \tag{4.7}$$

For obvious reasons, the first matrix in the right-hand side of (4.7) is called diffusion matrix, the second advection matrix, and the third reaction matrix. With expressive notation, we denote these three matrices by $A_{\boldsymbol{n},D}^{[\boldsymbol{p}]}, A_{\boldsymbol{n},A}^{[\boldsymbol{p}]}, A_{\boldsymbol{n},R}^{[\boldsymbol{p}]}$, respectively:

$$A_{\boldsymbol{n},D}^{[\boldsymbol{p}]} := \left[\int_\Omega \nabla N_{\boldsymbol{j}+\boldsymbol{1},[\boldsymbol{p}]} \cdot \nabla N_{\boldsymbol{i}+\boldsymbol{1},[\boldsymbol{p}]}\right]_{\boldsymbol{i},\boldsymbol{j}=\boldsymbol{1}}^{\boldsymbol{n}+\boldsymbol{p}-\boldsymbol{2}}, \tag{4.8}$$

$$A_{\boldsymbol{n},A}^{[\boldsymbol{p}]} := \left[\int_\Omega \boldsymbol{\beta} \cdot \nabla N_{\boldsymbol{j}+\boldsymbol{1},[\boldsymbol{p}]}\, N_{\boldsymbol{i}+\boldsymbol{1},[\boldsymbol{p}]}\right]_{\boldsymbol{i},\boldsymbol{j}=\boldsymbol{1}}^{\boldsymbol{n}+\boldsymbol{p}-\boldsymbol{2}}, \tag{4.9}$$

$$A_{\boldsymbol{n},R}^{[\boldsymbol{p}]} := \left[\int_\Omega \gamma\, N_{\boldsymbol{j}+\boldsymbol{1},[\boldsymbol{p}]} N_{\boldsymbol{i}+\boldsymbol{1},[\boldsymbol{p}]}\right]_{\boldsymbol{i},\boldsymbol{j}=\boldsymbol{1}}^{\boldsymbol{n}+\boldsymbol{p}-\boldsymbol{2}}. \tag{4.10}$$

The diffusion matrix is SPD, the reaction matrix is SPSD (SPD if $\gamma \neq 0$), while the advection matrix is skew-symmetric and is responsible for the non-symmetry of $A_{\boldsymbol{n}}^{[\boldsymbol{p}]}$. The real and imaginary parts of $A_{\boldsymbol{n}}^{[\boldsymbol{p}]}$ are

$$\Re(A_{\boldsymbol{n}}^{[\boldsymbol{p}]}) = A_{\boldsymbol{n},D}^{[\boldsymbol{p}]} + A_{\boldsymbol{n},R}^{[\boldsymbol{p}]}, \tag{4.11}$$

$$\Im(A_{\boldsymbol{n}}^{[\boldsymbol{p}]}) = -\mathrm{i} A_{\boldsymbol{n},A}^{[\boldsymbol{p}]}. \tag{4.12}$$

Before providing a construction of the Galerkin B-spline IgA stiffness matrix $A_{\boldsymbol{n}}^{[\boldsymbol{p}]}$, we introduce in the next section the cardinal B-splines. We also study some of their properties which are relevant for our purposes and, in particular, for obtaining a simplified expression of $A_{\boldsymbol{n}}^{[\boldsymbol{p}]}$.

4.2 Cardinal B-splines

The cardinal B-spline of degree p over the uniform knot sequence $\{0, 1, \ldots, p+1\}$ is denoted by $\phi_{[p]}$ and is defined recursively as follows [21]:

$$\phi_{[0]}(t) := \begin{cases} 1 & \text{if } t \in [0,1), \\ 0 & \text{elsewhere,} \end{cases} \tag{4.13}$$

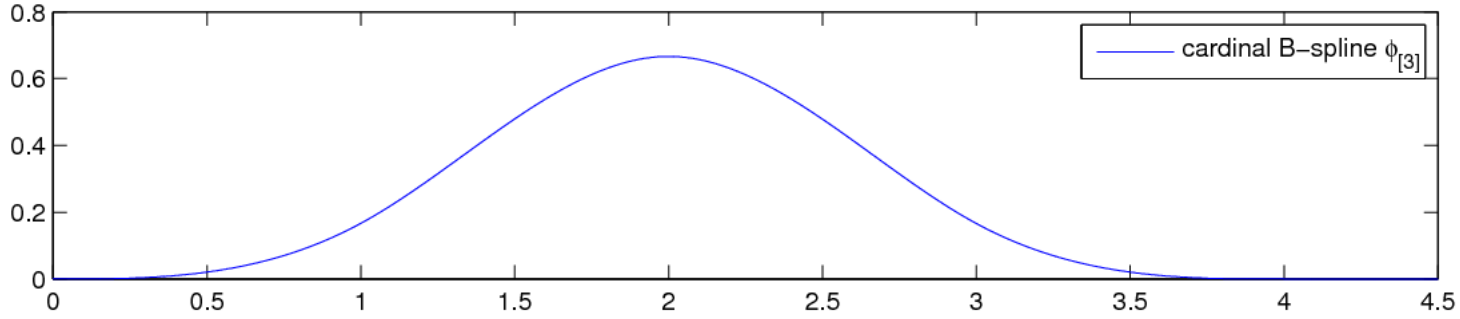


Figure 4.2: graph of the cubic cardinal B-spline $\phi_{[3]}$.

and

$$\phi_{[p]}(t) := \frac{t}{p}\phi_{[p-1]}(t) + \frac{p+1-t}{p}\phi_{[p-1]}(t-1), \quad p \geq 1. \tag{4.14}$$

The cardinal B-spline can also be expressed in terms of truncated powers [21]:

$$\phi_{[p]}(t) = \frac{1}{p!}\sum_{i=0}^{p+1}(-1)^i \binom{p+1}{i}(t-i)_+^p, \tag{4.15}$$

where $(t)_+^r := (\max(t,0))^r$. Figure 4.2 shows the graph of the cubic cardinal B-spline $\phi_{[3]}$. As usual in the literature, we will refer to cardinal B-splines of degree p as the set of integer translates of $\phi_{[p]}$, that is $\{\phi_{[p]}(\cdot - k),\ k \in \mathbb{Z}\}$. In the next subsections we collect some properties of cardinal B-splines and their Fourier transform that will be useful later on.

4.2.1 Properties of cardinal B-splines

It is known that $\phi_{[p]} \in C^{p-1}(\mathbb{R})$ and $\phi_{[p]}$ coincides with a polynomial in $\mathbb{P}_p$ over the intervals $[i, i+1]$, $i = 0, \ldots, p$. Moreover, $\phi_{[p]}$ possesses certain fundamental properties, some of which are briefly summarized below; see [21, 17].

- *Positivity:*
$$\phi_{[p]}(t) \geq 0, \quad t \in \mathbb{R}. \tag{4.16}$$
- *Minimal support:*
$$\operatorname{supp}(\phi_{[p]}) = [0, p+1] \quad \Rightarrow \quad \phi_{[p]}(t) = 0, \quad t \notin [0, p+1]. \tag{4.17}$$
- *Symmetry:*
$$\phi_{[p]}\left(\frac{p+1}{2} + t\right) = \phi_{[p]}\left(\frac{p+1}{2} - t\right). \tag{4.18}$$
- *Partition of unity:*
$$\sum_{k\in\mathbb{Z}} \phi_{[p]}(t-k) = 1, \tag{4.19}$$
which gives, in combination with the local support and continuity,
$$\sum_{k=1}^{p} \phi_{[p]}(k) = 1, \quad p \geq 1. \tag{4.20}$$

- *Recurrence relation for derivatives:*

$$\phi_{[p]}^{(r)}(t) = \phi_{[p-1]}^{(r-1)}(t) - \phi_{[p-1]}^{(r-1)}(t-1). \tag{4.21}$$

- *Convolution relation:*

$$\phi_{[p]}(t) = (\phi_{[p-1]} * \phi_{[0]})(t) := \int_{\mathbb{R}} \phi_{[p-1]}(t-s)\phi_{[0]}(s)\,ds = \int_0^1 \phi_{[p-1]}(t-s)\,ds. \tag{4.22}$$

In the remainder of this subsection we derive from the previous properties some results that are needed in later sections. The next lemma generalizes the symmetry property to derivatives of any order of the cardinal B-splines.

Lemma 4.1. *We have*

$$\phi_{[p]}^{(r)}\left(\frac{p+1}{2}+t\right) = (-1)^r\,\phi_{[p]}^{(r)}\left(\frac{p+1}{2}-t\right).$$

Proof. The result follows from repeated differentiations of the symmetry property (4.18). We can also prove it by induction on the order of derivatives using the recurrence relation (4.21), as outlined below. The base case ($r = 0$) is just the symmetry property (4.18). As inductive step we increase the order of derivative by one, i.e., $r \to r+1$. Using the recurrence relation for derivatives (4.21) and the induction hypothesis, we have

$$\begin{aligned}\phi_{[p]}^{(r+1)}\left(\frac{p+1}{2}+t\right) &= \phi_{[p-1]}^{(r)}\left(\frac{p+1}{2}+t\right) - \phi_{[p-1]}^{(r)}\left(\frac{p+1}{2}+t-1\right) = (-1)^r\left(\phi_{[p-1]}^{(r)}\left(\frac{p+1}{2}-t-1\right) - \phi_{[p-1]}^{(r)}\left(\frac{p+1}{2}-t\right)\right)\\ &= (-1)^{r+1}\,\phi_{[p]}^{(r+1)}\left(\frac{p+1}{2}-t\right).\end{aligned}$$

□

The following lemma provides an expression for inner products of derivatives of the cardinal B-spline and its integer translates. It generalizes the result given in [17, p. 89].

Lemma 4.2. *We have*

$$\int_{\mathbb{R}} \phi_{[p_1]}^{(r)}(t)\,\phi_{[p_2]}^{(s)}(t+k)\,dt = (-1)^r\,\phi_{[p_1+p_2+1]}^{(r+s)}(p_1+1+k) = (-1)^s\,\phi_{[p_1+p_2+1]}^{(r+s)}(p_2+1-k). \tag{4.23}$$

Proof. Because of the (anti-)symmetry of the higher order derivatives of the B-splines given by Lemma 4.1, we have

$$\begin{aligned}(-1)^r\,\phi_{[p_1+p_2+1]}^{(r+s)}(p_1+1+k) &= (-1)^r\,\phi_{[p_1+p_2+1]}^{(r+s)}\left(\frac{p_1+p_2+2}{2}+\frac{p_1-p_2}{2}+k\right)\\ &= (-1)^r\,(-1)^{r+s}\,\phi_{[p_1+p_2+1]}^{(r+s)}\left(\frac{p_1+p_2+2}{2}-\frac{p_1-p_2}{2}-k\right)\\ &= (-1)^s\,\phi_{[p_1+p_2+1]}^{(r+s)}(p_2+1-k).\end{aligned}$$

So, we only have to show one of the two equalities in (4.23).

We first address the case $r = s = 0$, namely

$$\int_{\mathbb{R}} \phi_{[p_1]}(t)\phi_{[p_2]}(t+k)\,dt = \phi_{[p_1+p_2+1]}(p_2+1-k). \tag{4.24}$$

Using the convolution relation of cardinal B-splines (4.22), we obtain

$$\phi_{[p_1+p_2+1]}(p_2+1-k) = \int_0^1 \phi_{[p_1+p_2]}(p_2+1-k-t_1)\,dt_1 = \int_0^1 \dots \int_0^1 \phi_{[p_2]}(p_2+1-k-(t_1+t_2+\dots+t_{p_1+1}))\,dt_1 \dots dt_{p_1+1}.$$

From [17, p. 85] we also know that for every continuous function f it holds

$$\int_{\mathbb{R}} f(t)\phi_{[p]}(t)\,dt = \int_0^1 \dots \int_0^1 f(t_1+t_2+\dots+t_{p+1})\,dt_1 \dots dt_{p+1},$$

and hence

$$\phi_{[p_1+p_2+1]}(p_2+1-k) = \int_{\mathbb{R}} \phi_{[p_2]}(p_2+1-k-t)\phi_{[p_1]}(t)\,dt. \tag{4.25}$$

Moreover, by symmetry of the cardinal B-splines, see (4.18), we have

$$\phi_{[p_2]}(p_2+1-k-t) = \phi_{[p_2]}(k+t). \tag{4.26}$$

The combination of (4.25) and (4.26) results in (4.24).

We now prove the general case, i.e.,

$$\int_{\mathbb{R}} \phi^{(r)}_{[p_1]}(t)\,\phi^{(s)}_{[p_2]}(t+k)\,dt = (-1)^r\,\phi^{(r+s)}_{[p_1+p_2+1]}(p_1+1+k), \tag{4.27}$$

by induction on the order of derivatives. We consider two inductive steps: in the first inductive step we increase the order of derivative of $\phi_{[p_1]}$ by one, i.e., $r \to r+1$, and in the second inductive step we increase the order of derivative of $\phi_{[p_2]}$ by one, i.e., $s \to s+1$.

1. $(r \to r+1)$. Using (4.21) and the induction hypothesis, we have

$$\begin{aligned}
\int_{\mathbb{R}} \phi^{(r+1)}_{[p_1]}(t)\,\phi^{(s)}_{[p_2]}(t+k)\,dt &= \int_{\mathbb{R}} \left(\phi^{(r)}_{[p_1-1]}(t) - \phi^{(r)}_{[p_1-1]}(t-1)\right)\,\phi^{(s)}_{[p_2]}(t+k)\,dt \\
&= \int_{\mathbb{R}} \phi^{(r)}_{[p_1-1]}(t)\phi^{(s)}_{[p_2]}(t+k)\,dt - \int_{\mathbb{R}} \phi^{(r)}_{[p_1-1]}(t-1)\phi^{(s)}_{[p_2]}(t+k)\,dt \\
&= \int_{\mathbb{R}} \phi^{(r)}_{[p_1-1]}(t)\phi^{(s)}_{[p_2]}(t+k)\,dt - \int_{\mathbb{R}} \phi^{(r)}_{[p_1-1]}(t)\phi^{(s)}_{[p_2]}(t+k+1)\,dt \\
&= (-1)^r \left(\phi^{(r+s)}_{[p_1+p_2]}(p_1+k) - \phi^{(r+s)}_{[p_1+p_2]}(p_1+1+k)\right) \\
&= (-1)^{r+1}\,\phi^{(r+s+1)}_{[p_1+p_2+1]}(p_1+1+k).
\end{aligned}$$

2. $(s \to s+1)$. This inductive step can be proved in a completely analogous way as the first inductive step.

□

Finally, we provide some relations about second derivatives of cardinal B-splines. We will denote by $\dot{\phi}_{[p]}$ and $\ddot{\phi}_{[p]}$ the first and second derivative of $\phi_{[p]}$.

Lemma 4.3. *We have*

$$\sum_{k=1}^{p} \ddot{\phi}_{[2p+1]}(p+1-k) = \dot{\phi}_{[2p]}(p) = -\frac{1}{2}\ddot{\phi}_{[2p+1]}(p+1), \qquad \sum_{k=1}^{p} k^2\,\ddot{\phi}_{[2p+1]}(p+1-k) = 1.$$

Proof. We first note that by (4.21) and (4.18) we have

$$-\ddot{\phi}_{[2p+1]}(p+1) = -2\dot{\phi}_{[2p]}(p+1) = 2\dot{\phi}_{[2p]}(p) > 0. \tag{4.28}$$

Using (4.21) and $\phi_{[2p-1]}(-1) = \phi_{[2p-1]}(0) = 0$, we obtain

$$\begin{aligned}\sum_{k=1}^{p} \ddot{\phi}_{[2p+1]}(p+1-k) &= \sum_{k=1}^{p} (\phi_{[2p-1]}(p+1-k) - 2\phi_{[2p-1]}(p-k) + \phi_{[2p-1]}(p-1-k)) \\ &= \phi_{[2p-1]}(p) - \phi_{[2p-1]}(p-1) = \dot{\phi}_{[2p]}(p).\end{aligned}$$

In a similar way, taking into account that

$$k^2 - 2(k+1)^2 + (k+2)^2 = 2, \quad k \geq 0,$$

we find that

$$\begin{aligned}\sum_{k=1}^{p} k^2\, \ddot{\phi}_{[2p+1]}(p+1-k) &= \sum_{k=1}^{p} k^2\, (\phi_{[2p-1]}(p+1-k) - 2\phi_{[2p-1]}(p-k) + \phi_{[2p-1]}(p-1-k)) \\ &= \phi_{[2p-1]}(p) + 2\sum_{k=2}^{p} \phi_{[2p-1]}(p+1-k) = \sum_{k=-p+2}^{p} \phi_{[2p-1]}(p+1-k) = \sum_{k=1}^{2p-1} \phi_{[2p-1]}(k) = 1.\end{aligned}$$

The last equalities follow from the symmetry property (4.18) and the partition of unity property (4.20) of cardinal B-splines. □

4.2.2 Fourier transform of cardinal B-splines

In this subsection we will address some relations between inner products of cardinal B-splines, and the Fourier transform of the cardinal B-spline. We will need the following result; see [17, Theorem 2.28]. Recall that, given any two functions $\xi, \zeta : \mathbb{R} \to \mathbb{C}$, the notation '$\xi(t) = O(\zeta(t))$ as $|t| \to \infty$' means that '$|\xi(t)| \leq C|\zeta(t)|$ for $|t| \geq T$', where C, T are positive constant independent of t.

Theorem 4.1. *Let $\psi \in L^2(\mathbb{R})$ and its Fourier transform $\widehat{\psi}$ satisfy*

$$\psi(t) = O(|t|^{-a}), \quad a > 1, \quad \text{as } |t| \to \infty, \tag{4.29}$$

and

$$\widehat{\psi}(\theta) = O(|\theta|^{-b}), \quad b > \frac{1}{2}, \quad \text{as } |\theta| \to \infty. \tag{4.30}$$

Then,

$$\sum_{k\in\mathbb{Z}} \left(\int_{\mathbb{R}} \psi(t-k)\overline{\psi(t)}\, dt \right) \mathrm{e}^{\mathrm{i}k\theta} = \sum_{k\in\mathbb{Z}} \left| \widehat{\psi}(\theta + 2k\pi) \right|^2, \quad \forall \theta \in [-\pi, \pi]. \tag{4.31}$$

By using the convolution relation (4.22), one can easily obtain a simple expression for the Fourier transform of the cardinal B-spline $\phi_{[p]}$ (see [17, p. 56]):

$$\widehat{\phi_{[p]}}(\theta) = \left(\frac{1 - \mathrm{e}^{-\mathrm{i}\theta}}{\mathrm{i}\theta} \right) \widehat{\phi_{[p-1]}}(\theta) = \left(\widehat{\phi_{[0]}}(\theta) \right)^{p+1} = \left(\frac{1 - \mathrm{e}^{-\mathrm{i}\theta}}{\mathrm{i}\theta} \right)^{p+1}. \tag{4.32}$$

It follows that

$$\left| \widehat{\phi_{[p]}}(\theta) \right| = \left| \widehat{\phi_{[0]}}(\theta) \right|^{p+1} = \left(\frac{2 - 2\cos\theta}{\theta^2} \right)^{\frac{p+1}{2}} = \left| \frac{\sin(\theta/2)}{\theta/2} \right|^{p+1}. \tag{4.33}$$

We also note that for the symmetrized function $\phi^*_{[0]}(t) := \phi_{[0]}(t + 1/2)$, we have

$$\widehat{\phi^*_{[0]}}(\theta) = \frac{\sin(\theta/2)}{\theta/2}. \tag{4.34}$$

Concerning the Fourier transform of $\dot{\phi}_{[p]}$, using the recurrence relation for derivatives (4.21), we obtain

$$\widehat{\dot{\phi}_{[p]}}(\theta) = \mathrm{i}\theta\widehat{\phi_{[p]}}(\theta) = \left(1 - \mathrm{e}^{-\mathrm{i}\theta}\right)\widehat{\phi_{[p-1]}}(\theta). \tag{4.35}$$

From (4.17) and (4.33) it follows that the cardinal B-spline $\phi_{[p]}$ satisfies the conditions (4.29)–(4.30). When using $\phi_{[p]}$ as the function ψ in Theorem 4.1, we can express the right-hand side in (4.31) by means of (4.33). This implies

$$\sum_{k\in\mathbb{Z}} \left|\widehat{\phi_{[p]}}(\theta + 2k\pi)\right|^2 \geq \left|\widehat{\phi_{[p]}}(\theta)\right|^2 = \left(\frac{2 - 2\cos\theta}{\theta^2}\right)^{p+1} \geq \left(\frac{4}{\pi^2}\right)^{p+1}, \quad \forall\theta \in [-\pi, \pi]. \tag{4.36}$$

A sharper lower bound can be found in [17]. It is formulated in terms of the roots of the so-called Euler-Frobenius polynomials of degree $2p$, but these roots are not provided in a closed form expression. On the other hand, to obtain an upper bound for (4.31), we make use of relations (4.24), (4.31) and the partition of unity property (4.19). In this way, we obtain

$$\sum_{k\in\mathbb{Z}} \left|\widehat{\phi_{[p]}}(\theta + 2k\pi)\right|^2 = \sum_{k\in\mathbb{Z}} \phi_{[2p+1]}(p + 1 - k)\,\mathrm{e}^{\mathrm{i}k\theta} \leq \sum_{k\in\mathbb{Z}} \phi_{[2p+1]}(p + 1 - k)\,|\mathrm{e}^{\mathrm{i}k\theta}| = 1. \tag{4.37}$$

Note that for the cardinal B-spline of degree p the left-hand side in (4.31) is a finite sum consisting of $2p+1$ terms.

4.3 Construction of the Galerkin B-spline IgA stiffness matrices $A^{[\boldsymbol{p}]}_{\boldsymbol{n}}$

Using the tensor structure of the tensor-product B-spline basis $\{N_{\boldsymbol{j}+\boldsymbol{1},[\boldsymbol{p}]} : \boldsymbol{j} = \boldsymbol{1}, \ldots, \boldsymbol{n} + \boldsymbol{p} - \boldsymbol{2}\}$ and the rectangularity of the domain Ω, we now prove the following result, analogous to Theorem 3.1, which highlights the tensor structure of the Galerkin B-spline IgA matrices (4.8)–(4.10).

Theorem 4.2. *Let $\boldsymbol{p}, \boldsymbol{n} \in \mathbb{N}^d$, then*

$$A^{[\boldsymbol{p}]}_{\boldsymbol{n},D} = \sum_{k=1}^{d} \frac{1}{n_1} M^{[p_1]}_{n_1} \otimes \cdots \otimes \frac{1}{n_{k-1}} M^{[p_{k-1}]}_{n_{k-1}} \otimes n_k K^{[p_k]}_{n_k} \otimes \frac{1}{n_{k+1}} M^{[p_{k+1}]}_{n_{k+1}} \otimes \cdots \otimes \frac{1}{n_d} M^{[p_d]}_{n_d}, \tag{4.38}$$

$$A^{[\boldsymbol{p}]}_{\boldsymbol{n},A} = \sum_{k=1}^{d} \frac{1}{n_1} M^{[p_1]}_{n_1} \otimes \cdots \otimes \frac{1}{n_{k-1}} M^{[p_{k-1}]}_{n_{k-1}} \otimes \beta_k H^{[p_k]}_{n_k} \otimes \frac{1}{n_{k+1}} M^{[p_{k+1}]}_{n_{k+1}} \otimes \cdots \otimes \frac{1}{n_d} M^{[p_d]}_{n_d}, \tag{4.39}$$

$$A^{[\boldsymbol{p}]}_{\boldsymbol{n},R} = \gamma \frac{1}{n_1} M^{[p_1]}_{n_1} \otimes \cdots \otimes \frac{1}{n_d} M^{[p_d]}_{n_d}, \tag{4.40}$$

where, for $p, n \geq 1$, $K^{[p]}_n$, $H^{[p]}_n$, $M^{[p]}_n$ are given by

$$nK^{[p]}_n := \left[\int_{(0,1)} N'_{j+1,[p]} N'_{i+1,[p]}\right]^{n+p-2}_{i,j=1}, \quad H^{[p]}_n := \left[\int_{(0,1)} N'_{j+1,[p]} N_{i+1,[p]}\right]^{n+p-2}_{i,j=1}, \quad \frac{1}{n} M^{[p]}_n := \left[\int_{(0,1)} N_{j+1,[p]} N_{i+1,[p]}\right]^{n+p-2}_{i,j=1}, \tag{4.41}$$

and we note that $K^{[p]}_n$ and $M^{[p]}_n$ are SPD, while $H^{[p]}_n$ is skew-symmetric.

Proof. Copy the proof of Theorem 3.1. □

4.3.1 Construction of $K_n^{[p]}$, $H_n^{[p]}$, $M_n^{[p]}$

We now provide the construction of the 'pieces' that compose the Galerkin B-spline IgA stiffness matrix $A_n^{[p]}$, i.e., the matrices defined in (4.41). We begin with some observations concerning the B-spline basis functions $N_{i,[p]}$, $j = 2, \dots, n+p-1$. First, using the notation introduced in Definition 4.1, the support of $N_{i,[p]}$ is $[t_i, t_{i+p+1}]$; see [21]. This immediately implies that $K_n^{[p]}$, $H_n^{[p]}$, $M_n^{[p]}$ have a $(2p+1)$-band structure, because the nonzero entries in the i-th row of these matrices are at most the entries with column index $j \in \{i-p, \dots, i+p\}$. Second, the 'central' basis functions $N_{i,[p]}$, $i = p+1, \dots, n$, are 'uniformly shifted and scaled versions' of the cardinal B-spline $\phi_{[p]}$. More precisely, we have

$$N_{i,[p]}(x) = \phi_{[p]}(nx - i + p + 1), \quad i = p+1, \dots, n, \tag{4.42}$$

and

$$N'_{i,[p]}(x) = n\dot{\phi}_{[p]}(nx - i + p + 1), \quad i = p+1, \dots, n.$$

We now focus on the construction of the 'central part' of the matrices $K_n^{[p]}$, $H_n^{[p]}$, $M_n^{[p]}$, which is the part determined only by the central basis functions in (4.42). In other words, we focus on the submatrices

$$[(K_n^{[p]})_{ij}]_{i,j=p}^{n-1}, \qquad [(H_n^{[p]})_{ij}]_{i,j=p}^{n-1}, \qquad [(M_n^{[p]})_{ij}]_{i,j=p}^{n-1}, \tag{4.43}$$

which are nonempty for $n \geq p+1$. For $i, j = p, \dots, n-1$,

$$\begin{aligned}
(K_n^{[p]})_{ij} &= \frac{1}{n}\int_0^1 N'_{j+1,[p]}(x)N'_{i+1,[p]}(x)dx = n\int_0^1 \dot{\phi}_{[p]}(nx-j+p)\dot{\phi}_{[p]}(nx-i+p)dx = \int_{-i+p}^{n-i+p} \dot{\phi}_{[p]}(t+i-j)\dot{\phi}_{[p]}(t)dt \\
&= \int_{\mathbb{R}} \dot{\phi}_{[p]}(t+i-j)\dot{\phi}_{[p]}(t)dt \qquad \text{(because } [-i+p, n-i+p] \supseteq [0, p+1] = \operatorname{supp}(\phi_{[p]}) \text{, since } i \in \{p, \dots, n-1\}) \\
&= -\ddot{\phi}_{[2p+1]}(p+1+i-j) \qquad \text{(by Lemma 4.2)} \\
&= -\ddot{\phi}_{[2p+1]}(p+1-i+j) \qquad \text{(by Lemma 4.1)},
\end{aligned}$$

and similarly we obtain

$$(H_n^{[p]})_{ij} = \dot{\phi}_{[2p+1]}(p+1+i-j) = -\dot{\phi}_{[2p+1]}(p+1-i+j),$$

$$(M_n^{[p]})_{ij} = \phi_{[2p+1]}(p+1+i-j) = \phi_{[2p+1]}(p+1-i+j).$$

Since the entries of the submatrices (4.43) only depend on the difference $i-j$, these submatrices are (1-level) Toeplitz matrices. In particular, we have

$$[(K_n^{[p]})_{ij}]_{i,j=p}^{n-1} = [-\ddot{\phi}_{[2p+1]}(p+1-i+j)]_{i,j=p}^{n-1} = T_{n-p}(f_p), \tag{4.44}$$

$$[(M_n^{[p]})_{ij}]_{i,j=p}^{n-1} = [\phi_{[2p+1]}(p+1-i+j)]_{i,j=p}^{n-1} = T_{n-p}(h_p), \tag{4.45}$$

where

$$f_p(\theta) := \sum_{k\in\mathbb{Z}} -\ddot{\phi}_{[2p+1]}(p+1-k)\mathrm{e}^{\mathrm{i}k\theta} = -\ddot{\phi}_{[2p+1]}(p+1) - 2\sum_{k=1}^{p} \ddot{\phi}_{[2p+1]}(p+1-k)\cos(k\theta), \tag{4.46}$$

$$h_p(\theta) := \sum_{k\in\mathbb{Z}} \phi_{[2p+1]}(p+1-k)\mathrm{e}^{\mathrm{i}k\theta} = \phi_{[2p+1]}(p+1) + 2\sum_{k=1}^{p} \phi_{[2p+1]}(p+1-k)\cos(k\theta). \tag{4.47}$$

We end this subsection by giving the definition of 'central rows' of $K_n^{[p]}$, $H_n^{[p]}$, $M_n^{[p]}$. They are defined as the rows corresponding to an index i such that $\{i-p, \dots, i+p\} \subseteq \{p, \dots, n-1\}$ or, equivalently, $i \in$

$\{2p, \dots, n-p-1\}$. Clearly, a central row exists if and only if $n \geq 3p+1$. As observed above, $\{i-p, \dots, i+p\}$ is the set of column indices j corresponding to the nonzero entries of $K_n^{[p]}$, $H_n^{[p]}$, $M_n^{[p]}$ in the i-th row, while $\{p, \dots, n-1\}$ is the range of indices identifying the submatrices (4.43). Therefore, the generic central row of $K_n^{[p]}$ can be expressed as

$$\begin{bmatrix} 0 & \cdots & 0 & -\ddot{\phi}_{[2p+1]}(1) & \cdots & -\ddot{\phi}_{[2p+1]}(p) & -\ddot{\phi}_{[2p+1]}(p+1) & -\ddot{\phi}_{[2p+1]}(p) & \cdots & -\ddot{\phi}_{[2p+1]}(1) & 0 & \cdots & 0 \end{bmatrix}, \quad (4.48)$$

and in particular, by (4.28), the diagonal element can be expressed as $(K_n^{[p]})_{ii} = 2\dot{\phi}_{[2p]}(p) > 0$. The generic central row of $H_n^{[p]}$ can be expressed as

$$\begin{bmatrix} 0 & \cdots & 0 & -\dot{\phi}_{[2p+1]}(1) & \cdots & -\dot{\phi}_{[2p+1]}(p) & 0 & \dot{\phi}_{[2p+1]}(p) & \cdots & \dot{\phi}_{[2p+1]}(1) & 0 & \cdots & 0 \end{bmatrix}, \quad (4.49)$$

where we remark that $(H_n^{[p]})_{ii} = \dot{\phi}_{[2p+1]}(p+1) = 0$ (see Lemma 4.1). The generic central row of $M_n^{[p]}$ can be expressed as

$$\begin{bmatrix} 0 & \cdots & 0 & \phi_{[2p+1]}(1) & \cdots & \phi_{[2p+1]}(p) & \phi_{[2p+1]}(p+1) & \phi_{[2p+1]}(p) & \cdots & \phi_{[2p+1]}(1) & 0 & \cdots & 0 \end{bmatrix}. \quad (4.50)$$

We note that, for all $i = 2p, \dots, n-p-1$, the i-th central row of $K_n^{[p]}$ coincides with the i-th row of the Toeplitz matrix $T_{n+p-2}(f_p)$. Similarly, the i-th central row of $M_n^{[p]}$ coincide with the i-th row of $T_{n+p-2}(h_p)$.

Remark 4.1. Considering the recurrence relations for derivatives (4.21), for the computation of the matrix elements in (4.48)–(4.50) we only need to evaluate cardinal B-splines at integer points. We sum up some possibilities to evaluate $\phi_{[p]}$ at integer positions.

1. The values of $\phi_{[p]}$ at the integers can be obtained by using the recurrence relation (4.14): we have $\phi_{[0]}(k) = \delta_{0k}$ and $\phi_{[1]}(k) = \delta_{1k}$ for all $k \in \mathbb{Z}$, and
$$\phi_{[p]}(k) = \frac{k}{p}\phi_{[p-1]}(k) + \frac{p+1-k}{p}\phi_{[p-1]}(k-1), \quad k \in \mathbb{Z}, \quad p \geq 1.$$
2. From (4.15) it follows that the non-zero values of $\phi_{[p]}$ at the integers are equal to
$$\phi_{[p]}(k) = \frac{1}{p!}\sum_{i=0}^{k-1}\binom{p+1}{i}(-1)^i(k-i)^p, \quad k = 1, \dots, p.$$

4.4 Properties of $f_p(\theta)$ and $h_p(\theta)$

The results in this section provide some interesting properties of the functions $f_p(\theta)$ and $h_p(\theta)$. We shall see later that these functions appear in the expression of the spectral symbol that characterizes the asymptotic spectral distribution of the Galerkin B-spline IgA stiffness matrices.

We begin with the observation that $f_p(\theta)$ and $h_p(\theta)$ are defined for all $p \geq 1$ by (4.46)–(4.47). However, the right-hand side of (4.47) is well-defined also in the case $p = 0$, and we take it as the definition of $h_0(\theta)$:

$$h_0(\theta) := 1.$$

On the contrary, we cannot extend the definition of $f_p(\theta)$ to the case $p = 0$, because the right-hand side of (4.46) has no meaning for $p = 0$, since $\ddot{\phi}_{[1]}(1)$ is not defined. So, while $h_p(\theta)$ is now defined for all $p \geq 0$, $f_p(\theta)$ is still defined only for $p \geq 1$.

Lemma 4.4. *Let $p \geq 0$, let $h_p : [-\pi, \pi] \to \mathbb{R}$ be the function defined in* (4.47)*, and let $m_{h_p} := \min_{\theta \in [-\pi,\pi]} h_p(\theta)$. Then the following properties hold.*

1. $h_p(\theta) = \sum_{k\in\mathbb{Z}} \left|\widehat{\phi_{[p]}}(\theta + 2k\pi)\right|^2$.

2. $\max_{\theta\in[-\pi,\pi]} h_p(\theta) = h_p(0) = 1$ *and* $m_{h_p} \geq \left(\frac{4}{\pi^2}\right)^{p+1}$.

3. $h_p(\frac{\pi}{2}) = 2^p h_p(\pi)$. *In particular,* $h_p(\pi) \to 0$ *exponentially as* $p \to \infty$.

Proof. From the symmetry property (4.18), relation (4.24) and Theorem 4.1, it follows that

$$h_p(\theta) = \sum_{k\in\mathbb{Z}} \phi_{[2p+1]}(p+1-k)\mathrm{e}^{\mathrm{i}k\theta} = \sum_{k\in\mathbb{Z}} \left(\int_{\mathbb{R}} \phi_{[p]}(t)\phi_{[p]}(t-k)dt\right)\mathrm{e}^{\mathrm{i}k\theta} = \sum_{k\in\mathbb{Z}} \left|\widehat{\phi_{[p]}}(\theta + 2k\pi)\right|^2.$$

The inequalities (4.36)–(4.37) imply that

$$\left(\frac{4}{\pi^2}\right)^{p+1} \leq h_p(\theta) \leq 1, \qquad \theta \in [-\pi,\pi]. \tag{4.51}$$

In addition, by the partition of unity property (4.20) we get

$$h_p(0) = \phi_{[2p+1]}(p+1) + 2\sum_{k=1}^{p} \phi_{[2p+1]}(p+1-k) = \sum_{k=1}^{2p+1} \phi_{[2p+1]}(k) = 1.$$

We now prove item 3. From item 1 and (4.33) we know that

$$h_p(\theta) = \sum_{k\in\mathbb{Z}} \left|\widehat{\phi_{[p]}}(\theta + 2k\pi)\right|^2 = \sum_{k\in\mathbb{Z}} \left(\frac{2-2\cos\theta}{(\theta+2k\pi)^2}\right)^{p+1}.$$

Hence,

$$h_p\left(\frac{\pi}{2}\right) = \sum_{k\in\mathbb{Z}} \left(\frac{2}{(\frac{\pi}{2}+2k\pi)^2}\right)^{p+1} = \frac{2^{3p+3}}{\pi^{2p+2}} \sum_{k\in\mathbb{Z}} \frac{1}{(4k+1)^{2p+2}}, \tag{4.52}$$

$$h_p(\pi) = \sum_{k\in\mathbb{Z}} \left(\frac{4}{(\pi+2k\pi)^2}\right)^{p+1} = \frac{2^{2p+2}}{\pi^{2p+2}} \sum_{k\in\mathbb{Z}} \frac{1}{(2k+1)^{2p+2}}. \tag{4.53}$$

By splitting the latter sum into a sum over the even integers and a sum over the odd integers, we get

$$\begin{aligned}\sum_{k\in\mathbb{Z}} \frac{1}{(2k+1)^{2p+2}} &= \sum_{l\in\mathbb{Z}} \frac{1}{(4l+1)^{2p+2}} + \sum_{l\in\mathbb{Z}} \frac{1}{(4l+3)^{2p+2}} = \sum_{l\in\mathbb{Z}} \frac{1}{(4l+1)^{2p+2}} + \sum_{m\in\mathbb{Z}} \frac{1}{(-4m-1)^{2p+2}} \\ &= \sum_{l\in\mathbb{Z}} \frac{1}{(4l+1)^{2p+2}} + \sum_{m\in\mathbb{Z}} \frac{1}{(4m+1)^{2p+2}} = 2\sum_{k\in\mathbb{Z}} \frac{1}{(4k+1)^{2p+2}}.\end{aligned} \tag{4.54}$$

Therefore, by combining (4.54) with (4.52) and (4.53), we obtain

$$\frac{h_p(\frac{\pi}{2})}{h_p(\pi)} = \frac{2^{3p+3}}{2^{2p+2}} \frac{\sum_{k\in\mathbb{Z}} \frac{1}{(4k+1)^{2p+2}}}{2\sum_{k\in\mathbb{Z}} \frac{1}{(4k+1)^{2p+2}}} = 2^p.$$

□

Lemma 4.5. *Let* $p \geq 1$, *let* $f_p : [-\pi,\pi] \to \mathbb{R}$ *be the function defined in* (4.46), *and let* $M_{f_p} := \max_{\theta\in[-\pi,\pi]} f_p(\theta)$. *Then the following properties hold.*

1. *For all* $\theta \in [-\pi, \pi]$,

$$f_p(\theta) = (2 - 2\cos\theta) \sum_{k\in\mathbb{Z}} \left|\widehat{\phi_{[p-1]}}(\theta + 2k\pi)\right|^2 = (2 - 2\cos\theta) h_{p-1}(\theta), \tag{4.55}$$

and

$$(2 - 2\cos\theta)\left(\frac{4}{\pi^2}\right)^p \leq f_p(\theta) \leq \min\left(2 - 2\cos\theta, (2 - 2\cos\theta)^{p+1}\left(\frac{1}{\theta^{2p}} + \frac{1}{6\,\pi^{2p-2}}\right)\right).$$

2. $\min_{\theta\in[-\pi,\pi]} f_p(\theta) = f_p(0) = 0$, $\theta = 0$ *is the unique zero of* f_p *over* $[-\pi, \pi]$ *and it has order 2, because*

$$\lim_{\theta\to 0} \frac{f_p(\theta)}{\theta^2} = 1. \tag{4.56}$$

Moreover,

$$M_{f_p} \leq \min\left(4, \frac{8}{p+1} + \frac{2\pi^2}{3}\left(\frac{4}{\pi^2}\right)^p, 2\dot{\phi}_{[2p]}(p) + 2\sum_{k=1}^{p} \left|\ddot{\phi}_{[2p+1]}(p+1-k)\right|\right).$$

In particular, $M_{f_p} \to 0$ *as* $p \to \infty$.

3. $f_p\left(\frac{\pi}{2}\right) = 2^{p-2} f_p(\pi)$. *In particular,* $\dfrac{f_p(\pi)}{M_{f_p}} \to 0$ *exponentially as* $p \to \infty$.

Proof. 1. We recall from (4.35) that, for every $\theta \in [-\pi, \pi]$,

$$\widehat{\dot{\phi}_{[p]}}(\theta) = (1 - \mathrm{e}^{-\mathrm{i}\theta})\widehat{\phi_{[p-1]}}(\theta)$$

and

$$\left|\widehat{\dot{\phi}_{[p]}}(\theta)\right|^2 = (2 - 2\cos\theta)\left|\widehat{\phi_{[p-1]}}(\theta)\right|^2.$$

This implies that

$$\sum_{k\in\mathbb{Z}} \left|\widehat{\dot{\phi}_{[p]}}(\theta + 2k\pi)\right|^2 = (2 - 2\cos\theta) \sum_{k\in\mathbb{Z}} \left|\widehat{\phi_{[p-1]}}(\theta + 2k\pi)\right|^2. \tag{4.57}$$

The equality (4.55) follows from (4.23), Theorem 4.1 and (4.57) in the following way:

$$\begin{aligned} f_p(\theta) &= \sum_{k\in\mathbb{Z}} -\ddot{\phi}_{[2p+1]}(p+1-k)\mathrm{e}^{\mathrm{i}k\theta} = \sum_{k\in\mathbb{Z}} \left(\int_{\mathbb{R}} \dot{\phi}_{[p]}(t)\dot{\phi}_{[p]}(t-k)dt\right)\mathrm{e}^{\mathrm{i}k\theta} \\ &= \sum_{k\in\mathbb{Z}} \left|\widehat{\dot{\phi}_{[p]}}(\theta + 2k\pi)\right|^2 = (2 - 2\cos\theta) \sum_{k\in\mathbb{Z}} \left|\widehat{\phi_{[p-1]}}(\theta + 2k\pi)\right|^2. \end{aligned}$$

From (4.55) and from the inequalities (4.36)–(4.37), we get

$$(2 - 2\cos\theta)\left(\frac{4}{\pi^2}\right)^p \leq f_p(\theta) \leq 2 - 2\cos\theta, \quad \forall\theta \in [-\pi, \pi]. \tag{4.58}$$

Furthermore, using (4.33) in the expression of $f_p(\theta)$ given by (4.55), we obtain

$$f_p(\theta) = (2 - 2\cos\theta) \sum_{k\in\mathbb{Z}} \left(\frac{2 - 2\cos(\theta + 2k\pi)}{(\theta + 2k\pi)^2}\right)^p = (2 - 2\cos\theta)^{p+1} \sum_{k\in\mathbb{Z}} \frac{1}{(\theta + 2k\pi)^{2p}}. \tag{4.59}$$

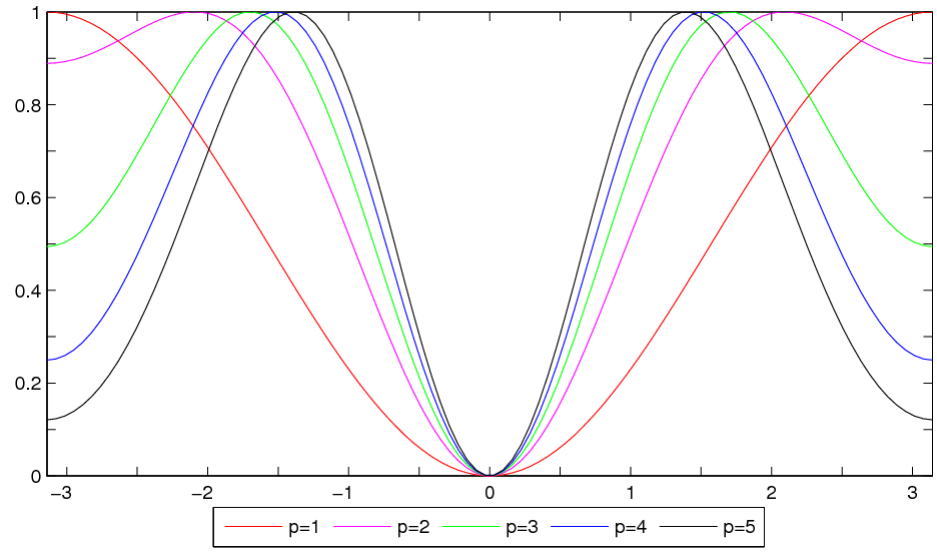


Figure 4.3: graph of f_p/M_{f_p} for $p = 1, \dots, 5$.

p	1	2	3	4	5	6	7	8	9	10	11	12
$\frac{f_p(\pi)}{M_{f_p}}$	1.0000	0.8889	0.4941	0.2494	0.1209	0.0570	0.0264	0.0120	0.0054	0.0024	0.0011	0.0005
$\frac{f_p(\frac{2\pi}{3})}{M_{f_p}}$	0.7500	1.0000	0.9034	0.7613	0.6209	0.4939	0.3853	0.2960	0.2247	0.1689	0.1259	0.0932

Table 4.1: values of $f_p(\pi)/M_{f_p}$ and $f_p(\frac{2\pi}{3})/M_{f_p}$ for $p = 1, \dots, 12$.

Now observe that

$$\sum_{k\in\mathbb{Z}} \frac{1}{(\theta+2k\pi)^{2p}} = \frac{1}{\theta^{2p}} + \sum_{k=1}^{\infty} \frac{1}{(\theta+2k\pi)^{2p}} + \sum_{k=1}^{\infty} \frac{1}{(-\theta+2k\pi)^{2p}} \le \frac{1}{\theta^{2p}} + \sum_{k=1}^{\infty} \frac{1}{(2k\pi)^{2p}} + \sum_{k=1}^{\infty} \frac{1}{(-\pi+2k\pi)^{2p}}$$
$$\le \frac{1}{\theta^{2p}} + \frac{1}{\pi^{2p}} \left(\sum_{k=1}^{\infty} \frac{1}{(2k)^2} + \sum_{k=1}^{\infty} \frac{1}{(2k-1)^2} \right) = \frac{1}{\theta^{2p}} + \frac{1}{6\,\pi^{2p-2}}.$$

By (4.59), the latter inequality yields

$$f_p(\theta) \le (2-2\cos\theta)^{p+1} \left(\frac{1}{\theta^{2p}} + \frac{1}{6\,\pi^{2p-2}} \right), \quad \forall \theta \in [-\pi, \pi]. \tag{4.60}$$

This proves the first statement in the lemma.

2. The inequalities in (4.58) imply that $\min_{\theta\in[-\pi,\pi]} f_p(\theta) = f_p(0) = 0$, that $\theta = 0$ is the only zero of f_p, and that $M_{f_p} \le 4$. Moreover, (4.55) together with the fact that $h_{p-1}(0) = 1$, gives (4.56). In order to prove that $M_{f_p} \le \frac{8}{p+1} + \frac{2\pi^2}{3} \left(\frac{4}{\pi^2}\right)^p$, we use the inequalities

$$2 - 2\cos\theta \le \theta^2 - \frac{\theta^4}{18} \le \theta^2, \quad \forall \theta \in [-\pi, \pi].$$

It follows that

$$(2-2\cos\theta) \left(\frac{2-2\cos\theta}{\theta^2} \right)^p \le \theta^2 \left(1 - \frac{\theta^2}{18} \right)^p, \quad \forall \theta \in [-\pi, \pi].$$

If $p \geq 2$, the maximum of $\theta^2\left(1-\frac{\theta^2}{18}\right)^p$ over $[-\pi,\pi]$ is located at $\theta^2 = \frac{18}{p+1}$ and its value is given by

$$\frac{18}{p+1}\left(1-\frac{1}{p+1}\right)^p \leq \frac{8}{p+1}.$$

Therefore, if $p \geq 2$, we have

$$\frac{(2-2\cos\theta)^{p+1}}{\theta^{2p}} \leq \frac{8}{p+1}, \qquad \forall \theta \in [-\pi,\pi]. \tag{4.61}$$

Moreover,

$$\frac{(2-2\cos\theta)^{p+1}}{6\,\pi^{2p-2}} \leq \frac{4^{p+1}}{6\,\pi^{2p-2}}, \qquad \forall \theta \in [-\pi,\pi]. \tag{4.62}$$

Recalling (4.60), the inequalities (4.61)–(4.62) prove that, for $p \geq 2$,

$$M_{f_p} \leq \frac{8}{p+1} + \frac{2\pi^2}{3}\left(\frac{4}{\pi^2}\right)^p. \tag{4.63}$$

In addition, (4.63) holds for $p = 1$ too, because $f_1(\theta) = 2-2\cos\theta$ and $M_{f_1} = 4$. To complete the proof of the second statement, we still have to show that

$$M_{f_p} \leq 2\dot{\phi}_{[2p]}(p) + 2\sum_{k=1}^{p} |\ddot{\phi}_{[2p+1]}(p+1-k)|, \tag{4.64}$$

which is easily obtained by using (4.28) and (4.46).

3. Item 3 follows from item 1 and item 3 in Lemma 4.4. □

Using item 1 in Lemma 4.5 and the definition (4.47) of h_{p-1}, we see that

$$f_p(\theta) = (2-2\cos\theta)h_{p-1}(\theta) = (2-2\cos\theta)\left(\phi_{[2p-1]}(p) + 2\sum_{k=1}^{p-1}\phi_{[2p-1]}(p-k)\cos(k\theta)\right).$$

This is a more elegant and efficient formula to evaluate f_p.

Figure 4.3 shows the graph of f_p normalized by its maximum M_{f_p}, for $p = 1,\dots,5$. As predicted by Lemma 4.5, the value $f_p(\pi)/M_{f_p}$ decreases exponentially to zero as $p \to \infty$; cf. Table 4.1. From a numerical viewpoint, we can say that, for large p, the normalized function f_p/M_{f_p} vanishes not only at $\theta = 0$ but also at $\theta = \pi$. In reality, we see from Figure 4.3 and Table 4.1 that f_p/M_{f_p} approaches zero for very large p in a whole interval containing $[2\pi/3,\pi]$.

Figure 4.4 shows the graph of h_p for $p = 0,\dots,4$. We see that the behavior of h_p over the interval $[2\pi/3,\pi]$ is analogous to the one of f_p/M_{f_p} over the same interval.

4.5 Spectral analysis and spectral symbol

In this section we follow the same program as in Section 3.4. We study the spectral properties of the stiffness matrix $A_{\boldsymbol{n}}^{[\boldsymbol{p}]}$ in (4.6), focusing on the asymptotic behavior as the fineness parameters $\boldsymbol{n} \to \infty$. In particular, we give estimates for the eigenvalues and for the spectral condition number $\kappa(A_{\boldsymbol{n}}^{[\boldsymbol{p}]})$. Moreover, assuming $\boldsymbol{n} = \boldsymbol{\nu}n = (\nu_1 n,\dots,\nu_d n) \in \mathbb{N}^d$ for a fixed $\boldsymbol{\nu} \in \mathbb{Q}_+^d$, we prove that the sequence $\{n^{d-2}A_{\boldsymbol{n}}^{[\boldsymbol{p}]}\}_n$ has an asymptotic spectral distribution characterized by the real function $f_{\boldsymbol{p}}^{(\boldsymbol{\nu})} : [-\pi,\pi]^d \to \mathbb{R}$,

$$\begin{aligned} f_{\boldsymbol{p}}^{(\boldsymbol{\nu})}(\boldsymbol{\theta}) &:= \sum_{k=1}^{d} c_k(\boldsymbol{\nu})\,(h_{p_1}\otimes\cdots\otimes h_{p_{k-1}}\otimes f_{p_k}\otimes h_{p_{k+1}}\otimes\cdots\otimes h_{p_d})(\boldsymbol{\theta}), \\ &= \sum_{k=1}^{d} c_k(\boldsymbol{\nu})\,h_{p_1}(\theta_1)\cdots h_{p_{k-1}}(\theta_{k-1})f_{p_k}(\theta_k)h_{p_{k+1}}(\theta_{k+1})\cdots h_{p_d}(\theta_d), \end{aligned} \tag{4.65}$$

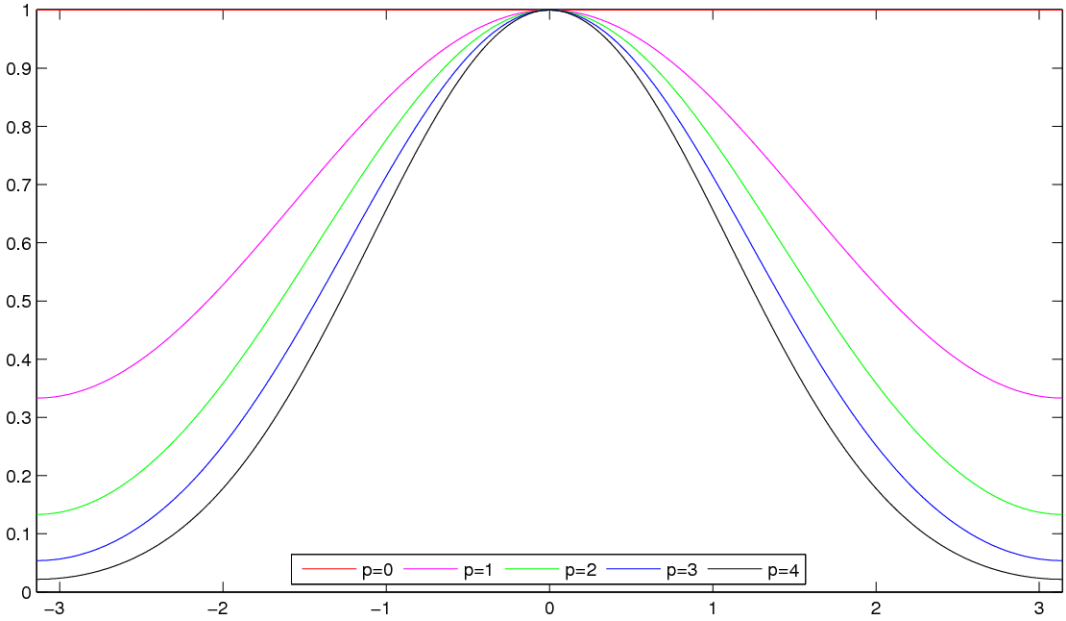


Figure 4.4: graph of h_p for $p = 0, \ldots, 4$.

where f_p and h_p are defined in in (4.46)–(4.47) and

$$c_k(\boldsymbol{\nu}) := \frac{\nu_k}{\nu_1 \cdots \nu_{k-1} \nu_{k+1} \cdots \nu_d}, \quad k = 1, \ldots, d. \tag{4.66}$$

4.5.1 Estimates for the eigenvalues, localization of the spectrum and conditioning of $A_{\boldsymbol{n}}^{[\boldsymbol{p}]}$

We first provide suitable estimates for the minimal eigenvalues of $M_n^{[p]}$ and $K_n^{[p]}$. These will be employed to obtain a lower bound for $\lambda_{\min}(\Re(A_{\boldsymbol{n}}^{[\boldsymbol{p}]}))$, which, in turn, will be used in combination with the Fan-Hoffman theorem (Theorem 1.1) to obtain an upper bound for the spectral condition number $\kappa(A_{\boldsymbol{n}}^{[\boldsymbol{p}]})$. We begin with the following result [54].

Lemma 4.6. *Let $p, n \geq 1$ and $\mathbf{x} = (x_1, \ldots, x_{n+p-2}) \in \mathbb{R}^{n+p-2}$, then*

$$C_p \frac{\|\mathbf{x}\|^2}{n} \leq \left\| \sum_{i=1}^{n+p-2} x_i N_{i+1,[p]} \right\|_{L^2(0,1)}^2 \leq \bar{C}_p \frac{\|\mathbf{x}\|^2}{n}, \tag{4.67}$$

where $C_p, \bar{C}_p > 0$ are constants that do not depend on n and $\mathbf{x}$.

The inequalities in (4.67) are a special instance for the L^2-norm of the results stated in [54, Theorem 9.27]. We remark that the quantity $\bar{\Delta}$ used in the cited theorem in our context has the value $\frac{1}{n}$; see [54, Eq. (6.3)].

Theorem 4.3. *Let $C_p > 0$ be a constant for which the left inequality in* (4.67) *is satisfied. Then, for all $p, n \geq 1$ the following properties hold.*

1. $\lambda_{\min}(M_n^{[p]}) \geq C_p$.
2. $K_n^{[p]} \geq \dfrac{\pi^2}{n^2} M_n^{[p]}$ *and* $\lambda_{\min}(K_n^{[p]}) \geq \dfrac{\pi^2 C_p}{n^2}$.

Proof. Using the definition of $M_n^{[p]}$, see (4.41), for all $\mathbf{y} \in \mathbb{R}^{n+p-2}$ we have

$$\begin{aligned}\mathbf{y}^T \left(\frac{1}{n} M_n^{[p]}\right) \mathbf{y} &= \sum_{i,j=1}^{n+p-2} \left(\frac{1}{n} M_n^{[p]}\right)_{i,j} y_i y_j = \sum_{i,j=1}^{n+p-2} \int_0^1 y_i y_j N_{j+1,[p]}(x) N_{i+1,[p]}(x) dx \\ &= \int_0^1 \sum_{i=1}^{n+p-2} y_i N_{i+1,[p]}(x) \sum_{j=1}^{n+p-2} y_j N_{j+1,[p]}(x) dx = \int_0^1 \left(\sum_{i=1}^{n+p-2} y_i N_{i+1,[p]}(x)\right)^2 dx \\ &= \left\| \sum_{i=1}^{n+p-2} y_i N_{i+1,[p]} \right\|_{L^2(0,1)}^2 \geq C_p \frac{\|\mathbf{y}\|^2}{n}, \end{aligned} \tag{4.68}$$

where the last inequality holds by (4.67). Hence, we get $\mathbf{y}^T M_n^{[p]} \mathbf{y} \geq C_p \|\mathbf{y}\|^2$, and from the minimax principle it follows that

$$\lambda_{\min}(M_n^{[p]}) = \min_{\mathbf{y} \neq \mathbf{0}} \frac{\mathbf{y}^T M_n^{[p]} \mathbf{y}}{\|\mathbf{y}\|^2} \geq C_p. \tag{4.69}$$

This proves the first statement. To prove the second statement, we follow the same argument used in the proof of Lemma 3.5, which, in fact, can be extended to a much more general setting; see [31, Proposition 1]. Using the definition of $K_n^{[p]}$, see (4.41), for all $\mathbf{y} \in \mathbb{R}^{n+p-2}$ we obtain

$$\begin{aligned}\mathbf{y}^T \left(n K_n^{[p]}\right) \mathbf{y} &= \sum_{i,j=1}^{n+p-2} \left(n K_n^{[p]}\right)_{i,j} y_i y_j = \sum_{i,j=1}^{n+p-2} \int_0^1 y_i y_j N'_{j+1,[p]}(x) N'_{i+1,[p]}(x) dx \\ &= \int_0^1 \sum_{i=1}^{n+p-2} y_i N'_{i+1,[p]}(x) \sum_{j=1}^{n+p-2} y_j N'_{j+1,[p]}(x) dx = \int_0^1 \left(\sum_{i=1}^{n+p-2} y_i N'_{i+1,[p]}(x)\right)^2 dx \\ &= \left\| \sum_{i=1}^{n+p-2} y_i N'_{i+1,[p]} \right\|_{L^2(0,1)}^2 = \|v'_{\mathbf{y}}\|_{L^2(0,1)}^2, \end{aligned} \tag{4.70}$$

where $v_{\mathbf{y}} := \sum_{i=1}^{n+p-2} y_i N_{i+1,[p]} \in W_n^{[p]}$; see Section 4.1 for the definition of $W_n^{[p]}$. Since $W_n^{[p]} \subset H_0^1(0,1)$, we may apply the Poincaré inequality (3.42). From (3.42) and (4.68) it follows that

$$\mathbf{y}^T \left(n K_n^{[p]}\right) \mathbf{y} = \|v'_{\mathbf{y}}\|_{L^2(0,1)}^2 \geq \pi^2 \|v_{\mathbf{y}}\|_{L^2(0,1)}^2 = \mathbf{y}^T \left(\frac{\pi^2}{n} M_n^{[p]}\right) \mathbf{y}.$$

Dividing both sides by n we obtain, for all $\mathbf{y} \in \mathbb{R}^{n+p-2}$,

$$\mathbf{y}^T K_n^{[p]} \mathbf{y} \geq \mathbf{y}^T \left(\frac{\pi^2}{n^2} M_n^{[p]}\right) \mathbf{y}.$$

This proves that $K_n^{[p]} \geq \frac{\pi^2}{n^2} M_n^{[p]}$. The application of the minimax principle and (4.69) yields

$$\lambda_{\min}(K_n^{[p]}) = \min_{\mathbf{y} \neq \mathbf{0}} \frac{\mathbf{y}^T K_n^{[p]} \mathbf{y}}{\|\mathbf{y}\|^2} \geq \min_{\mathbf{y} \neq \mathbf{0}} \frac{\mathbf{y}^T \left(\frac{\pi^2}{n^2} M_n^{[p]}\right) \mathbf{y}}{\|\mathbf{y}\|^2} = \frac{\pi^2}{n^2} \lambda_{\min}(M_n^{[p]}) \geq \frac{\pi^2 C_p}{n^2},$$

which concludes the proof. □

Remark 4.2. Suppose that, for a given $p \geq 1$, we are able to find a constant $\tilde{C}_p > 0$ such that[1]

$$\lambda_{\min}(M_n^{[p]}) \geq \tilde{C}_p.$$

[1]Such a constant $\tilde{C}_p$ may be found, e.g., by using the Gershgorin theorems [8]. We refer to Remark 4.6 for an example.

n	$\lambda_{\min}(M_n^{[2]})$	$\lambda_{\min}(M_n^{[3]})$	$\lambda_{\min}(M_n^{[4]})$	$n^2\,\lambda_{\min}(K_n^{[2]})$	$n^2\,\lambda_{\min}(K_n^{[3]})$	$n^2\,\lambda_{\min}(K_n^{[4]})$
20	0.1333333	0.0482607	0.0171864	9.8089070	9.7834046	9.7507398
40	0.1333333	0.0486447	0.0173795	9.8543957	9.8486563	9.8419964
80	0.1333333	0.0486538	0.0173821	9.8658001	9.8644478	9.8629796
160	0.1333333	0.0486538	0.0173821	9.8686532	9.8683256	9.8679834
320	0.1333333	0.0486538	0.0173821	9.8693666	9.8692860	9.8692036
640	0.1333333	0.0486538	0.0173821	9.8695450	9.8695250	9.8695048
1280	0.1333333	0.0486538	0.0173821	9.8695896	9.8695846	9.8695796

Table 4.2: computation of $\lambda_{\min}(M_n^{[p]})$ and $n^2\,\lambda_{\min}(K_n^{[p]})$ for $p = 2, 3, 4$ and for increasing values of n.

In this case, items 1 and 2 in Theorem 4.3 hold with $\tilde{C}_p$ in place of C_p. Moreover, the left inequality in (4.67) also holds with $\tilde{C}_p$ in place of C_p. Indeed, by using a similar argument as in the proof of Theorem 4.3, we obtain

$$\frac{n\left\|\sum_{i=1}^{n+p-2} x_i N_{i+1,[p]}\right\|^2_{L^2(0,1)}}{\|\mathbf{x}\|^2} = \frac{\mathbf{x}^T M_n^{[p]}\mathbf{x}}{\|\mathbf{x}\|^2} \geq \min_{\mathbf{y}\neq 0} \frac{\mathbf{y}^T M_n^{[p]}\mathbf{y}}{\|\mathbf{y}\|^2} = \lambda_{\min}(M_n^{[p]}) \geq \tilde{C}_p.$$

Table 4.2 shows the results of some numerical experiments performed on the matrices $M_n^{[p]}$ and $K_n^{[p]}$ for $p = 2, 3, 4$ and for increasing values of n. From these results it seems that

$$\lambda_{\min}(M_n^{[p]}) \overset{n\to\infty}{\sim} \tilde{m}_p, \tag{4.71}$$

with $\tilde{m}_2 = \frac{2}{15}$, $\tilde{m}_3 \approx 0.0486538$ and $\tilde{m}_4 \approx 0.0173821$. Apparently, the sequence $\lambda_{\min}(M_n^{[p]})$ converges to $\tilde{m}_p$ very quickly as $n \to \infty$. In addition, it seems that[2]

$$\lambda_{\min}(K_n^{[p]}) \overset{n\to\infty}{\sim} \frac{\pi^2}{n^2}. \tag{4.72}$$

Since $K_n^{[1]} = \text{Tridiagonal}(-1, 2, -1) \in \mathbb{R}^{(n-1)\times(n-1)}$, we note that for $\lambda_{\min}(K_n^{[1]})$ the asymptotic formula (4.72) holds, because it is known that

$$\lambda_{\min}(K_n^{[1]}) = 4\left(\sin\frac{\pi}{2n}\right)^2 \overset{n\to\infty}{\sim} \frac{\pi^2}{n^2};$$

see Theorem 1.9. The numerical experiments show that, for $p = 2, 3, 4$, the eigenvalue $\lambda_{\min}(K_n^{[p]})$ converges to 0 as n^{-2}, which means that the lower estimate $\frac{\pi^2 C_p}{n^2}$ obtained in Theorem 4.3 is asymptotically of the same order as $\lambda_{\min}(K_n^{[p]})$ when $n \to \infty$.

In addition, referring to Table 4.3, we can formulate a deeper conjecture than (4.72).

Conjecture 4.1. *For every $p \geq 1$ and for each fixed $j \geq 1$,*

$$\lim_{n\to\infty} n^2\,\lambda_{n+p-1-j}(K_n^{[p]}) = j^2\pi^2, \tag{4.73}$$

where $\lambda_{n+p-1-j}(K_n^{[p]})$ is the j-th smallest eigenvalue of $K_n^{[p]}$ (recall that $K_n^{[p]}$ has size $n+p-2$ and its eigenvalues are arranged in non-increasing order). This conjecture has a motivation completely analogous to the one given in Conjecture 3.2.

We now derive upper bounds for the infinity norms of $K_n^{[p]}$, $H_n^{[p]}$, $M_n^{[p]}$. They are needed for giving a localization of the spectrum of $A_{\boldsymbol{n}}^{[\boldsymbol{p}]}$ and for providing an upper bound for the spectral condition number $\kappa(A_{\boldsymbol{n}}^{[\boldsymbol{p}]})$.

[2]The constant π^2 is precisely $c_{1,1}$; see Remarks 1.3–1.5.

n	$n^2\lambda_{n-1}(K_n^{[2]})$	$n^2\lambda_n(K_n^{[3]})$	$n^2\lambda_{n+1}(K_n^{[4]})$	$n^2\lambda_{n-2}(K_n^{[2]})$	$n^2\lambda_{n-1}(K_n^{[3]})$	$n^2\lambda_n(K_n^{[4]})$
20	38.51599640	38.11745811	37.61719616	84.02689324	82.08515027	79.68696933
40	39.23562801	39.14429699	39.03869339	87.60193868	87.14374691	86.61710221
80	39.41758280	39.39597383	39.37252923	88.51875348	88.40959481	88.29129852
160	39.46320030	39.45796001	39.45248834	88.74942017	88.72290521	88.69522615
320	39.47461274	39.47332339	39.47200503	88.80717862	88.80065213	88.79397917
640	39.47746635	39.47714663	39.47682328	88.82162398	88.82000545	88.81836856
1280	39.47817979	39.47810019	39.47802013	88.82523569	88.82483270	88.82442742

Table 4.3: computation of $n^2\lambda_{n+p-1-j}(K_n^{[p]})$ for $p = 2, 3, 4$, for $j = 2, 3$ and for increasing values of n.

Lemma 4.7. *Let $p, n \geq 1$, then*

$$\|M_n^{[p]}\|_\infty \leq 1, \qquad \|H_n^{[p]}\|_\infty \leq 2, \qquad \|K_n^{[p]}\|_\infty \leq 4p.$$

Proof. We first note that the derivative and integral of a B-spline $N_{i,[p]}(x)$ are given by

$$N'_{i,[p]}(x) = p\left(\frac{N_{i,[p-1]}(x)}{t_{i+p} - t_i} - \frac{N_{i+1,[p-1]}(x)}{t_{i+p+1} - t_{i+1}}\right) \tag{4.74}$$

and

$$\int_{\mathbb{R}} N_{i,[p]}(x)\,dx = \frac{t_{i+p+1} - t_i}{p+1}; \tag{4.75}$$

see [21, 54]. The sequence of knots (4.2)–(4.3) implies that the maximum length of the support of any $N_{i,[p]}$ is $\frac{p+1}{n}$. Recalling (4.41), by the positivity property and the partition of unity property of B-splines, we obtain

$$\begin{aligned}\left\|\frac{1}{n}M_n^{[p]}\right\|_\infty &= \max_{i=1,\dots,n+p-2} \sum_{j=1}^{n+p-2} \int_0^1 N_{j+1,[p]}(x)N_{i+1,[p]}(x)\,dx = \max_{i=1,\dots,n+p-2} \int_0^1 \left(\sum_{j=1}^{n+p-2} N_{j+1,[p]}(x)\right) N_{i+1,[p]}(x)\,dx \\ &\leq \max_{i=1,\dots,n+p-2} \int_0^1 N_{i+1,[p]}(x)\,dx = \max_{i=1,\dots,n+p-2} \frac{t_{i+p+2} - t_{i+1}}{p+1} \leq \frac{1}{n}.\end{aligned}$$

Recalling (4.41) and using the skew-symmetry of the matrix $H_n^{[p]}$, we obtain

$$\begin{aligned}\|H_n^{[p]}\|_\infty &= \max_{i=1,\dots,n+p-2} \sum_{j=1}^{n+p-2} \left|\int_0^1 N_{j+1,[p]}(x)N'_{i+1,[p]}(x)\,dx\right| \\ &= \max_{i=1,\dots,n+p-2} p \sum_{j=1}^{n+p-2} \left|\int_0^1 N_{j+1,[p]}(x)\left(\frac{N_{i+1,[p-1]}(x)}{t_{i+p+1} - t_{i+1}} - \frac{N_{i+2,[p-1]}(x)}{t_{i+p+2} - t_{i+2}}\right) dx\right|.\end{aligned} \tag{4.76}$$

Using the partition of unity property and (4.75), we have

$$\sum_{j=1}^{n+p-2} \int_0^1 N_{j+1,[p]}(x)\frac{N_{i+1,[p-1]}(x)}{t_{i+p+1} - t_{i+1}}\,dx = \int_0^1 \left(\sum_{j=1}^{n+p-2} N_{j+1,[p]}(x)\right) \frac{N_{i+1,[p-1]}(x)}{t_{i+p+1} - t_{i+1}}\,dx \leq \frac{1}{p},$$

and a similar bound holds for the remaining term in (4.76). It follows that $\|H_n^{[p]}\|_\infty \leq 2$.

Recalling (4.41), we obtain

$$\begin{aligned}\|nK_n^{[p]}\|_\infty &= \max_{i=1,\dots,n+p-2} \sum_{j=1}^{n+p-2} \left|\int_0^1 N'_{j+1,[p]}(x)N'_{i+1,[p]}(x)\,dx\right| \\ &= \max_{i=1,\dots,n+p-2} p^2 \sum_{j=1}^{n+p-2} \left|\int_0^1 \left(\frac{N_{j+1,[p-1]}(x)}{t_{j+p+1} - t_{j+1}} - \frac{N_{j+2,[p-1]}(x)}{t_{j+p+2} - t_{j+2}}\right)\left(\frac{N_{i+1,[p-1]}(x)}{t_{i+p+1} - t_{i+1}} - \frac{N_{i+2,[p-1]}(x)}{t_{i+p+2} - t_{i+2}}\right) dx\right|.\end{aligned} \tag{4.77}$$

In addition, we have

$$\sum_{j=1}^{n+p-2}\int_0^1 \frac{N_{j+1,[p-1]}(x)}{t_{j+p+1}-t_{j+1}}\,\frac{N_{i+1,[p-1]}(x)}{t_{i+p+1}-t_{i+1}}\,dx = \int_0^1\left(\sum_{j=1}^{n+p-2}\frac{N_{j+1,[p-1]}(x)}{t_{j+p+1}-t_{j+1}}\right)\frac{N_{i+1,[p-1]}(x)}{t_{i+p+1}-t_{i+1}}\,dx \le n\int_0^1 \frac{N_{i+1,[p-1]}(x)}{t_{i+p+1}-t_{i+1}}\,dx = \frac{n}{p},$$

and in a similar way we can also bound the remaining terms in (4.77). This results in

$$\|nK_n^{[p]}\|_\infty \le \max_{i=1,\dots,n+p-2} p^2\left(4\frac{n}{p}\right) = 4pn.$$

□

Remark 4.3. A consequence of Lemma 4.7 is that we can take $\bar{C}_p = 1$ in (4.67), independently of p. Indeed, Lemma 4.7 implies that $\lambda_{\max}(M_n^{[p]}) \le \|M_n^{[p]}\|_\infty \le 1$ for all $p, n \ge 1$. Thus, by the minimax principle,

$$\frac{n\left\|\sum_{i=1}^{n+p-2} x_i N_{i+1,[p]}\right\|^2_{L^2(0,1)}}{\|\mathbf{x}\|^2} = \frac{\mathbf{x}^T M_n^{[p]}\mathbf{x}}{\|\mathbf{x}\|^2} \le \max_{\mathbf{y}\neq\mathbf{0}}\frac{\mathbf{y}^T M_n^{[p]}\mathbf{y}}{\|\mathbf{y}\|^2} = \lambda_{\max}(M_n^{[p]}) \le 1.$$

Theorem 4.4 (localization of the spectrum of $\Re(A_{\boldsymbol{n}}^{[\boldsymbol{p}]})$). *Let $\boldsymbol{p}, \boldsymbol{n} \in \mathbb{N}^d$, then*

$$\lambda_{\min}(\Re(A_{\boldsymbol{n}}^{[\boldsymbol{p}]})) \ge \frac{\pi^2 d+\gamma}{n_1\cdots n_d} C_{p_1}\cdots C_{p_d}, \tag{4.78}$$

$$\lambda_{\max}(\Re(A_{\boldsymbol{n}}^{[\boldsymbol{p}]})) \le \frac{\sum_{k=1}^d 4p_k n_k^2+\gamma}{n_1\cdots n_d}, \tag{4.79}$$

where C_p is a constant satisfying the left inequality in Lemma 4.6.

Proof. We recall from (4.11) and Theorem 4.2 that

$$\Re(A_{\boldsymbol{n}}^{[\boldsymbol{p}]}) = \sum_{k=1}^d \frac{1}{n_1}M_{n_1}^{[p_1]}\otimes\cdots\otimes\frac{1}{n_{k-1}}M_{n_{k-1}}^{[p_{k-1}]}\otimes n_k K_{n_k}^{[p_k]}\otimes\frac{1}{n_{k+1}}M_{n_{k+1}}^{[p_{k+1}]}\otimes\cdots\otimes\frac{1}{n_d}M_{n_d}^{[p_d]} + \gamma\frac{1}{n_1}M_{n_1}^{[p_1]}\otimes\cdots\otimes\frac{1}{n_d}M_{n_d}^{[p_d]}.$$

Now apply (1.8), (1.16) and Theorem 4.3 to obtain (4.78). Then, apply (1.9), (1.16) and Lemma 4.7 to obtain (4.79). □

Theorem 4.5 (localization of the spectrum of $A_{\boldsymbol{n}}^{[\boldsymbol{p}]}$). *Let $\boldsymbol{p}, \boldsymbol{n} \in \mathbb{N}^d$, then*

$$\begin{aligned}\Lambda(A_{\boldsymbol{n}}^{[\boldsymbol{p}]}) &\subseteq \left[\lambda_{\min}(\Re(A_{\boldsymbol{n}}^{[\boldsymbol{p}]})), \lambda_{\max}(\Re(A_{\boldsymbol{n}}^{[\boldsymbol{p}]}))\right]\times\left[\lambda_{\min}(\Im(A_{\boldsymbol{n}}^{[\boldsymbol{p}]})), \lambda_{\max}(\Im(A_{\boldsymbol{n}}^{[\boldsymbol{p}]}))\right]\\ &\subseteq \left[\frac{\pi^2 d+\gamma}{n_1\cdots n_d}C_{p_1}\cdots C_{p_d}, \frac{\sum_{k=1}^d 4p_k n_k^2+\gamma}{n_1\cdots n_d}\right]\times\left[-\frac{2\sum_{k=1}^d|\beta_k|n_k}{n_1\cdots n_d}, \frac{2\sum_{k=1}^d|\beta_k|n_k}{n_1\cdots n_d}\right],\end{aligned}$$

where C_p is a constant satisfying the left inequality in Lemma 4.6.

Proof. From Theorem 4.4 we have

$$\frac{\pi^2 d+\gamma}{n_1\cdots n_d}C_{p_1}\cdots C_{p_d} \le \lambda_{\min}(\Re(A_{\boldsymbol{n}}^{[\boldsymbol{p}]})) \le \lambda_{\max}(\Re(A_{\boldsymbol{n}}^{[\boldsymbol{p}]})) \le \frac{\sum_{k=1}^d 4p_k n_k^2+\gamma}{n_1\cdots n_d}. \tag{4.80}$$

Taking into account that $H_n^{[p]}$ is normal, we have $\|H_n^{[p]}\| = \rho(H_n^{[p]}) \le \|H_n^{[p]}\|_\infty$. Similarly, $\|M_n^{[p]}\| = \rho(M_n^{[p]}) \le \|M_n^{[p]}\|_\infty$. Therefore, using (4.12), (4.39), (1.13) and Lemma 4.7, we get

$$\begin{aligned}\|\Im(A_{\boldsymbol{n}}^{[\boldsymbol{p}]})\| = \|A_{\boldsymbol{n},A}^{[\boldsymbol{p}]}\| &= \left\|\sum_{k=1}^d \frac{1}{n_1}M_{n_1}^{[p_1]}\otimes\cdots\otimes\frac{1}{n_{k-1}}M_{n_{k-1}}^{[p_{k-1}]}\otimes\beta_k H_{n_k}^{[p_k]}\otimes\frac{1}{n_{k+1}}M_{n_{k+1}}^{[p_{k+1}]}\otimes\cdots\otimes\frac{1}{n_d}M_{n_d}^{[p_d]}\right\|\\ &\le \frac{1}{n_1\cdots n_d}\sum_{k=1}^d \|M_{n_1}^{[p_1]}\|\cdots\|M_{n_{k-1}}^{[p_{k-1}]}\|\,n_k|\beta_k|\,\|H_{n_k}^{[p_k]}\|\,\|M_{n_{k+1}}^{[p_{k+1}]}\|\cdots\|M_{n_d}^{[p_d]}\| \le \frac{2\sum_{k=1}^d|\beta_k|n_k}{n_1\cdots n_d}.\end{aligned}$$

n	$\kappa(A_{\boldsymbol{n}}^{[\boldsymbol{p}]})/n^2$	$\kappa(A_{\boldsymbol{n}})/n^2$	$\kappa(A_{\boldsymbol{n}}^{(\boldsymbol{p})})/n^2$
8	0.0916	0.2278	1.2597
16	0.0772	0.2173	1.2573
32	0.0763	0.2101	1.2553
64	0.0761	0.2065	1.2543
128	0.0760	0.2049	1.2539
256	0.0760	0.2041	1.2538

Table 4.4: computation of $\kappa(A_{\boldsymbol{n}}^{[\boldsymbol{p}]})/n^2$, $\kappa(A_{\boldsymbol{n}})/n^2$ and $\kappa(A_{\boldsymbol{n}}^{(\boldsymbol{p})})/n^2$ in the case $d = 2$, $\boldsymbol{\beta} = \mathbf{0}$, $\gamma = 0$, $\boldsymbol{p} = (2,2)$, $\boldsymbol{n} = (n, \log_2 n)$, for increasing values of n. Note that we are in the presence of a non-uniform mesh refinement.

It follows that

$$-\frac{2\sum_{k=1}^{d}|\beta_k|n_k}{n_1\cdots n_d} \le -\|\mathfrak{I}(A_{\boldsymbol{n}}^{[\boldsymbol{p}]})\| \le \lambda_{\min}(\mathfrak{I}(A_{\boldsymbol{n}}^{[\boldsymbol{p}]})) \le \lambda_{\max}(\mathfrak{I}(A_{\boldsymbol{n}}^{[\boldsymbol{p}]})) \le \|\mathfrak{I}(A_{\boldsymbol{n}}^{[\boldsymbol{p}]})\| \le \frac{2\sum_{k=1}^{d}|\beta_k|n_k}{n_1\cdots n_d}. \tag{4.81}$$

Combining (4.80)–(4.81) with (1.7) we get the thesis. □

Theorem 4.6 (**conditioning**). *For every $\boldsymbol{p} \in \mathbb{N}^d$ there exists a constant $\alpha_{\boldsymbol{p}}$ such that, for all $\boldsymbol{n} \in \mathbb{N}^d$,*

$$\kappa(A_{\boldsymbol{n}}^{[\boldsymbol{p}]}) \le \alpha_{\boldsymbol{p}} \sum_{k=1}^{d} n_k^2. \tag{4.82}$$

Proof. The proof is exactly the same as the proof of Theorem 3.7. From $A_{\boldsymbol{n}}^{[\boldsymbol{p}]} = \mathfrak{R}(A_{\boldsymbol{n}}^{[\boldsymbol{p}]}) + \mathrm{i}\mathfrak{I}(A_{\boldsymbol{n}}^{[\boldsymbol{p}]})$ and from the fact that $\mathfrak{R}(A_{\boldsymbol{n}}^{[\boldsymbol{p}]})$, $\mathfrak{I}(A_{\boldsymbol{n}}^{[\boldsymbol{p}]})$ are Hermitian, we have

$$\sigma_{\max}(A_{\boldsymbol{n}}^{[\boldsymbol{p}]}) = \|A_{\boldsymbol{n}}^{[\boldsymbol{p}]}\| \le \|\mathfrak{R}(A_{\boldsymbol{n}}^{[\boldsymbol{p}]})\| + \|\mathfrak{I}(A_{\boldsymbol{n}}^{[\boldsymbol{p}]})\| = \rho(\mathfrak{R}(A_{\boldsymbol{n}}^{[\boldsymbol{p}]})) + \rho(\mathfrak{I}(A_{\boldsymbol{n}}^{[\boldsymbol{p}]})).$$

Hence, by Theorem 4.5,

$$\|A_{\boldsymbol{n}}^{[\boldsymbol{p}]}\| \le \hat{\alpha}_{\boldsymbol{p}} \frac{\sum_{k=1}^{d} n_k^2}{n_1\cdots n_d}$$

for some constant $\hat{\alpha}_{\boldsymbol{p}}$ independent of $\boldsymbol{n}$. Furthermore, by Theorem 4.4 and by the Fan-Hoffman theorem,

$$\sigma_{\min}(A_{\boldsymbol{n}}^{[\boldsymbol{p}]}) \ge \lambda_{\min}(\mathfrak{R}(A_{\boldsymbol{n}}^{[\boldsymbol{p}]})) \ge \frac{\tilde{\alpha}_{\boldsymbol{p}}}{n_1\cdots n_d}$$

for some constant $\tilde{\alpha}_{\boldsymbol{p}} > 0$ independent of $\boldsymbol{n}$. Thus, $\kappa(A_{\boldsymbol{n}}^{[\boldsymbol{p}]}) = \frac{\sigma_{\max}(A_{\boldsymbol{n}}^{[\boldsymbol{p}]})}{\sigma_{\min}(A_{\boldsymbol{n}}^{[\boldsymbol{p}]})} \le \alpha_{\boldsymbol{p}} \sum_{k=1}^{d} n_k^2$, with $\alpha_{\boldsymbol{p}} = \hat{\alpha}_{\boldsymbol{p}}/\tilde{\alpha}_{\boldsymbol{p}}$. □

(4.82) says that $\kappa(A_{\boldsymbol{n}}^{[\boldsymbol{p}]})$ is bounded from above by $\max(\boldsymbol{n}^2) = \max(n_1^2, \ldots, n_d^2)$ multiplied by some constant independent of $\boldsymbol{n}$ (for instance $\alpha_{\boldsymbol{p}} d$). This upper bound is the sharpest possible, as shown by the numerical experiments in Table 4.4, where we fixed $d = 2$, $\boldsymbol{\beta} = \mathbf{0}$, $\gamma = 0$, $\boldsymbol{p} = (2,2)$, and we computed $\kappa(A_{\boldsymbol{n}}^{[\boldsymbol{p}]}) = \kappa(A_{\boldsymbol{n},D}^{[\boldsymbol{p}]})$ (normalized by n^2) for $\boldsymbol{n} = (n, \log_2 n)$ and for increasing values of n. For a nice comparison with Finite Differences and Lagrangian Finite Elements, in the third and fourth column of Table 4.4 we reported the values of $\kappa(A_{\boldsymbol{n}})/n^2$ and $\kappa(A_{\boldsymbol{n}}^{[\boldsymbol{p}]})/n^2$, with $A_{\boldsymbol{n}}$ and $A_{\boldsymbol{n}}^{[\boldsymbol{p}]}$ as in Table 3.2; see also (3.58) for the expression of the FD diffusion matrix $A_{\boldsymbol{n}}$. We note that the smallest asymptotic growth of the condition number when $n \to \infty$ is obtained in correspondence of the Galerkin IgA stiffness matrices $A_{\boldsymbol{n}}^{[\boldsymbol{p}]}$. Note also that the best asymptotic constant 0.0760 is about 0.06 times the worst asymptotic constant 1.2538, associated with the Finite Element stiffness matrices $A_{\boldsymbol{n}}^{(\boldsymbol{p})}$.

4.5.2 Spectral distribution and symbol of the normalized sequence $\{n^{d-2}A_{\boldsymbol{n}}^{[\boldsymbol{p}]}\}_n$

In this subsection we assume that $n_j = \nu_j n$ for all $j = 1, \ldots, d$, i.e. $\boldsymbol{n} = \boldsymbol{\nu} n = (\nu_1 n, \ldots, \nu_d n) \in \mathbb{N}^d$, where $\boldsymbol{\nu} \in \mathbb{Q}_+^d$ is fixed and n varies in the set of natural numbers such that $\boldsymbol{n} \in \mathbb{N}^d$. In Theorem 4.7 we prove that the sequence of matrices $\{n^{d-2}A_{\boldsymbol{n}}^{[\boldsymbol{p}]}\}_n$ is distributed, in the sense of the eigenvalues, like the real function $f_{\boldsymbol{p}}^{(\boldsymbol{\nu})}$ in (4.65), which is therefore the symbol of the sequence $\{n^{d-2}A_{\boldsymbol{n}}^{[\boldsymbol{p}]}\}_n$. Note that $\{n^{d-2}A_{\boldsymbol{n}}^{[\boldsymbol{p}]}\}_n$ is really a sequence of matrices, due to the assumption $\boldsymbol{n} = \boldsymbol{\nu} n$, which must be kept in mind while reading this subsection. Note also that, by the properties of f_p and h_p obtained in Section 4.4, $f_{\boldsymbol{p}}^{(\boldsymbol{\nu})}(\boldsymbol{\theta}) \geq 0$ for all $\boldsymbol{\theta} \in [-\pi, \pi]^d$ and $f_{\boldsymbol{p}}^{(\boldsymbol{\nu})}(\boldsymbol{\theta}) > 0$ for all $\boldsymbol{\theta} \in [-\pi, \pi]^d \setminus \{\boldsymbol{0}\}$.

In order to prove Theorem 4.7, some preliminary work is needed. Let us decompose the matrix $K_n^{[p]}$ into

$$K_n^{[p]} = T_{n+p-2}(f_p) + R_n^{[p]}, \tag{4.83}$$

where $T_{n+p-2}(f_p)$, the $(n+p-2)$-th Toeplitz matrix associated with f_p, is nothing else than the symmetric $(2p+1)$-band matrix whose generic central row is given by (4.48), while $R_n^{[p]} := K_n^{[p]} - T_{n+p-2}(f_p)$ is a low-rank correction term. Indeed, we know from Subsection 4.3.1, see (4.44), that $[K_n^{[p]}]_{i,j=p}^{n-1} = T_{n-p}(f_p) = [T_{n+p-2}(f_p)]_{i,j=p}^{n-1}$, hence

$$\operatorname{rank}(R_n^{[p]}) \leq 4(p-1). \tag{4.84}$$

Similarly, we decompose the matrix $M_n^{[p]}$ into

$$M_n^{[p]} = T_{n+p-2}(h_p) + S_n^{[p]}, \tag{4.85}$$

where $T_{n+p-2}(h_p)$ is just the symmetric $(2p+1)$-band matrix whose generic central row is given by (4.50), while $S_n^{[p]} := M_n^{[p]} - T_{n+p-2}(h_p)$ is a low-rank correction term analogous to $R_n^{[p]}$:

$$\operatorname{rank}(S_n^{[p]}) \leq 4(p-1). \tag{4.86}$$

The next lemma analyzes the spectral properties of $T_{n+p-2}(f_p)$. Besides being interesting in its own right, some of the given properties are needed for the proof of Theorem 4.7, which yields the spectral distribution of the sequence $\{n^{d-2}A_{\boldsymbol{n}}^{[\boldsymbol{p}]}\}_n$.

Lemma 4.8. *Let f_p and M_{f_p} be defined as in Lemma 4.5. Then, the following properties hold.*

1. $\Lambda(T_{n+p-2}(f_p)) \subset (0, M_{f_p})$ *for all* n.
2. $\lambda_{\min}(T_{n+p-2}(f_p)) \searrow 0$ *and* $\lambda_{\max}(T_{n+p-2}(f_p)) \nearrow M_{f_p}$ *as* $n \to \infty$.
3. $\{T_{n+p-2}(f_p)\}_n \sim_\lambda f_p$.
4. *For each fixed* $j \geq 1$,
$$\lambda_{n+p-1-j}(T_{n+p-2}(f_p)) \overset{n\to\infty}{\sim} \frac{j^2\pi^2}{n^2}.$$

Proof. The first three statements are consequences of Theorem 1.8 and Lemma 4.5, except for the monotone convergence in item 2, which, however, follows from the Cauchy interlacing Theorem 1.3 and from the fact that $T_{n+p-2}(f_p)$ is a principal submatrix of $T_{(n+1)+p-2}(f_p)$.

We now prove the last statement. From Lemma 4.5 we know that $\theta = 0$ is the unique zero of f_p over $[-\pi, \pi]$. Furthermore, from the definition of f_p, see (4.46), it follows immediately that $f_p'(0) = 0$. Moreover, by using Lemma 4.3, we get

$$f_p''(0) = 2\sum_{k=1}^{p} k^2 \ddot{\phi}_{[2p+1]}(p+1-k) = 2.$$

This means that the function f_p satisfies all the hypotheses of Theorem 1.10 with $s = 1$, $\theta_{\min} = 0$ and $f_p^{(2s)}(\theta_{\min}) = 2$. Then, for each fixed $j \geq 1$,

$$\lambda_{n+p-1-j}(T_{n+p-2}(f_p)) \overset{n\to\infty}{\sim} \frac{c_{1,j}}{(n+p-2)^2} \overset{n\to\infty}{\sim} \frac{j^2\pi^2}{n^2},$$

where the last asymptotic equivalence holds because $c_{1,j} = j^2\pi^2$; see Remarks 1.4–1.5. □

Remark 4.4. In Subsection 4.5.1, looking at the numerical results summarized in the Tables 4.2–4.3, we conjectured that (4.73) holds for all $p, j \geq 1$. In Lemma 4.8 we have seen that (4.73) holds with $\lambda_{n+p-1-j}(T_{n+p-2}(f_p))$ in place of $\lambda_{n+p-1-j}(K_n^{[p]})$. Furthermore, using the Cauchy interlacing Theorem 1.3 and the fact that $T_{n+p-2}(f_p)$ is a principal submatrix of $K_{n+2p-2}^{[p]}$, we have

$$\lambda_j(K_{n+2p-2}^{[p]}) \geq \lambda_j(T_{n+p-2}(f_p)) \geq \lambda_{j+2p-2}(K_{n+2p-2}^{[p]}), \quad \forall p, n \geq 1,\ \forall j = 1, \ldots, n+p-2.$$

Hence, if, for a fixed $j \geq 1$, there exists a constant $\tilde{k}_{p,j}$ such that

$$\lambda_{n+p-1-j}(K_n^{[p]}) \overset{n\to\infty}{\sim} \frac{\tilde{k}_{p,j}}{n^2},$$

then

$$\lambda_{(n+p-1-j)+2p-2}(K_{n+2p-2}^{[p]}) = \lambda_{n+p-1-(j-2p+2)}(K_{n+2p-2}^{[p]}) = \lambda_{(n+2p-2)+p-1-j}(K_{n+2p-2}^{[p]}) \overset{n\to\infty}{\sim} \frac{\tilde{k}_{p,j}}{n^2},$$

and it follows that:

- $\tilde{k}_{p,j} \leq j^2\pi^2$;
- if $j > 2p-2$, then $\tilde{k}_{p,j} \geq (j-2p+2)^2$.

Theorem 4.7. *Let $\boldsymbol{p} \in \mathbb{N}^d$, $\boldsymbol{\nu} \in \mathbb{Q}_+^d$ and $\boldsymbol{n} = \boldsymbol{\nu}n$, then $\{n^{d-2}A_{\boldsymbol{n}}^{[\boldsymbol{p}]}\}_n \sim_\lambda f_{\boldsymbol{p}}^{(\boldsymbol{\nu})}$, with $f_{\boldsymbol{p}}^{(\boldsymbol{\nu})}$ defined in* (4.65)*. In particular, $\{n^{d-2}A_{\boldsymbol{n}}^{[\boldsymbol{p}]}\}_n$ is weakly clustered at the range $[0, M_{f_{\boldsymbol{p}}^{(\boldsymbol{\nu})}}]$ of $f_{\boldsymbol{p}}^{(\boldsymbol{\nu})}$, where $M_{f_{\boldsymbol{p}}^{(\boldsymbol{\nu})}} := \max_{\boldsymbol{\theta}\in[-\pi,\pi]^d} f_{\boldsymbol{p}}^{(\boldsymbol{\nu})}(\boldsymbol{\theta})$, and every point of $[0, M_{f_{\boldsymbol{p}}^{(\boldsymbol{\nu})}}]$ strongly attracts $\Lambda(n^{d-2}A_{\boldsymbol{n}}^{[\boldsymbol{p}]})$ with infinite order (cf. Theorem 1.5).*

Proof. From (4.7)–(4.10) we have

$$\begin{aligned} n^{d-2}A_{\boldsymbol{n}}^{[\boldsymbol{p}]} &= n^{d-2}A_{\boldsymbol{n},D}^{[\boldsymbol{p}]} + n^{d-2}A_{\boldsymbol{n},A}^{[\boldsymbol{p}]} + n^{d-2}A_{\boldsymbol{n},R}^{[\boldsymbol{p}]} \\ &= T_{\boldsymbol{n}+\boldsymbol{p}-\boldsymbol{2}}(f_{\boldsymbol{p}}^{(\boldsymbol{\nu})}) + n^{d-2}A_{\boldsymbol{n},D}^{[\boldsymbol{p}]} - T_{\boldsymbol{n}+\boldsymbol{p}-\boldsymbol{2}}(f_{\boldsymbol{p}}^{(\boldsymbol{\nu})}) + n^{d-2}A_{\boldsymbol{n},A}^{[\boldsymbol{p}]} + n^{d-2}A_{\boldsymbol{n},R}^{[\boldsymbol{p}]}. \end{aligned} \tag{4.87}$$

We prove that the hypotheses of Theorem 1.6 are satisfied with

$$Z_n = n^{d-2}A_{\boldsymbol{n}}^{[\boldsymbol{p}]}, \quad X_n = T_{\boldsymbol{n}+\boldsymbol{p}-\boldsymbol{2}}(f_{\boldsymbol{p}}^{(\boldsymbol{\nu})}), \quad Y_n = n^{d-2}A_{\boldsymbol{n}}^{[\boldsymbol{p}]} - T_{\boldsymbol{n}+\boldsymbol{p}-\boldsymbol{2}}(f_{\boldsymbol{p}}^{(\boldsymbol{\nu})}) = n^{d-2}A_{\boldsymbol{n},D}^{[\boldsymbol{p}]} - T_{\boldsymbol{n}+\boldsymbol{p}-\boldsymbol{2}}(f_{\boldsymbol{p}}^{(\boldsymbol{\nu})}) + n^{d-2}A_{\boldsymbol{n},A}^{[\boldsymbol{p}]} + n^{d-2}A_{\boldsymbol{n},R}^{[\boldsymbol{p}]}. \tag{4.88}$$

Clearly, by Theorem 1.8 we have $\{T_{\boldsymbol{n}+\boldsymbol{p}-\boldsymbol{2}}(f_{\boldsymbol{p}}^{(\boldsymbol{\nu})})\}_n \sim_\lambda f_{\boldsymbol{p}}^{(\boldsymbol{\nu})}$. Moreover, $T_{\boldsymbol{n}+\boldsymbol{p}-\boldsymbol{2}}(f_{\boldsymbol{p}}^{(\boldsymbol{\nu})})$ is Hermitian (because $f_{\boldsymbol{p}}^{(\boldsymbol{\nu})}$ is real-valued) and (1.37) ensures that

$$\|T_{\boldsymbol{n}+\boldsymbol{p}-\boldsymbol{2}}(f_{\boldsymbol{p}}^{(\boldsymbol{\nu})})\| \leq M_{f_{\boldsymbol{p}}^{(\boldsymbol{\nu})}}, \tag{4.89}$$

where $M_{f_{\boldsymbol{p}}^{(\boldsymbol{\nu})}}$ is defined in the statement of the theorem and is a constant independent of n.

Concerning the spectral norms of $n^{d-2}A_{\boldsymbol{n},D}^{[\boldsymbol{p}]}$, $n^{d-2}A_{\boldsymbol{n},A}^{[\boldsymbol{p}]}$, $n^{d-2}A_{\boldsymbol{n},R}^{[\boldsymbol{p}]}$, we have the following bounds, which were obtained by using (4.38)–(4.40), the equality $\boldsymbol{n} = \boldsymbol{\nu}n$, the property (1.13), the fact that $K_n^{[p]}$, $H_n^{[p]}$, $M_n^{[p]}$ are

normal and Lemma 4.7.

$$\|n^{d-2}A_{\boldsymbol{n},D}^{[\boldsymbol{p}]}\| = \left\| \sum_{k=1}^{d} \frac{n^{d-2}n_k}{n_1\cdots n_{k-1}n_{k+1}\cdots n_d}\, M_{n_1}^{[p_1]}\otimes\cdots\otimes M_{n_{k-1}}^{[p_{k-1}]}\otimes K_{n_k}^{[p_k]}\otimes M_{n_{k+1}}^{[p_{k+1}]}\otimes\cdots\otimes M_{n_d}^{[p_d]} \right\|$$

$$\leq 4\sum_{k=1}^{d} p_k c_k(\boldsymbol{\nu}), \qquad \text{(see (4.66) for the definition of } c_k(\boldsymbol{\nu})) \tag{4.90}$$

$$\|n^{d-2}A_{\boldsymbol{n},A}^{[\boldsymbol{p}]}\| = \left\| \sum_{k=1}^{d} \frac{n^{d-2}}{n_1\cdots n_{k-1}n_{k+1}\cdots n_d}\, M_{n_1}^{[p_1]}\otimes\cdots\otimes M_{n_{k-1}}^{[p_{k-1}]}\otimes \beta_k H_{n_k}^{[p_k]}\otimes M_{n_{k+1}}^{[p_{k+1}]}\otimes\cdots\otimes M_{n_d}^{[p_d]} \right\|$$

$$\leq \frac{2}{n}\sum_{k=1}^{d} \frac{1}{\nu_1\cdots\nu_{k-1}\nu_{k+1}\cdots\nu_d}|\beta_k| = \frac{2\sum_{k=1}^{d}\nu_k|\beta_k|}{\nu_1\cdots\nu_d n}, \tag{4.91}$$

$$\|n^{d-2}A_{\boldsymbol{n},R}^{[\boldsymbol{p}]}\| = \left\| \frac{n^{d-2}}{n_1\cdots n_d} M_{n_1}^{[p_1]}\otimes\cdots\otimes M_{n_d}^{[p_d]} \right\| \leq \frac{1}{\nu_1\cdots\nu_d n^2}. \tag{4.92}$$

From (4.90)–(4.92), it follows that $\|n^{d-2}A_{\boldsymbol{n}}^{[\boldsymbol{p}]}\| \leq C$ for some constant C independent of n.

To complete the proof, it only remains to show that $\left|\!\left|\!\left| n^{d-2}A_{\boldsymbol{n}}^{[\boldsymbol{p}]} - T_{\boldsymbol{n}+\boldsymbol{p}-\boldsymbol{2}}(f_{\boldsymbol{p}}^{(\boldsymbol{\nu})}) \right|\!\right|\!\right| = o(N(\boldsymbol{n}+\boldsymbol{p}-\boldsymbol{2})) = o(n^d)$, where we recall that $N(\boldsymbol{n}+\boldsymbol{p}-\boldsymbol{2}) = \prod_{i=1}^{d}(n_i+p_i-2)$ is the size of $A_{\boldsymbol{n}}^{[\boldsymbol{p}]}$ and $\boldsymbol{n} = \boldsymbol{\nu}n$. We first note that, by definition of $f_{\boldsymbol{p}}^{(\boldsymbol{\nu})}$, see (4.65), by the linearity of $T_{\boldsymbol{n}+\boldsymbol{p}-\boldsymbol{2}}(\cdot)$ and by Lemma 1.8, we have

$$T_{\boldsymbol{n}+\boldsymbol{p}-\boldsymbol{2}}(f_{\boldsymbol{p}}^{(\boldsymbol{\nu})}) = T_{\boldsymbol{n}+\boldsymbol{p}-\boldsymbol{2}}\left(\sum_{k=1}^{d} c_k(\boldsymbol{\nu})\, h_{p_1}\otimes\cdots\otimes h_{p_{k-1}}\otimes f_{p_k}\otimes h_{p_{k+1}}\otimes\cdots\otimes h_{p_d} \right)$$

$$= \sum_{k=1}^{d} c_k(\boldsymbol{\nu})\, T_{n_1+p_1-2}(h_{p_1})\otimes\cdots\otimes T_{n_{k-1}+p_{k-1}-2}(h_{p_{k-1}})\otimes T_{n_k+p_k-2}(f_{p_k})\otimes T_{n_{k+1}+p_{k+1}-2}(h_{p_{k+1}})\otimes\cdots\otimes T_{n_d+p_d-2}(h_{p_d}).$$

Moreover,

$$n^{d-2}A_{\boldsymbol{n},D}^{[\boldsymbol{p}]} = \sum_{k=1}^{d} c_k(\boldsymbol{\nu})\, M_{n_1}^{[p_1]}\otimes\cdots\otimes M_{n_{k-1}}^{[p_{k-1}]}\otimes K_{n_k}^{[p_k]}\otimes M_{n_{k+1}}^{[p_{k+1}]}\otimes\cdots\otimes M_{n_d}^{[p_d]},$$

and we recall from (4.84), (4.86) that

$$\operatorname{rank}(K_n^{[p]} - T_{n+p-2}(f_p)) \leq 4(p-1), \qquad \operatorname{rank}(M_n^{[p]} - T_{n+p-2}(h_p)) \leq 4(p-1).$$

Therefore, by (1.5) and by the property (1.18) of tensor products we obtain

$$\left|\!\left|\!\left| n^{d-2}A_{\boldsymbol{n}}^{[\boldsymbol{p}]} - T_{\boldsymbol{n}+\boldsymbol{p}-\boldsymbol{2}}(f_{\boldsymbol{p}}^{(\boldsymbol{\nu})}) \right|\!\right|\!\right|_1 \leq \left|\!\left|\!\left| n^{d-2}A_{\boldsymbol{n},D}^{[\boldsymbol{p}]} - T_{\boldsymbol{n}+\boldsymbol{p}-\boldsymbol{2}}(f_{\boldsymbol{p}}^{(\boldsymbol{\nu})}) \right|\!\right|\!\right|_1 + \left|\!\left|\!\left| n^{d-2}A_{\boldsymbol{n},A}^{[\boldsymbol{p}]} \right|\!\right|\!\right|_1 + \left|\!\left|\!\left| A_{\boldsymbol{n},R}^{[\boldsymbol{p}]} \right|\!\right|\!\right|_1$$

$$\leq d\,N(\boldsymbol{n}+\boldsymbol{p}-\boldsymbol{2}) \sum_{i=1}^{d} \frac{4(p_i-1)}{n_i+p_i-2} \left\| n^{d-2}A_{\boldsymbol{n},D}^{[\boldsymbol{p}]} - T_{\boldsymbol{n}+\boldsymbol{p}-\boldsymbol{2}}(f_{\boldsymbol{p}}^{(\boldsymbol{\nu})}) \right\| + N(\boldsymbol{n}+\boldsymbol{p}-\boldsymbol{2}) \left\| n^{d-2}A_{\boldsymbol{n},A}^{[\boldsymbol{p}]} \right\| + N(\boldsymbol{n}+\boldsymbol{p}-\boldsymbol{2}) \left\| n^{d-2}A_{\boldsymbol{n},R}^{[\boldsymbol{p}]} \right\| \tag{4.93}$$

and the latter is $o(n^d)$ by (4.89)–(4.92). □

Remark 4.5. In Theorem 4.7, assume that $d = 1$ and take $\boldsymbol{p} = p$, $\boldsymbol{\nu} = 1$. Then $\boldsymbol{n} = n$, $f_{\boldsymbol{p}}^{(\boldsymbol{\nu})}(\boldsymbol{\theta}) = f_p(\theta)$ is just the function analyzed in Section 4.4, and Theorem 4.7 gives $\{\frac{1}{n}A_n^{[p]}\}_n \sim_\lambda f_p$, where

$$A_n^{[p]} = nK_n^{[p]} + \beta H_n^{[p]} + \frac{\gamma}{n}M_n^{[p]}, \qquad n = 1, 2, \ldots$$

is the sequence of 1D Galerkin B-spline IgA stiffness matrices.[3] Moreover, (4.93) together with (4.89)–(4.92) shows that

$$\left\|\left|\frac{1}{n}A_n^{[p]} - T_{n+p-2}(f_p)\right|\right\|_1 \le C$$

for some constant C independent of n. Then, all the hypotheses of Theorem 1.7 are satisfied with Z_n, X_n, Y_n as in (4.88), and so $\{\frac{1}{n}A_n^{[p]}\}_n$ is strongly clustered at $[0, M_{f_p}]$.

In the next subsections, we will consider some cases that will be analyzed in more detail. We first focus on the spectral properties of the matrices $\frac{1}{n}A_n^{[p]}$ associated with the linear ($p = 1$) and quadratic ($p = 2$) B-spline IgA approximation of problem (4.1) in 1D. Then, we will address the 2D discretization matrices $A_{n_1,n_2}^{[p_1,p_2]}$ associated with the bilinear ($p_1 = p_2 = 1$) and biquadratic ($p_1 = p_2 = 2$) B-spline IgA approximation of (4.1).

4.5.3 The linear case $p = 1$

In the case $p = 1$, the matrix $A_n^{[1]}$ is of size $(n-1)\times(n-1)$ and is given by

$$A_n^{[1]} = nK_n^{[1]} + \beta H_n^{[1]} + \frac{\gamma}{n}M_n^{[1]}, \tag{4.94}$$

where, for $n \ge 4$,

$$K_n^{[1]} = \begin{bmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 2 \end{bmatrix}, \quad H_n^{[1]} = \frac{1}{2}\begin{bmatrix} 0 & 1 & & & \\ -1 & 0 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 0 & 1 \\ & & & -1 & 0 \end{bmatrix}, \quad M_n^{[1]} = \frac{1}{6}\begin{bmatrix} 4 & 1 & & & \\ 1 & 4 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & 4 & 1 \\ & & & 1 & 4 \end{bmatrix}.$$

The matrix $A_n^{[1]}$ is nothing else than the stiffness matrix arising from classical FEM with linear elements. In other words, using the notation of Chapter 3, we have $A_n^{[1]} = A_n^{(1)}$. Observe that the scaled matrix

$$\frac{1}{n}A_n^{[1]} = K_n^{[1]} + \frac{\beta}{n}H_n^{[1]} + \frac{\gamma}{n^2}M_n^{[1]} \tag{4.95}$$

is a real Toeplitz tridiagonal matrix, which is given explicitly by

$$\frac{1}{n}A_n^{[1]} = \text{Tridiagonal}\left(-1 - \frac{\beta}{2n} + \frac{\gamma}{6n^2},\ 2 + \frac{2\gamma}{3n^2},\ -1 + \frac{\beta}{2n} + \frac{\gamma}{6n^2}\right).$$

Moreover, for n large enough, the elements $-1 - \frac{\beta}{2n} + \frac{\gamma}{6n^2}$ and $-1 + \frac{\beta}{2n} + \frac{\gamma}{6n^2}$ are both negative. This means that, for n large enough, all the eigenvalues of $\frac{1}{n}A_n^{[p]}$ are real and can be computed by means of Theorem 1.9.

Proposition 4.1. *Let $n \ge 4$ be such that $-1 - \frac{\beta}{2n} + \frac{\gamma}{6n^2}$ and $-1 + \frac{\beta}{2n} + \frac{\gamma}{6n^2}$ are both negative. Then, $\frac{1}{n}A_n^{[1]}$ has $n-1$ real distinct eigenvalues*

$$\lambda_j\left(\frac{1}{n}A_n^{[1]}\right) = 2 + \frac{2\gamma}{3n^2} + 2\sqrt{1 - \left(\frac{\gamma}{3} + \frac{\beta^2}{4}\right)\frac{1}{n^2} + \frac{\gamma^2}{36n^4}}\,\cos\frac{j\pi}{n}, \quad j = 1, \dots, n-1. \tag{4.96}$$

[3] $A_n^{[p]}$ is given by (4.6)–(4.10) and Theorem 4.2 for $d = 1$, $\boldsymbol{n} = n$, $\boldsymbol{p} = p$, $\boldsymbol{\beta} = \beta$.

By using the expression (4.96) for the eigenvalues, it can be proved by direct computation (without even invoking Theorem 4.7 or Remark 4.5) that the sequence $\{\frac{1}{n}A_n^{[1]}\}$ is distributed like the function $f_1(\theta) = 2-2\cos\theta$ in the sense of the eigenvalues. In addition, by using (4.96) and some asymptotic expansion, one can prove that

$$\lambda_{\min}\left(\frac{1}{n}A_n^{[1]}\right) \geq 4\left(\sin\frac{\pi}{2n}\right)^2 + \frac{2\gamma}{3n^2}.$$

Furthermore, by Gershgorin's first theorem [8], we have $\lambda_{\min}(\frac{1}{n}A_n^{[1]}) \geq \frac{\gamma}{n^2}$. Hence,

$$\Lambda\left(\frac{1}{n}A_n^{[1]}\right) \subset \left[\max\left(4\left(\sin\frac{\pi}{2n}\right)^2 + \frac{2\gamma}{3n^2}, \frac{\gamma}{n^2}\right), 4 + \frac{\gamma}{3n^2}\right).$$

This gives a sharper lower bound for $\lambda_{\min}(\frac{1}{n}A_n^{[p]})$ than the one provided in Theorem 4.5, if we take into account that $C_1 = \frac{1}{3}$ is the best constant satisfying (4.67) for $p = 1$. The latter is true because:

- $\lambda_{\min}(M_n^{[1]}) = \lambda_{\min}(T_{n-1}(\frac{2}{3} + \frac{1}{3}\cos\theta)) \searrow \frac{1}{3} = \min_{\theta\in[-\pi,\pi]}(\frac{2}{3} + \frac{1}{3}\cos\theta)$ when $n \to \infty$ (this follows from Theorems 1.8 and 1.3);
- if C_1 is a constant satisfying the left-hand side inequality in (4.67), then $\lambda_{\min}(M_n^{[1]}) \geq C_1$ for all n (see Theorem 4.3);
- if C_1 is a constant satisfying $\lambda_{\min}(M_n^{[1]}) \geq C_1$ for all n, then it also satisfies the left-hand side inequality in (4.67) (see Remark 4.2).

From (4.96) it also follows that

$$n^2\lambda_{\min}\left(\frac{1}{n}A_n^{[1]}\right) = n^2\lambda_{n-1}\left(\frac{1}{n}A_n^{[1]}\right) \overset{n\to\infty}{\longrightarrow} \pi^2 + \gamma + \frac{\beta^2}{4},$$
$$n^2\left(4 - \lambda_{\max}\left(\frac{1}{n}A_n^{[1]}\right)\right) = n^2\left(4 - \lambda_1\left(\frac{1}{n}A_n^{[1]}\right)\right) \overset{n\to\infty}{\longrightarrow} \pi^2 - \frac{\gamma}{3} + \frac{\beta^2}{4}.$$

In particular, $\{\frac{1}{n}A_n^{[1]}\}$ is strongly clustered at $[0,4]$ according to Definition 1.3. Note that $[0,4]$ is precisely the range of the function $f_1(\theta) = 2 - 2\cos\theta$ (cf. Remark 4.5).

We conclude this subsection by collecting in the next lemma some results which can be derived by the Gershgorin theorems and will be used later on.

Lemma 4.9. *For all $n \geq 4$,*

- *$H_n^{[1]}$ is skew-symmetric, irreducible, and $\Lambda(H_n^{[1]}) \subset \{0\} \times (-1,1)$;*
- *$M_n^{[1]}$ is symmetric, irreducible, and $\Lambda(M_n^{[1]}) \subset \left(\frac{1}{3}, 1\right)$.*

4.5.4 The quadratic case $p = 2$

The spectral analysis of $\frac{1}{n}A_n^{[1]}$ has not been difficult because Theorem 1.9 provided us with the explicit expression (4.96) for the eigenvalues of $\frac{1}{n}A_n^{[1]}$. For $p \geq 2$ such an expression for the eigenvalues of $\frac{1}{n}A_n^{[p]}$ is not available and so our spectral analysis must rely on other considerations. In the case $p = 2$, the matrix $\frac{1}{n}A_n^{[2]}$ is of size $n \times n$ and is given by

$$\frac{1}{n}A_n^{[2]} = K_n^{[2]} + \frac{\beta}{n}H_n^{[2]} + \frac{\gamma}{n^2}M_n^{[2]},$$

where, for $n \geq 5$,

$$K_n^{[2]} = \frac{1}{6}\begin{bmatrix} 8 & -1 & -1 & & & & \\ -1 & 6 & -2 & -1 & & & \\ -1 & -2 & 6 & -2 & -1 & & \\ & \ddots & \ddots & \ddots & \ddots & \ddots & \\ & & -1 & -2 & 6 & -2 & -1 \\ & & & -1 & -2 & 6 & -1 \\ & & & & -1 & -1 & 8 \end{bmatrix}, \qquad H_n^{[2]} = \frac{1}{24}\begin{bmatrix} 0 & 9 & 1 & & & & \\ -9 & 0 & 10 & 1 & & & \\ -1 & -10 & 0 & 10 & 1 & & \\ & \ddots & \ddots & \ddots & \ddots & \ddots & \\ & & -1 & -10 & 0 & 10 & 1 \\ & & & -1 & -10 & 0 & 9 \\ & & & & -1 & -9 & 0 \end{bmatrix},$$

$$M_n^{[2]} = \frac{1}{120}\begin{bmatrix} 40 & 25 & 1 & & & & & \\ 25 & 66 & 26 & 1 & & & & \\ 1 & 26 & 66 & 26 & 1 & & & \\ & \ddots & \ddots & \ddots & \ddots & \ddots & & \\ & & 1 & 26 & 66 & 26 & 1 & \\ & & & 1 & 26 & 66 & 25 & \\ & & & & 1 & 25 & 40 & \end{bmatrix}.$$

For the case $p = 2$, Theorem 4.7 (or Remark 4.5) gives $\{\frac{1}{n}A_n^{[2]}\} \sim_\lambda f_2$, with

$$f_2(\theta) = 1 - \frac{2}{3}\cos\theta - \frac{1}{3}\cos(2\theta).$$

Moreover, $\{\frac{1}{n}A_n^{[2]}\}$ is strongly clustered at $\left[0, \frac{3}{2}\right]$, which is the range of f_2 (see Remark 4.5). In the remainder of this subsection we provide more specific results about the spectral properties of $\frac{1}{n}A_n^{[2]}$.

Localization of the eigenvalues

We look for a more precise localization of $\Lambda(\frac{1}{n}A_n^{[2]})$ than the one provided by Theorem 4.5. We first note that, by using the Gershgorin theorems, we can derive the following bounds for the spectra of the matrices $K_n^{[2]}$, $H_n^{[2]}$ and $M_n^{[2]}$.

Lemma 4.10. *For all $n \geq 5$,*

- *$K_n^{[2]}$ is symmetric, irreducible, and $\Lambda(K_n^{[2]}) \subset (0, 2)$;*
- *$H_n^{[2]}$ is skew-symmetric, irreducible, and $\Lambda(H_n^{[2]}) \subset \{0\} \times \left(-\frac{11}{12}, \frac{11}{12}\right)$;*
- *$M_n^{[2]}$ is symmetric, irreducible, and $\Lambda(M_n^{[2]}) \subset \left(\frac{1}{10}, 1\right)$;*
- *if $\frac{25\gamma}{120n^2} < \frac{1}{6}$, then $K_n^{[2]} + \frac{\gamma}{n^2}M_n^{[2]}$ is symmetric, irreducible, and*

$$\Lambda\left(K_n^{[2]} + \frac{\gamma}{n^2}M_n^{[2]}\right) \subset \left(\frac{\gamma}{n^2}, 2 + \frac{\gamma}{10n^2}\right).$$

Remark 4.6. Lemma 4.10 implies that $\lambda_{\min}(M_n^{[2]}) > \frac{1}{10}$ for all $n \geq 5$. Using the arguments shown in Remark 4.2 and Theorem 4.3, it follows that

$$\lambda_{\min}(K_n^{[2]}) > \frac{\pi^2}{10n^2}, \quad \forall n \geq 5, \tag{4.97}$$

$$\left\| \sum_{i=1}^{n} x_i N_{i+1,[2]} \right\|_{L^2(0,1)}^2 > \frac{\|\mathbf{x}\|^2}{10n}, \quad \forall n \geq 5, \quad \forall \mathbf{x} \in \mathbb{R}^n.$$

In particular, the left-hand side inequality in (4.67) holds for $p = 2$ with $C_2 = \frac{1}{10}$.

Using the constant $C_2 = \frac{1}{10}$, the localization of $\Lambda(\frac{1}{n}A_n^{[2]})$ provided by Theorem 4.5 gives

$$\Lambda\left(\frac{1}{n}A_n^{[2]}\right) \subseteq \left[\frac{\pi^2+\gamma}{n^2}\,\frac{1}{10}, 8+\frac{\gamma}{n^2}\right] \times \left[-\frac{2|\beta|}{n}, \frac{2|\beta|}{n}\right]. \tag{4.98}$$

We will prove in Theorem 4.8 a better localization of $\Lambda(\frac{1}{n}A_n^{[2]})$ than (4.98).

Lemma 4.11. *Let $R_n^{[2]}$ be the low-rank matrix $R_n^{[2]}$ introduced in* (4.83)*. Then, for every $n \geq 5$,*

$$R_n^{[2]} = \frac{1}{6}\begin{bmatrix} 2 & 1 & & & \\ 1 & 0 & & & \\ & & & & \\ & & & 0 & 1 \\ & & & 1 & 2 \end{bmatrix} \in \mathbb{R}^{n\times n},$$

and the characteristic polynomial of $R_n^{[2]}$ is given by $\frac{1}{1296}\lambda^{n-4}(36\lambda^2 - 12\lambda - 1)^2$. Hence, the eigenvalues of $R_n^{[2]}$ are $\frac{1+\sqrt{2}}{6}$ (with multiplicity 2), $\frac{1-\sqrt{2}}{6}$ (with multiplicity 2) and 0 (with multiplicity $n-4$).

Theorem 4.8. *For every $n \geq 5$ such that $\frac{25\gamma}{120n^2} < \frac{1}{6}$,*

$$\Lambda\left(\frac{1}{n}A_n^{[2]}\right) \subset \left(\max\left(\frac{\gamma}{n^2}, \frac{\pi^2+\gamma}{10n^2}\right), \min\left(\frac{3}{2}+\frac{1+\sqrt{2}}{6}+\frac{\gamma}{n^2}, 2+\frac{\gamma}{10n^2}\right)\right) \times \left[-\frac{11|\beta|}{12n}, \frac{11|\beta|}{12n}\right]. \tag{4.99}$$

Proof. Fix $n \geq 5$ such that the condition $\frac{25\gamma}{120n^2} < \frac{1}{6}$ is met. The real and imaginary part of $\frac{1}{n}A_n^{[2]}$ are given by

$$\Re\left(\frac{1}{n}A_n^{[2]}\right) = K_n^{[2]} + \frac{\gamma}{n^2}M_n^{[2]} = T_{n+p-2}(f_p) + R_n^{[2]} + \frac{\gamma}{n^2}M_n^{[2]}, \qquad \Im\left(\frac{1}{n}A_n^{[2]}\right) = \frac{\beta}{\mathrm{i}n}H_n^{[2]}.$$

We aim at localizing the spectra $\Lambda(\Re(\frac{1}{n}A_n^{[2]}))$ and $\Lambda(\Im(\frac{1}{n}A_n^{[2]}))$. We begin with $\Lambda(\Re(\frac{1}{n}A_n^{[2]}))$. Since n satisfies the condition $\frac{25\gamma}{120n^2} < \frac{1}{6}$, by Lemma 4.10 we have

$$\Lambda\left(\Re\left(\frac{1}{n}A_n^{[2]}\right)\right) \subset \left(\frac{\gamma}{n^2}, 2+\frac{\gamma}{10n^2}\right). \tag{4.100}$$

We can improve (4.100) as follows. By combining (1.9) with Lemmas 4.8, 4.10 and 4.11, and taking into account that $M_{f_2} = \frac{3}{2}$, we obtain

$$\begin{aligned}\lambda_{\max}\left(\Re\left(\frac{1}{n}A_n^{[2]}\right)\right) &= \lambda_{\max}\left(T_{n+p-2}(f_p) + R_n^{[2]} + \frac{\gamma}{n^2}M_n^{[2]}\right) \leq \lambda_{\max}(T_{n+p-2}(f_p)) + \lambda_{\max}(R_n^{[2]}) + \frac{\gamma}{n^2}\lambda_{\max}(M_n^{[2]}) \\ &< \frac{3}{2}+\frac{1+\sqrt{2}}{6}+\frac{\gamma}{n^2}.\end{aligned}$$

Similarly, by using (4.97) and Lemma 4.10,

$$\lambda_{\min}\left(\Re\left(\frac{1}{n}A_n^{[2]}\right)\right) = \lambda_{\min}\left(K_n^{[2]} + \frac{\gamma}{n^2}M_n^{[2]}\right) \geq \lambda_{\min}(K_n^{[2]}) + \frac{\gamma}{n^2}\lambda_{\min}(M_n^{[2]}) > \frac{\pi^2+\gamma}{10n^2}.$$

Thus, we can replace (4.100) with

$$\Lambda\left(\Re\left(\frac{1}{n}A_n^{[2]}\right)\right) \subset \left(\max\left(\frac{\gamma}{n^2}, \frac{\pi^2+\gamma}{10n^2}\right), \min\left(\frac{3}{2}+\frac{1+\sqrt{2}}{6}+\frac{\gamma}{n^2}, 2+\frac{\gamma}{10n^2}\right)\right). \tag{4.101}$$

Now we localize the spectrum $\Lambda(\Im(\frac{1}{n}A_n^{[2]}))$. Since $\Im(\frac{1}{n}A_n^{[2]})$ is Hermitian, from Lemma 4.10 we obtain[4]

$$\Lambda\left(\Im\left(\frac{1}{n}A_n^{[2]}\right)\right) \subset \left[-\frac{11|\beta|}{12n}, \frac{11|\beta|}{12n}\right]. \tag{4.102}$$

Combining (4.101)–(4.102) with (1.7), we obtain (4.99). □

Clustering

We now deal with the clustering properties of the sequence $\{\frac{1}{n}A_n^{[2]}\}$. We have already mentioned that $\{\frac{1}{n}A_n^{[2]}\}$ is strongly clustered at $\left[0, \frac{3}{2}\right]$, but we have no bounds on the number of outliers, i.e., those eigenvalues of $\frac{1}{n}A_n^{[2]}$ lying outside the rectangular ε-expansion $\left[0, \frac{3}{2}\right]_\varepsilon = \left[-\varepsilon, \frac{3}{2} + \varepsilon\right] \times [-\varepsilon, \varepsilon]$. Theorem 4.9 allows us to provide an estimate for the number of outliers.

Theorem 4.9. *For all $\varepsilon \in (0,1)$ and $n \geq \max(5, \frac{\sqrt{2\gamma}}{\varepsilon})$, it holds that*

$$q_n^+(\varepsilon) \leq \frac{1+\sqrt{2}}{3\varepsilon}, \tag{4.103}$$

where $q_n^+(\varepsilon)$ is the number of eigenvalues of $\frac{1}{n}A_n^{[2]}$ whose real parts are $\geq \frac{3}{2} + \varepsilon$.

Proof. For every $n \geq 5$, we consider again the decomposition $K_n^{[2]} = T_n(f_2) + R_n^{[2]}$ introduced in (4.83). The matrix $R_n^{[2]}$ is symmetric and we know the eigenvalues of $R_n^{[2]}$ from Lemma 4.11. In particular, $R_n^{[2]}$ has two positive and two negative eigenvalues, and so, by Theorem 1.4,

$$\lambda_{j-2}(T_n(f_2)) \geq \lambda_j(K_n^{[2]}) \geq \lambda_{j+2}(T_n(f_2)), \qquad j = 3, \ldots, n-2.$$

From Lemma 4.8 and $M_{f_2} = \frac{3}{2}$, we have $\Lambda(T_n(f_2)) \subset \left(0, \frac{3}{2}\right)$, hence

$$\frac{3}{2} > \lambda_1(T_n(f_2)) \geq \lambda_3(K_n^{[2]}) \geq \ldots \geq \lambda_n(K_n^{[2]}) > 0, \tag{4.104}$$

where the last inequality is a consequence of Lemma 4.10 (or, simply, of the positive definiteness of $K_n^{[2]}$; see Theorem 4.2). Moreover, by (1.9),

$$\lambda_{\max}(K_n^{[2]}) = \lambda_{\max}(T_n(f_2) + R_n^{[2]}) \leq \lambda_{\max}(T_n(f_2)) + \lambda_{\max}(R_n^{[2]}) < \frac{3}{2} + \frac{1+\sqrt{2}}{6}. \tag{4.105}$$

Finally, recalling from that $\lambda_{\max}(M_n^{[2]}) \leq 1$ and applying the minimax principle, for every $j = 1, \ldots, n$ we have

$$\begin{aligned}\lambda_j(\Re(\tfrac{1}{n}A_n^{[2]})) &= \min_{\substack{V \subseteq \mathbb{C}^n \\ \dim V = n+1-j}} \max_{\substack{\mathbf{x} \in V \\ \|\mathbf{x}\|=1}} \left(\mathbf{x}^* \,\Re(\tfrac{1}{n}A_n^{[2]})\,\mathbf{x}\right) = \min_{\substack{V \subseteq \mathbb{C}^n \\ \dim V = n+1-j}} \max_{\substack{\mathbf{x} \in V \\ \|\mathbf{x}\|=1}} \left(\mathbf{x}^*(K_n^{[2]} + \frac{\gamma}{n^2}M_n^{[2]})\mathbf{x}\right) \\ &< \min_{\substack{V \subseteq \mathbb{C}^n \\ \dim V = n+1-j}} \max_{\substack{\mathbf{x} \in V \\ \|\mathbf{x}\|=1}} \left(\mathbf{x}^* K_n^{[2]}\mathbf{x} + \frac{\gamma}{n^2}\right) = \lambda_j(K_n^{[2]}) + \frac{\gamma}{n^2}, \qquad j = 1, \ldots, n.\end{aligned} \tag{4.106}$$

Now fix $\varepsilon > 0$ and let $q_n^+(\varepsilon)$ be the number of eigenvalues of $\frac{1}{n}A_n^{[2]}$ whose real parts are $\geq \frac{3}{2} + \varepsilon$. Label the eigenvalues of $\frac{1}{n}A_n^{[2]}$ and $\Re(\frac{1}{n}A_n^{[2]})$ in the following way:

$$\Re\left(\lambda_1\left(\tfrac{1}{n}A_n^{[2]}\right)\right) \geq \ldots \geq \Re\left(\lambda_n\left(\tfrac{1}{n}A_n^{[2]}\right)\right),$$

[4] If $\beta \neq 0$ then $\Im(\frac{1}{n}A_n^{[2]})$ is irreducible and $\Lambda(\Im(\frac{1}{n}A_n^{[2]})) \subset (-\frac{11|\beta|}{12n}, \frac{11|\beta|}{12n})$. In (4.102) we have included the endpoints $\pm\frac{11|\beta|}{12n}$ to cover the case $\beta = 0$.

and

$$\lambda_1\left(\Re\left(\tfrac{1}{n}A_n^{[2]}\right)\right) \geq \ldots \geq \lambda_n\left(\Re\left(\tfrac{1}{n}A_n^{[2]}\right)\right).$$

Following the argument used in [37, proof of Theorem 3.5] and keeping in mind (4.104)–(4.106), we apply the Ky-Fan Theorem 1.2 to obtain

$$\begin{aligned}\left(\frac{3}{2}+\varepsilon\right)q_n^+(\varepsilon) &\leq \sum_{j=1}^{q_n^+(\varepsilon)} \Re\left(\lambda_j\left(\frac{1}{n}A_n^{[2]}\right)\right) \leq \sum_{j=1}^{q_n^+(\varepsilon)} \lambda_j\left(\Re\left(\frac{1}{n}A_n^{[2]}\right)\right) \leq \sum_{j=1}^{q_n^+(\varepsilon)} \left(\lambda_j(K_n^{[2]})+\frac{\gamma}{n^2}\right)\\ &= \sum_{j=1}^{q_n^+(\varepsilon)} \lambda_j(K_n^{[2]}) + \frac{\gamma\, q_n^+(\varepsilon)}{n^2} = \lambda_1(K_n^{[2]}) + \lambda_2(K_n^{[2]}) + \sum_{j=3}^{q_n^+(\varepsilon)} \lambda_j(K_n^{[2]}) + \frac{\gamma\, q_n^+(\varepsilon)}{n^2}\\ &< 2\left(\frac{3}{2}+\frac{1+\sqrt{2}}{6}\right) + (q_n^+(\varepsilon)-2)\,\frac{3}{2} + \frac{\gamma\, q_n^+(\varepsilon)}{n^2} = \left(\frac{3}{2}+\frac{\gamma}{n^2}\right)q_n^+(\varepsilon) + \frac{1+\sqrt{2}}{3},\end{aligned}$$

and so, for every $\varepsilon > 0$ and $n \geq 5$ such that $\frac{\gamma}{n^2} < \varepsilon$, we have

$$q_n^+(\varepsilon) < \frac{1+\sqrt{2}}{3\left(\varepsilon - \frac{\gamma}{n^2}\right)}. \tag{4.107}$$

If $0 < \varepsilon < 1$ and $n > \max\left(5, \sqrt{\frac{\gamma}{\varepsilon}}\right)$, then

$$\frac{1+\sqrt{2}}{3\left(\varepsilon - \frac{\gamma}{n^2}\right)} \leq \frac{1+\sqrt{2}}{3\varepsilon} + 1 \quad \Leftrightarrow \quad n \geq \sqrt{\frac{(1+\sqrt{2}+3\varepsilon)\gamma}{3\varepsilon^2}}.$$

From the inequality

$$\sqrt{\frac{(1+\sqrt{2}+3\varepsilon)\gamma}{3\varepsilon^2}} \leq \frac{\sqrt{2\gamma}}{\varepsilon},$$

and from (4.107) it follows that (4.103) holds $\forall \varepsilon \in (0,1)$ and $\forall n \geq \max(5, \frac{\sqrt{2\gamma}}{\varepsilon})$. □

Let $q_n(\varepsilon)$ be the number of eigenvalues of $\frac{1}{n}A_n^{[2]}$ lying outside the rectangular ε-expansion $\left[0, \frac{3}{2}\right]_\varepsilon$. By combining (4.99) and (4.103), we are able to find an upper bound for $q_n(\varepsilon)$. Indeed, $\forall \varepsilon \in (0,1)$ and $\forall n > \max\left(5, \frac{11|\beta|}{12\varepsilon}, \frac{\sqrt{2\gamma}}{\varepsilon}\right) = O\left(\frac{1}{\varepsilon}\right)$,

$$q_n(\varepsilon) \leq \frac{1+\sqrt{2}}{3\varepsilon}.$$

Note that, by Theorem 4.8, $\forall \varepsilon \in (0,1)$ and $\forall n \geq \max\left(5, \frac{11|\beta|}{12\varepsilon}, \sqrt{\frac{5\gamma}{4\varepsilon}}\right)$, there are no eigenvalues of $\frac{1}{n}A_n^{[2]}$ lying outside the ε-expansion $\left[0, \frac{3}{2}+\frac{1+\sqrt{2}}{6}\right]_\varepsilon$. Thus, $\forall \varepsilon \in (0,1)$ and $\forall n \geq \max\left(5, \frac{11|\beta|}{12\varepsilon}, \frac{\sqrt{2\gamma}}{\varepsilon}\right)$, $q_n(\varepsilon)$ is just the number of eigenvalues of $\frac{1}{n}A_n^{[2]}$ lying in

$$\left[0, \tfrac{3}{2}+\tfrac{1+\sqrt{2}}{6}\right]_\varepsilon \setminus \left[0, \tfrac{3}{2}\right]_\varepsilon = \left(\tfrac{3}{2}+\varepsilon, \tfrac{3}{2}+\tfrac{1+\sqrt{2}}{6}+\varepsilon\right] \times [-\varepsilon, \varepsilon].$$

4.5.5 The bilinear case $p_1 = p_2 = 1$

In the case $p_1 = p_2 = 1$, the matrix $A^{[1,1]}_{n,n}$ is $(n-1)^2 \times (n-1)^2$ and is given by

$$A^{[1,1]}_{n,n} = A^{[1,1]}_{n,n,D} + \frac{\beta_1}{n} H^{[1]}_n \otimes M^{[1]}_n + \frac{\beta_2}{n} M^{[1]}_n \otimes H^{[1]}_n + \frac{\gamma}{n^2} M^{[1]}_n \otimes M^{[1]}_n, \tag{4.108}$$

where

$$A^{[1,1]}_{n,n,D} = K^{[1]}_n \otimes M^{[1]}_n + M^{[1]}_n \otimes K^{[1]}_n.$$

In this case, Theorem 4.7 gives $\{A^{[1,1]}_{n,n}\} \sim_\lambda f^{(1,1)}_{1,1} =: f_{1,1}$, with

$$f_{1,1}(\theta_1, \theta_2) = (f_1 \otimes h_1)(\theta_1, \theta_2) + (h_1 \otimes f_1)(\theta_1, \theta_2) = \frac{8}{3} - \frac{2}{3}\cos(\theta_1) - \frac{2}{3}\cos(\theta_2) - \frac{4}{3}\cos(\theta_1)\cos(\theta_2).$$

Localization of the eigenvalues and clustering

As in the previous subsection, we look for a precise localization of the spectrum $\Lambda(A^{[1,1]}_{n,n})$ as well as for the clustering properties of the matrix-sequence $\{A^{[1,1]}_{n,n}\}$.

Theorem 4.10. *For every $n \geq 4$ such that $\frac{\gamma}{9n^2} < \frac{1}{3}$,*

$$\Lambda(A^{[1,1]}_{n,n}) \subset \left(\max\left(\frac{\gamma}{n^2}, \frac{8}{3}\left(\sin\frac{\pi}{2n}\right)^2 + \frac{\gamma}{9n^2}\right), \min\left(4 + \frac{\gamma}{n^2}, \frac{16}{3} - \frac{\gamma}{9n^2}\right)\right) \times \left[-\frac{|\beta_1| + |\beta_2|}{n}, \frac{|\beta_1| + |\beta_2|}{n}\right]. \tag{4.109}$$

Proof. Fix $n \geq 4$. The real and imaginary part of $A^{[1,1]}_{n,n}$ are

$$\Re(A^{[1,1]}_{n,n}) = A^{[1,1]}_{n,n,D} + \frac{\gamma}{n^2} M^{[1]}_n \otimes M^{[1]}_n, \qquad \Im(A^{[1,1]}_{n,n}) = \frac{\beta_1}{\mathrm{i}n} H^{[1]}_n \otimes M^{[1]}_n + \frac{\beta_2}{\mathrm{i}n} M^{[1]}_n \otimes H^{[1]}_n.$$

The target is the localization of $\Lambda(\Re(A^{[1,1]}_{n,n}))$ and $\Lambda(\Im(A^{[1,1]}_{n,n}))$.

We begin with $\Lambda(\Re(A^{[1,1]}_{n,n}))$. By performing some computations, we have found that, since n satisfies the condition $\frac{\gamma}{9n^2} < \frac{1}{3}$, $\Re(A^{[1,1]}_{n,n})$ is Hermitian, irreducible and (by the Gershgorin theorems)

$$\Lambda(\Re(A^{[1,1]}_{n,n})) \subset \left(\frac{\gamma}{n^2}, \frac{16}{3} - \frac{\gamma}{9n^2}\right).$$

We can improve this estimate as follows. The matrix $A^{[1,1]}_{n,n,D}$ is equal to $T_{n-1,n-1}(f_{1,1})$ by Lemma 1.8 and by the fact that, as we have seen in Subsection 4.5.3, $K^{[1]}_n = T_{n-1}(f_1)$ and $M^{[1]}_n = T_{n-1}(h_1)$ (note that $h_1(\theta) = \frac{2}{3} + \frac{1}{3}\cos\theta$). The range of $f_{1,1}$ is $[0, 4]$ and so, by Theorem 1.8, $\Lambda(A^{[1,1]}_{n,n,D}) \subset (0, 4)$. Moreover, from the properties of tensors in Subsection 1.2.1 and from Lemma 4.9 it follows that $M^{[1]}_n \otimes M^{[1]}_n$ is symmetric and $\Lambda(M^{[1]}_n \otimes M^{[1]}_n) \subset (\frac{1}{9}, 1)$. By (1.9) we then have

$$\lambda_{\max}(\Re(A^{[1,1]}_{n,n})) = \lambda_{\max}(A^{[1,1]}_{n,n,D} + \frac{\gamma}{n^2} M^{[1]}_n \otimes M^{[1]}_n) \leq \lambda_{\max}(A^{[1,1]}_{n,n,D}) + \frac{\gamma}{n^2}\lambda_{\max}(M^{[1]}_n \otimes M^{[1]}_n) < 4 + \frac{\gamma}{n^2}.$$

In addition, by (1.8), by the properties of tensors in Subsection 1.2.1, by Lemma 4.9, and by the fact that $\lambda_{\min}(K^{[1]}_n) = 4\left(\sin\frac{\pi}{2n}\right)^2$, we get

$$\begin{aligned}\lambda_{\min}(\Re(A^{[1,1]}_{n,n})) &= \lambda_{\min}(A^{[1,1]}_{n,n,D} + \frac{\gamma}{n^2} M^{[1]}_n \otimes M^{[1]}_n) = \lambda_{\min}(K^{[1]}_n \otimes M^{[1]}_n + M^{[1]}_n \otimes K^{[1]}_n + \frac{\gamma}{n^2} M^{[1]}_n \otimes M^{[1]}_n) \\ &\geq \lambda_{\min}(K^{[1]}_n)\lambda_{\min}(M^{[1]}_n) + \lambda_{\min}(M^{[1]}_n)\lambda_{\min}(K^{[1]}_n) + \frac{\gamma}{n^2}\,\lambda_{\min}(M^{[1]}_n)^2 > 2 \cdot 4\left(\sin\frac{\pi}{2n}\right)^2 \frac{1}{3} + \frac{\gamma}{9n^2}.\end{aligned}$$

Therefore, we obtain for $\Lambda(\Re(A_{n,n}^{[1,1]}))$ the localization

$$\Lambda(\Re(A_{n,n}^{[1,1]})) \subset \left(\max\left(\frac{\gamma}{n^2}, \frac{8}{3}\left(\sin\frac{\pi}{2n}\right)^2 + \frac{\gamma}{9n^2}\right), \min\left(4 + \frac{\gamma}{n^2}, \frac{16}{3} - \frac{\gamma}{9n^2}\right)\right). \tag{4.110}$$

We now localize the spectrum $\Lambda(\Im(A_{n,n}^{[1,1]}))$. The matrices $H_n^{[1]} \otimes M_n^{[1]}$ and $M_n^{[1]} \otimes H_n^{[1]}$ are skew-symmetric. This follows from the properties of tensors in Subsection 1.2.1, taking into account that $H_n^{[1]}$ is skew-symmetric, while $M_n^{[1]}$ is symmetric; see Theorem 4.2. As a consequence, the matrices $\mathrm{i}\, H_n^{[1]} \otimes M_n^{[1]}$ and $\mathrm{i}\, M_n^{[1]} \otimes H_n^{[1]}$ are Hermitian, proving that all the eigenvalues of $H_n^{[1]} \otimes M_n^{[1]}$ and $M_n^{[1]} \otimes H_n^{[1]}$ are purely imaginary. Moreover, again by the properties of tensors and by Lemma 4.9, $\Lambda(H_n^{[1]} \otimes M_n^{[1]}) = \Lambda(M_n^{[1]} \otimes H_n^{[1]}) \subset \{0\} \times (-1, 1)$. Hence,

$$\begin{aligned}\lambda_{\min}(\Im(A_{n,n}^{[1,1]})) &= \lambda_{\min}\left(\frac{\beta_1}{n}\frac{1}{\mathrm{i}}H_n^{[1]} \otimes M_n^{[1]} + \frac{\beta_2}{n}\frac{1}{\mathrm{i}}M_n^{[1]} \otimes H_n^{[1]}\right) \geq \lambda_{\min}\left(\frac{\beta_1}{n}\frac{1}{\mathrm{i}}H_n^{[1]} \otimes M_n^{[1]}\right) + \lambda_{\min}\left(\frac{\beta_2}{n}\frac{1}{\mathrm{i}}M_n^{[1]} \otimes H_n^{[1]}\right) \\ &\geq -\frac{|\beta_1|}{n} - \frac{|\beta_2|}{n},\end{aligned}$$

and, similarly,

$$\lambda_{\max}(\Im(A_{n,n}^{[1,1]})) \leq \frac{|\beta_1|}{n} + \frac{|\beta_2|}{n}.$$

Therefore, we obtain for $\Lambda(\Im(A_{n,n}^{[1,1]}))$ the localization

$$\Lambda(\Im(A_{n,n}^{[1,1]})) \subseteq \left[-\frac{|\beta_1| + |\beta_2|}{n}, \frac{|\beta_1| + |\beta_2|}{n}\right]. \tag{4.111}$$

Combining (1.7) with (4.110)–(4.111), we obtain (4.109). □

In addition to providing a localization for $\Lambda(A_{n,n}^{[1,1]})$, Theorem 4.10 also shows that $\{A_{n,n}^{[1,1]}\}$ is strongly clustered at $[0, 4]$, the range of the function $f_{1,1}$. This is confirmed by the following corollary.

Corollary 4.1. *$\forall \varepsilon \in (0, 1)$ and $\forall n \geq \max\left(4, \sqrt{\frac{\gamma}{\varepsilon}}, \frac{|\beta_1|+|\beta_2|}{\varepsilon}\right)$, we have*

$$q_n(\varepsilon) = 0,$$

where $q_n(\varepsilon)$ is the number of eigenvalues of $A_{n,n}^{[1,1]}$ lying outside the rectangular ε-expansion $[0, 4]_\varepsilon$.

Proof. Fix $\varepsilon \in (0, 1)$ and $n \geq \max\left(4, \sqrt{\frac{\gamma}{\varepsilon}}, \frac{|\beta_1|+|\beta_2|}{\varepsilon}\right)$. Since n satisfies the conditions $\frac{\gamma}{9n^2} < \frac{1}{3}$, $\frac{\gamma}{n^2} \leq \varepsilon$ and $\frac{|\beta_1|+|\beta_2|}{n} \leq \varepsilon$, by Theorem 4.10 we have

$$\Lambda(A_{n,n}^{[1,1]}) \subset \left(\frac{\gamma}{n^2}, 4 + \frac{\gamma}{n^2}\right) \times \left[-\frac{|\beta_1| + |\beta_2|}{n}, \frac{|\beta_1| + |\beta_2|}{n}\right] \subset [-\varepsilon, 4 + \varepsilon] \times [-\varepsilon, \varepsilon] = [0, 4]_\varepsilon.$$

Hence, $q_n(\varepsilon) = 0$. □

4.5.6 The biquadratic case $p_1 = p_2 = 2$

In the case $p_1 = p_2 = 2$, the matrix $A_{n,n}^{[2,2]}$ is $n^2 \times n^2$ and

$$A_{n,n}^{[2,2]} = A_{n,n,D}^{[2,2]} + \frac{\beta_1}{n}H_n^{[2]} \otimes M_n^{[2]} + \frac{\beta_2}{n}M_n^{[2]} \otimes H_n^{[2]} + \frac{\gamma}{n^2}M_n^{[2]} \otimes M_n^{[2]},$$

where

$$A_{n,n,D}^{[2,2]} = K_n^{[2]} \otimes M_n^{[2]} + M_n^{[2]} \otimes K_n^{[2]}.$$

In this case, Theorem 4.7 gives $\{A_{n,n}^{[2,2]}\} \sim_\lambda f_{2,2}^{(1,1)} =: f_{2,2}$, with

$$\begin{aligned} f_{2,2}(\theta_1,\theta_2) &= (f_2 \otimes h_2)(\theta_1,\theta_2) + (h_2 \otimes f_2)(\theta_1,\theta_2) \\ &= \frac{1}{90}[99 + 6\cos(\theta_1) + 6\cos(\theta_2) - 15\cos(2\theta_1) - 15\cos(2\theta_2) - 52\cos(\theta_1)\cos(\theta_2) \\ &\qquad - 14\cos(\theta_1)\cos(2\theta_2) - 14\cos(\theta_2)\cos(2\theta_1) - \cos(2\theta_1)\cos(2\theta_2)]. \end{aligned}$$

Localization of the eigenvalues

Theorem 4.11 establishes a localization result analogous to the one that we have seen in the previous subsection.

Theorem 4.11. *For every $n \geq 5$ such that $\frac{25\gamma}{120n^2} < \frac{1}{6}$*

$$\Lambda(A_{n,n}^{[2,2]}) \subset \left(\max\left(\frac{\pi^2+10\gamma}{100n^2}, \frac{2\pi^2+\gamma}{100n^2}\right), \frac{49}{24} + \frac{\gamma}{n^2}\right) \times \left[-\frac{11}{12}\,\frac{|\beta_1|+|\beta_2|}{n}, \frac{11}{12}\,\frac{|\beta_1|+|\beta_2|}{n}\right]. \tag{4.112}$$

Proof. Fix $n \geq 5$ such that the condition $\frac{25\gamma}{120n^2} < \frac{1}{6}$ is met. The real and imaginary part of $A_{n,n}^{[2,2]}$ are given by

$$\Re(A_{n,n}^{[2,2]}) = A_{n,n,D}^{[2,2]} + \frac{\gamma}{n^2} M_n^{[2]} \otimes M_n^{[2]}, \qquad \Im(A_{n,n}^{[2,2]}) = \frac{\beta_1}{\mathrm{i}n} H_n^{[2]} \otimes M_n^{[2]} + \frac{\beta_2}{\mathrm{i}n} M_n^{[2]} \otimes H_n^{[2]}.$$

The target is now the localization of $\Lambda(\Re(A_{n,n}^{[2,2]}))$ and $\Lambda(\Im(A_{n,n}^{[2,2]}))$.

First we localize the spectrum of $\Re(A_{n,n}^{[2,2]})$. Note that

$$\Re(A_{n,n}^{[2,2]}) = A_{n,n,D}^{[2,2]} + \frac{\gamma}{n^2} M_n^{[2]} \otimes M_n^{[2]} = K_n^{[2]} \otimes M_n^{[2]} + M_n^{[2]} \otimes K_n^{[2]} + \frac{\gamma}{n^2} M_n^{[2]} \otimes M_n^{[2]} = M_n^{[2]} \otimes K_n^{[2]} + \left(K_n^{[2]} + \frac{\gamma}{n^2} M_n^{[2]}\right) \otimes M_n^{[2]}.$$

Therefore, by (1.8), the properties of tensors in Subsection 1.2.1, Lemma 4.10 and (4.97),

$$\begin{aligned} \lambda_{\min}(\Re(A_{n,n}^{[2,2]})) &\geq \lambda_{\min}(K_n^{[2]} \otimes M_n^{[2]}) + \lambda_{\min}(M_n^{[2]} \otimes K_n^{[2]}) + \frac{\gamma}{n^2}\lambda_{\min}(M_n^{[2]} \otimes M_n^{[2]}) \\ &= \lambda_{\min}(K_n^{[2]})\lambda_{\min}(M_n^{[2]}) + \lambda_{\min}(M_n^{[2]})\lambda_{\min}(K_n^{[2]}) + \frac{\gamma}{n^2}\lambda_{\min}(M_n^{[2]})\lambda_{\min}(M_n^{[2]}) \\ &> 2 \cdot \frac{\pi^2}{10n^2}\,\frac{1}{10} + \frac{\gamma}{100n^2} = \frac{2\pi^2+\gamma}{100n^2}. \end{aligned} \tag{4.113}$$

Moreover, recalling that $n \geq 5$ satisfies the condition $\frac{25\gamma}{120n^2} < \frac{1}{6}$, we can use the estimate provided in Lemma 4.10 for the spectrum of the matrix $K_n^{[2]} + \frac{\gamma}{n^2} M_n^{[2]}$. Hence, by (1.8), the properties of tensors, Lemma 4.10 and (4.97),

$$\begin{aligned} \lambda_{\min}(\Re(A_{n,n}^{[2,2]})) &\geq \lambda_{\min}(M_n^{[2]} \otimes K_n^{[2]}) + \lambda_{\min}((K_n^{[2]} + \frac{\gamma}{n^2} M_n^{[2]}) \otimes M_n^{[2]}) \\ &= \lambda_{\min}(M_n^{[2]})\lambda_{\min}(K_n^{[2]}) + \lambda_{\min}(K_n^{[2]} + \frac{\gamma}{n^2} M_n^{[2]})\lambda_{\min}(M_n^{[2]}) \\ &> \frac{1}{10}\,\frac{\pi^2}{10n^2} + \frac{\gamma}{n^2}\,\frac{1}{10} = \frac{\pi^2+10\gamma}{100n^2}. \end{aligned} \tag{4.114}$$

Furthermore, we can write

$$A_{n,n,D}^{[2,2]} = T_{n,n}(f_{2,2}) + (A_{n,n,D}^{[2,2]} - T_{n,n}(f_{2,2})) = T_{n,n}(f_{2,2}) + R_{n,n}^{[2,2]},$$

where $R_{n,n}^{[2,2]} := A_{n,n,D}^{[2,2]} - T_{n,n}(f_{2,2})$, and we can decompose $\Re(A_{n,n}^{[2,2]})$ as

$$\Re(A_{n,n}^{[2,2]}) = A_{n,n,D}^{[2,2]} + \frac{\gamma}{n^2} M_n^{[2]} \otimes M_n^{[2]} = T_{n,n}(f_{2,2}) + R_{n,n}^{[2,2]} + \frac{\gamma}{n^2} M_n^{[2]} \otimes M_n^{[2]}.$$

The range of $f_{2,2}$ is $\left[0, \frac{3}{2}\right]$, and so, by Theorem 1.8, we obtain $\Lambda(T_{n,n}(f_{2,2})) \subset \left(0, \frac{3}{2}\right)$. Concerning the symmetric matrix $R_{n,n}^{[2,2]}$, we have found by computer and by Gershgorin's first theorem that $\Lambda(R_{n,n}^{[2,2]}) \subset \left[-\frac{269}{360}, \frac{13}{24}\right]$. Using the properties of tensors in Subsection 1.2.1 and Lemma 4.10, we have $\Lambda(M_n^{[2]} \otimes M_n^{[2]}) \subset \left(\frac{1}{100}, 1\right)$. Then, we apply (1.9) to obtain an upper bound for $\lambda_{\max}(\Re(A_{n,n}^{[2,2]}))$:

$$\lambda_{\max}(\Re(A_{n,n}^{[2,2]})) \leq \lambda_{\max}(T_{n,n}(f_{2,2})) + \lambda_{\max}(R_{n,n}^{[2,2]}) + \frac{\gamma}{n^2}\lambda_{\max}(M_n^{[2]} \otimes M_n^{[2]}) < \frac{3}{2} + \frac{13}{24} + \frac{\gamma}{n^2} = \frac{49}{24} + \frac{\gamma}{n^2}. \tag{4.115}$$

Combining (4.113)–(4.115) we obtain

$$\Lambda(\Re(A_{n,n}^{[2,2]})) \subset \left(\max\left(\frac{\pi^2 + 10\gamma}{100n^2}, \frac{2\pi^2 + \gamma}{100n^2}\right), \frac{49}{24} + \frac{\gamma}{n^2}\right). \tag{4.116}$$

Now we to localize the spectrum of $\Im(A_{n,n}^{[2,2]})$. The matrices $H_n^{[2]} \otimes M_n^{[2]}$, $M_n^{[2]} \otimes H_n^{[2]}$ are skew-symmetric and, by the properties of tensors and Lemma 4.10, we have $\Lambda(H_n^{[2]} \otimes M_n^{[2]}) = \Lambda(M_n^{[2]} \otimes H_n^{[2]}) \subset \{0\} \times (-\frac{11}{12}, \frac{11}{12})$. Hence,

$$\begin{aligned}\lambda_{\min}(\Im(A_{n,n}^{[2,2]})) &= \lambda_{\min}\left(\frac{\beta_1}{n}\frac{1}{\mathrm{i}} H_n^{[2]} \otimes M_n^{[2]} + \frac{\beta_2}{n}\frac{1}{\mathrm{i}} M_n^{[2]} \otimes H_n^{[2]}\right) \\ &\geq \lambda_{\min}\left(\frac{\beta_1}{n}\frac{1}{\mathrm{i}} H_n^{[2]} \otimes M_n^{[2]}\right) + \lambda_{\min}\left(\frac{\beta_2}{n}\frac{1}{\mathrm{i}} M_n^{[2]} \otimes H_n^{[2]}\right) \geq -\frac{|\beta_1|}{n}\frac{11}{12} - \frac{|\beta_2|}{n}\frac{11}{12},\end{aligned}$$

and, similarly,

$$\lambda_{\max}(\Im(A_{n,n}^{[2,2]})) \leq \frac{|\beta_1|}{n}\frac{11}{12} + \frac{|\beta_2|}{n}\frac{11}{12}.$$

Thus,

$$\Lambda(\Im(A_{n,n}^{[2,2]})) \subseteq \left[-\frac{11}{12}\frac{|\beta_1| + |\beta_2|}{n}, \frac{11}{12}\frac{|\beta_1| + |\beta_2|}{n}\right]. \tag{4.117}$$

Using (1.7) in combination with (4.116) and (4.117), we obtain (4.112). □

Chapter 5

Spectral distribution and spectral symbol of B-spline IgA collocation matrices

This chapter, as well as the previous two, is devoted to the spectral analysis of the discretization matrices coming from a specific numerical technique for approximating the solution of a differential problem. However, as suggested by the title, in this case our spectral analysis will be focused only on the asymptotic spectral distribution of these matrices and on the associated spectral symbol. The numerical technique investigated in this chapter is the B-spline IgA Collocation Method, which has been recently introduced in [3, 53] and will be described later on. As for the differential problem, we consider the following linear full elliptic second-order PDE with homogeneous Dirichlet boundary conditions:

$$\begin{cases} -\nabla \cdot K\nabla u + \boldsymbol{\alpha} \cdot \nabla u + \gamma u = f & \text{in } \Omega, \\ u = 0 & \text{on } \partial\Omega, \end{cases} \tag{5.1}$$

where Ω is a bounded open domain in $\mathbb{R}^d$, $K : \overline{\Omega} \to \mathbb{R}^{d\times d}$ is a SPD matrix of functions in $C^1(\Omega) \cap C(\overline{\Omega})$, $\boldsymbol{\alpha} : \overline{\Omega} \to \mathbb{R}^d$ is a vector of functions in $C(\overline{\Omega})$, $\gamma, f \in C(\overline{\Omega})$ and $\gamma \geq 0$. Note that problem (5.1) is more complex than the ones considered in Chapters 3 and 4, due to the presence of the diffusion coefficient K and to the arbitrary shape of the domain Ω, which is no longer supposed to be rectangular.

As in the previous two chapters, we first describe the B-spline IgA Collocation Method and we give a construction of the inherently non-symmetric matrices arising from this approximation technique. After this, we find and study the associated spectral symbol, which describes their asymptotic spectral distribution when the matrix size tends to infinity or, equivalently, when the fineness parameters tend to zero. The specific properties of the symbol studied in this chapter will be used in Chapter 7 to design a fast multi-iterative solver of multigrid type for the B-spline IgA collocation matrices.

5.1 B-spline IgA Collocation Method

Problem (5.1) can be reformulated as follows:

$$\begin{cases} -1(K \circ Pu)1^T + \boldsymbol{\beta} \cdot \nabla u + \gamma u = f, & \text{in } \Omega, \\ u = 0, & \text{on } \partial\Omega, \end{cases} \tag{5.2}$$

where $1 := (1, \ldots, 1) \in \mathbb{N}^d$, Pu denotes the Hessian of u, i.e.

$$(Pu)_{i,j} := \frac{\partial^2 u}{\partial x_i \partial x_j}, \tag{5.3}$$

and $\circ$ denotes the componentwise Hadamard matrix product; see Subsection 1.2.2. Moreover, $\boldsymbol{\beta}$ collects the coefficients of the first order derivatives in (5.1), namely

$$\beta_j := \alpha_j - \sum_{i=1}^{d} \frac{\partial \kappa_{ij}}{\partial x_i}, \tag{5.4}$$

where κ_{ij} are the entries of the matrix $K := [\kappa_{ij}]_{i,j=1}^{d}$.

We consider the approximation of the solution of problem (5.2) by the standard collocation approach, as explained briefly in the following. Let W be a finite dimensional vector space of sufficiently smooth functions defined on $\overline{\Omega}$ and vanishing on the boundary $\partial\Omega$. We call W the approximation space. Then, we introduce a set of $N := \dim W$ collocation points in Ω,

$$\{\boldsymbol{\tau}_i \in \Omega, \quad i = 1, \dots, N\},$$

and we look for a function $u_W \in W$ such that

$$-1(K(\boldsymbol{\tau}_i) \circ Pu_W(\boldsymbol{\tau}_i))1^T + \boldsymbol{\beta}(\boldsymbol{\tau}_i) \cdot \nabla u_W(\boldsymbol{\tau}_i) + \gamma(\boldsymbol{\tau}_i)u_W(\boldsymbol{\tau}_i) = f(\boldsymbol{\tau}_i), \quad \forall \boldsymbol{\tau}_i. \tag{5.5}$$

If we fix a basis $\{\varphi_1, \dots, \varphi_N\}$ for W, then each $v \in W$ can be written as $v = \sum_{j=1}^{N} v_j \varphi_j$, and the collocation problem (5.5) is equivalent to solving the linear system

$$A\mathbf{u} = \mathbf{f}, \tag{5.6}$$

where

$$A := \left[-1(K(\boldsymbol{\tau}_i) \circ P\varphi_j(\boldsymbol{\tau}_i))1^T + \boldsymbol{\beta}(\boldsymbol{\tau}_i) \cdot \nabla\varphi_j(\boldsymbol{\tau}_i) + \gamma(\boldsymbol{\tau}_i)\varphi_j(\boldsymbol{\tau}_i)\right]_{i,j=1}^{N} \in \mathbb{R}^{N\times N} \tag{5.7}$$

is the collocation matrix and $\mathbf{f} := [f(\boldsymbol{\tau}_i)]_{i=1}^{N}$. Once we find $\mathbf{u} := [u_1 \ \cdots \ u_N]^T$, we know $u_W = \sum_{j=1}^{N} u_j \varphi_j$.

Let us now describe the isogeometric collocation approach. Let

$$\{\hat{\varphi}_1, \dots, \hat{\varphi}_{N+N_b}\} \tag{5.8}$$

be a set of basis functions defined on the parametric domain $\hat{\Omega} := [0,1]^d$, and assume that the physical domain Ω in (5.2) can be described by a global geometry function $\mathbf{G}$ expressed in terms of the functions $\hat{\varphi}_i$ as follows:

$$\mathbf{G} : \hat{\Omega} \to \overline{\Omega}, \qquad \mathbf{G}(\hat{\mathbf{x}}) := \sum_{i=1}^{N+N_b} \hat{\varphi}_i(\hat{\mathbf{x}})\, \mathbf{p}_i, \qquad \mathbf{p}_i \in \mathbb{R}^d. \tag{5.9}$$

We assume that the map $\mathbf{G}$ is invertible in $\hat{\Omega}$ and $\mathbf{G}(\partial\hat{\Omega}) = \partial\overline{\Omega}$. If $\{\hat{\varphi}_1, \dots, \hat{\varphi}_N\}$ is defined as the subset of the functions in (5.8) which vanish on the boundary $\partial\hat{\Omega}$, then the approximation space W is defined as the vector space spanned by

$$\varphi_i(\mathbf{x}) := \hat{\varphi}_i(\mathbf{G}^{-1}(\mathbf{x})) = \hat{\varphi}_i(\hat{\mathbf{x}}), \qquad i = 1, \dots, N, \qquad \mathbf{x} = \mathbf{G}(\hat{\mathbf{x}}). \tag{5.10}$$

Moreover, we introduce a set of collocation points in the parametric domain $\hat{\Omega}$,

$$\{\hat{\boldsymbol{\tau}}_i \in \hat{\Omega}, \quad i = 1, \dots, N\}, \tag{5.11}$$

and we define the collocation points in the physical domain Ω as follows:

$$\boldsymbol{\tau}_i := \mathbf{G}(\hat{\boldsymbol{\tau}}_i), \qquad i = 1, \dots, N. \tag{5.12}$$

In the isogeometric collocation approach, we solve the linear system (5.6) with the basis functions and the collocation points given by (5.10) and (5.12), respectively. In the most common formulation of IgA, the functions $\hat{\varphi}_i$ in (5.8) are tensor-product B-splines or NURBS, since they allow an exact representation – by definition – of an arbitrary domain designed in a (NURBS-based) CAD system. Nonetheless, other kinds of functions can be used as well.

In this chapter, we study the asymptotic spectral distribution and the symbol of the B-spline IgA collocation matrices (5.7), obtained from the approximation of problem (5.2) by isogeometric collocation methods

based on tensor-product B-splines with equally spaced knots. This means that the role of the functions $\hat{\varphi}_i$ in (5.8) will be played by tensor-product B-splines over uniform knot sequences. In addition, we do not confine ourselves to the isoparametric approach, since we will not require the geometry map $\mathbf{G}$ to be expressed in terms of the $\hat{\varphi}_i$ as in (5.9). As for the choice of the collocation points, which is crucial for the stability and good behavior of the discrete problem, we follow [3]: our collocation points $\hat{\boldsymbol{\tau}}_i$ in (5.11) are chosen as the Greville abscissae corresponding to the used B-splines.

We now provide the explicit expression of our basis functions $\hat{\varphi}_i$ and of our collocation points $\hat{\boldsymbol{\tau}}_i$. For $p, n \geq 2$, consider the (uniform) B-splines $N_{i,[p]}$, $i = 2, \dots, n+p-1$, corresponding to the knot sequence (4.2)–(4.3) (see Definition 4.1), whose associated Greville abscissae are

$$\xi_{i,[p]} := \frac{t_{i+1} + t_{i+2} + \ldots + t_{i+p}}{p}, \qquad i = 2, \dots, n+p-1. \tag{5.13}$$

For any $\boldsymbol{p} = (p_1, \dots, p_d)$ and $\boldsymbol{n} = (n_1, \dots, n_d) \in \mathbb{N}^d$, with $p_i, n_i \geq 2$ for all $i = 1, \dots, d$, let

$$N_{\boldsymbol{i},[\boldsymbol{p}]} := N_{i_1,[p_1]} \otimes N_{i_2,[p_2]} \otimes \cdots \otimes N_{i_d,[p_d]} : \hat{\Omega} \to \mathbb{R}, \qquad \mathbf{2} \leq \boldsymbol{i} \leq \boldsymbol{n}+\boldsymbol{p}-\mathbf{1}, \tag{5.14}$$

$$\boldsymbol{\xi}_{\boldsymbol{i},[\boldsymbol{p}]} := (\xi_{i_1,[p_1]}, \xi_{i_2,[p_2]}, \dots, \xi_{i_d,[p_d]}), \qquad \mathbf{2} \leq \boldsymbol{i} \leq \boldsymbol{n}+\boldsymbol{p}-\mathbf{1}. \tag{5.15}$$

In the framework of the B-spline IgA Collocation Method, the functions $\hat{\varphi}_i$, $i = 1, \dots, N$, in (5.8) are chosen as the tensor-product B-splines in (5.14) and the collocation points $\hat{\boldsymbol{\tau}}_i$, $i = 1, \dots, N$, in (5.11) are chosen as the Greville abscissae in (5.15). In this case, $N = \prod_{k=1}^{d}(n_k + p_k - 2) = N(\boldsymbol{n}+\boldsymbol{p}-\mathbf{2})$. Of course, we adopt for the tensor-product B-splines (5.14) and for the associated Greville abscissae (5.15) the standard lexicographic ordering, which is obtained by varying the multi-index $\boldsymbol{i}$ from $\mathbf{2}$ to $\boldsymbol{n}+\boldsymbol{p}-\mathbf{1}$ according to the rule in (1.1). By definition, for every $i = 1, \dots, N$, the i-th tensor-product B-spline in (5.14) and the i-th Greville abscissa in (5.15) according to the ordering (1.1) are, respectively, $\hat{\varphi}_i$ and $\hat{\boldsymbol{\tau}}_i$. This should be taken into consideration when assembling the collocation matrix (5.7).

5.2 Construction of the B-spline IgA collocation matrices $A_{\boldsymbol{n}}^{[\boldsymbol{p}]}$

In the case where $\mathbf{G}$ is the identity map (and so $\overline{\Omega} = \hat{\Omega} = [0,1]^d$), the collocation matrix (5.7) resulting from (5.10), (5.12) and from the choices of $\hat{\varphi}_i$ and $\hat{\boldsymbol{\tau}}_i$ as in (5.14)–(5.15) is

$$A_{\boldsymbol{n}}^{[\boldsymbol{p}]} = A_{\boldsymbol{n},D}^{[\boldsymbol{p}]} + A_{\boldsymbol{n},A}^{[\boldsymbol{p}]} + A_{\boldsymbol{n},R}^{[\boldsymbol{p}]}, \tag{5.16}$$

where

$$\begin{aligned} A_{\boldsymbol{n},D}^{[\boldsymbol{p}]} &:= \left[-\mathbf{1}(K(\boldsymbol{\tau}_i) \circ P\varphi_j(\boldsymbol{\tau}_i))\mathbf{1}^T\right]_{i,j=1}^{N} = \left[-\mathbf{1}(K(\boldsymbol{\xi}_{\boldsymbol{i}+\mathbf{1},[\boldsymbol{p}]}) \circ PN_{\boldsymbol{j}+\mathbf{1},[\boldsymbol{p}]}(\boldsymbol{\xi}_{\boldsymbol{i}+\mathbf{1},[\boldsymbol{p}]}))\mathbf{1}^T\right]_{\boldsymbol{i},\boldsymbol{j}=\mathbf{1}}^{\boldsymbol{n}+\boldsymbol{p}-\mathbf{2}} \\ &= \sum_{r=1}^{d} n_r^2 D_{\boldsymbol{n}}^{[\boldsymbol{p}]}(\kappa_{rr})(M_{n_1}^{[p_1]} \otimes \cdots \otimes M_{n_{r-1}}^{[p_{r-1}]} \otimes K_{n_r}^{[p_r]} \otimes M_{n_{r+1}}^{[p_{r+1}]} \otimes \cdots \otimes M_{n_d}^{[p_d]}) \\ &\quad - \sum_{\substack{r,s=1 \\ r<s}}^{d} n_r n_s D_{\boldsymbol{n}}^{[\boldsymbol{p}]}(\kappa_{rs}+\kappa_{sr})(M_{n_1}^{[p_1]} \otimes \cdots \otimes M_{n_{r-1}}^{[p_{r-1}]} \otimes H_{n_r}^{[p_r]} \otimes M_{n_{r+1}}^{[p_{r+1}]} \otimes \cdots \otimes M_{n_{s-1}}^{[p_{s-1}]} \otimes H_{n_s}^{[p_s]} \otimes M_{n_{s+1}}^{[p_{s+1}]} \otimes \cdots \otimes M_{n_d}^{[p_d]}), \end{aligned} \tag{5.17}$$

$$\begin{aligned} A_{\boldsymbol{n},A}^{[\boldsymbol{p}]} &:= \left[\boldsymbol{\beta}(\boldsymbol{\tau}_i) \cdot \nabla\varphi_j(\boldsymbol{\tau}_i)\right]_{i,j=1}^{N} = \left[\boldsymbol{\beta}(\boldsymbol{\xi}_{\boldsymbol{i}+\mathbf{1},[\boldsymbol{p}]}) \cdot \nabla N_{\boldsymbol{j}+\mathbf{1},[\boldsymbol{p}]}(\boldsymbol{\xi}_{\boldsymbol{i}+\mathbf{1},[\boldsymbol{p}]})\right]_{\boldsymbol{i},\boldsymbol{j}=\mathbf{1}}^{\boldsymbol{n}+\boldsymbol{p}-\mathbf{2}} \\ &= \sum_{r=1}^{d} n_r D_{\boldsymbol{n}}^{[\boldsymbol{p}]}(\beta_r)(M_{n_1}^{[p_1]} \otimes \cdots \otimes M_{n_{r-1}}^{[p_{r-1}]} \otimes H_{n_r}^{[p_r]} \otimes M_{n_{r+1}}^{[p_{r+1}]} \otimes \cdots \otimes M_{n_d}^{[p_d]}), \end{aligned} \tag{5.18}$$

$$A_{\boldsymbol{n},R}^{[\boldsymbol{p}]} := \left[\gamma(\boldsymbol{\tau}_i)\varphi_j(\boldsymbol{\tau}_i)\right]_{i,j=1}^{N} = \left[\gamma(\boldsymbol{\xi}_{\boldsymbol{i}+\mathbf{1},[\boldsymbol{p}]})N_{\boldsymbol{j}+\mathbf{1},[\boldsymbol{p}]}(\boldsymbol{\xi}_{\boldsymbol{i}+\mathbf{1},[\boldsymbol{p}]})\right]_{\boldsymbol{i},\boldsymbol{j}=\mathbf{1}}^{\boldsymbol{n}+\boldsymbol{p}-\mathbf{2}} = D_{\boldsymbol{n}}^{[\boldsymbol{p}]}(\gamma)(M_{n_1}^{[p_1]} \otimes \cdots \otimes M_{n_d}^{[p_d]}), \tag{5.19}$$

$D_{\boldsymbol{n}}^{[\boldsymbol{p}]}(a)$ denotes the d-level diagonal sampling matrix containing the samples of the function a at the Greville abscissae, i.e.

$$D_{\boldsymbol{n}}^{[\boldsymbol{p}]}(a) := \underset{\boldsymbol{i}=\boldsymbol{1},\ldots,\boldsymbol{n}+\boldsymbol{p}-\boldsymbol{2}}{\operatorname{diag}} a(\boldsymbol{\xi}_{\boldsymbol{i}+\boldsymbol{1},[\boldsymbol{p}]}) = \operatorname{diag}\left(a(\boldsymbol{\xi}_{\boldsymbol{2},[\boldsymbol{p}]}),\ldots,a(\boldsymbol{\xi}_{\boldsymbol{n}+\boldsymbol{p}-\boldsymbol{1},[\boldsymbol{p}]})\right), \tag{5.20}$$

and the matrices $K_n^{[p]}$, $H_n^{[p]}$, $M_n^{[p]}$ are defined for all $p, n \geq 2$ by

$$n^2 K_n^{[p]} := \left[-N''_{j+1,[p]}(\xi_{i+1,[p]})\right]_{i,j=1}^{n+p-2}, \qquad nH_n^{[p]} := \left[N'_{j+1,[p]}(\xi_{i+1,[p]})\right]_{i,j=1}^{n+p-2}, \qquad M_n^{[p]} := \left[N_{j+1,[p]}(\xi_{i+1,[p]})\right]_{i,j=1}^{n+p-2}. \tag{5.21}$$

This result can be proved once again by using the fundamental property (1.12) of tensor products, and the proof follows the same pattern as the proof of Theorem 3.1. Let us derive, for instance, the expression of $A_{\boldsymbol{n},D}^{[\boldsymbol{p}]}$ in (5.17): for all $\boldsymbol{i},\boldsymbol{j} = \boldsymbol{1},\ldots,\boldsymbol{n}+\boldsymbol{p}-\boldsymbol{2}$, we have

$$\begin{aligned}
(A_{\boldsymbol{n},D}^{[\boldsymbol{p}]})_{\boldsymbol{ij}} &= -\mathbf{1}(K(\boldsymbol{\xi}_{\boldsymbol{i}+\boldsymbol{1},[\boldsymbol{p}]}) \circ PN_{\boldsymbol{j}+\boldsymbol{1},[\boldsymbol{p}]}(\boldsymbol{\xi}_{\boldsymbol{i}+\boldsymbol{1},[\boldsymbol{p}]}))\mathbf{1}^T = -\sum_{r,s=1}^{d} \kappa_{rs}(\boldsymbol{\xi}_{\boldsymbol{i}+\boldsymbol{1},[\boldsymbol{p}]}) \frac{\partial^2 N_{\boldsymbol{j}+\boldsymbol{1},[\boldsymbol{p}]}}{\partial \hat{x}_r \partial \hat{x}_s}(\boldsymbol{\xi}_{\boldsymbol{i}+\boldsymbol{1},[\boldsymbol{p}]}) \\
&= -\sum_{r=1}^{d} \kappa_{rr}(\boldsymbol{\xi}_{\boldsymbol{i}+\boldsymbol{1},[\boldsymbol{p}]}) \frac{\partial^2 N_{\boldsymbol{j}+\boldsymbol{1},[\boldsymbol{p}]}}{\partial \hat{x}_r^2}(\boldsymbol{\xi}_{\boldsymbol{i}+\boldsymbol{1},[\boldsymbol{p}]}) - \sum_{\substack{r,s=1 \\ r<s}}^{d} (\kappa_{rs}(\boldsymbol{\xi}_{\boldsymbol{i}+\boldsymbol{1},[\boldsymbol{p}]}) + \kappa_{sr}(\boldsymbol{\xi}_{\boldsymbol{i}+\boldsymbol{1},[\boldsymbol{p}]})) \frac{\partial^2 N_{\boldsymbol{j}+\boldsymbol{1},[\boldsymbol{p}]}}{\partial \hat{x}_r \partial \hat{x}_s}(\boldsymbol{\xi}_{\boldsymbol{i}+\boldsymbol{1},[\boldsymbol{p}]}) \\
&= -\sum_{r=1}^{d} \kappa_{rr}(\boldsymbol{\xi}_{\boldsymbol{i}+\boldsymbol{1},[\boldsymbol{p}]})(N_{j_1+1,[p_1]} \otimes \cdots \otimes N_{j_{r-1}+1,[p_{r-1}]} \otimes N''_{j_r+1,[p_r]} \otimes N_{j_{r+1}+1,[p_{r+1}]} \otimes \cdots \otimes N_{j_d+1,[p_d]})(\boldsymbol{\xi}_{\boldsymbol{i}+\boldsymbol{1},[\boldsymbol{p}]}) \\
&\quad - \sum_{\substack{r,s=1 \\ r<s}}^{d} (\kappa_{rs}(\boldsymbol{\xi}_{\boldsymbol{i}+\boldsymbol{1},[\boldsymbol{p}]}) + \kappa_{sr}(\boldsymbol{\xi}_{\boldsymbol{i}+\boldsymbol{1},[\boldsymbol{p}]}))(N_{j_1+1,[p_1]} \otimes \cdots \otimes N_{j_{r-1}+1,[p_{r-1}]} \otimes N'_{j_r+1,[p_r]} \otimes N_{j_{r+1}+1,[p_{r+1}]} \otimes \cdots \otimes N_{j_{s-1}+1,[p_{s-1}]} \\
&\qquad\qquad \otimes N'_{j_s+1,[p_s]} \otimes N_{j_{s+1}+1,[p_{s+1}]} \otimes \cdots \otimes N_{j_d+1,[p_d]})(\boldsymbol{\xi}_{\boldsymbol{i}+\boldsymbol{1},[\boldsymbol{p}]}) \\
&= \sum_{r=1}^{d} \kappa_{rr}(\boldsymbol{\xi}_{\boldsymbol{i}+\boldsymbol{1},[\boldsymbol{p}]})(-N''_{j_r+1,[p_r]}(\xi_{i_r+1,[p_r]})) \prod_{\substack{t=1 \\ t\neq j}}^{d} N_{j_t+1,[p_t]}(\xi_{i_t+1,[p_t]}) \\
&\quad - \sum_{\substack{r,s=1 \\ r<s}}^{d} (\kappa_{rs}(\boldsymbol{\xi}_{\boldsymbol{i}+\boldsymbol{1},[\boldsymbol{p}]}) + \kappa_{sr}(\boldsymbol{\xi}_{\boldsymbol{i}+\boldsymbol{1},[\boldsymbol{p}]})) N'_{j_r+1,[p_r]}(\xi_{i_r+1,[p_r]}) N'_{j_s+1,[p_s]}(\xi_{i_s+1,[p_s]}) \prod_{\substack{t=1 \\ t\neq r,s}}^{d} N_{j_t+1,[p_t]}(\xi_{i_t+1,[p_t]}) \\
&= \sum_{r=1}^{d} [D_{\boldsymbol{n}}^{[\boldsymbol{p}]}(\kappa_{rr})]_{\boldsymbol{ii}} (n_r^2 K_{n_r}^{[p_r]})_{i_r j_r} \prod_{\substack{t=1 \\ t\neq j}}^{d} (M_{n_t}^{[p_t]})_{i_t j_t} - \sum_{\substack{r,s=1 \\ r<s}}^{d} [D_{\boldsymbol{n}}^{[\boldsymbol{p}]}(\kappa_{rs}+\kappa_{sr})]_{\boldsymbol{ii}} (n_r H_{n_r}^{[p_r]})_{i_r j_r} (n_s H_{n_s}^{[p_s]})_{i_s j_s} \prod_{\substack{t=1 \\ t\neq r,s}}^{d} (M_{n_t}^{[p_t]})_{i_t j_t} \\
&= \sum_{r=1}^{d} [D_{\boldsymbol{n}}^{[\boldsymbol{p}]}(\kappa_{rr})]_{\boldsymbol{ii}} (M_{n_1}^{[p_1]} \otimes \cdots \otimes M_{n_{r-1}}^{[p_{r-1}]} \otimes n_r^2 K_{n_r}^{[p_r]} \otimes M_{n_{r+1}}^{[p_{r+1}]} \otimes \cdots \otimes M_{n_d}^{[p_d]})_{\boldsymbol{ij}} \\
&\quad - \sum_{\substack{r,s=1 \\ r<s}}^{d} [D_{\boldsymbol{n}}^{[\boldsymbol{p}]}(\kappa_{rs}+\kappa_{sr})]_{\boldsymbol{ii}} (M_{n_1}^{[p_1]} \otimes \cdots \otimes M_{n_{r-1}}^{[p_{r-1}]} \otimes n_r H_{n_r}^{[p_r]} \otimes M_{n_{r+1}}^{[p_{r+1}]} \otimes \cdots \otimes M_{n_{s-1}}^{[p_{s-1}]} \otimes n_s H_{n_s}^{[p_s]} \otimes M_{n_{s+1}}^{[p_{s+1}]} \otimes \cdots \otimes M_{n_d}^{[p_d]})_{\boldsymbol{ij}} \\
&= \sum_{r=1}^{d} [D_{\boldsymbol{n}}^{[\boldsymbol{p}]}(\kappa_{rr})(M_{n_1}^{[p_1]} \otimes \cdots \otimes M_{n_{r-1}}^{[p_{r-1}]} \otimes n_r^2 K_{n_r}^{[p_r]} \otimes M_{n_{r+1}}^{[p_{r+1}]} \otimes \cdots \otimes M_{n_d}^{[p_d]})]_{\boldsymbol{ij}} \\
&\quad - \sum_{\substack{r,s=1 \\ r<s}}^{d} [D_{\boldsymbol{n}}^{[\boldsymbol{p}]}(\kappa_{rs}+\kappa_{sr})(M_{n_1}^{[p_1]} \otimes \cdots \otimes M_{n_{r-1}}^{[p_{r-1}]} \otimes n_r H_{n_r}^{[p_r]} \otimes M_{n_{r+1}}^{[p_{r+1}]} \otimes \cdots \otimes M_{n_{s-1}}^{[p_{s-1}]} \otimes n_s H_{n_s}^{[p_s]} \otimes M_{n_{s+1}}^{[p_{s+1}]} \otimes \cdots \otimes M_{n_d}^{[p_d]})]_{\boldsymbol{ij}},
\end{aligned}$$

and (5.17) follows. In the case $d = 1$, from (5.16)–(5.19) we have

$$A_n^{[p]} = n^2 D_n^{[p]}(\kappa) K_n^{[p]} + n D_n^{[p]}(\beta) H_n^{[p]} + D_n^{[p]}(\gamma) M_n^{[p]}, \tag{5.22}$$

with $D_n^{[p]}(a) = \mathrm{diag}_{j=1,\dots,n+p-2}(a(\xi_{j+1,[p]}))$, and for $d = 2$ we have

$$\begin{aligned} A_{n_1,n_2}^{[p_1,p_2]} &= n_1^2 D_{n_1,n_2}^{[p_1,p_2]}(\kappa_{11}) M_{n_1}^{[p_1]} \otimes K_{n_2}^{[p_2]} + n_2^2 D_{n_1,n_2}^{[p_1,p_2]}(\kappa_{22}) K_{n_1}^{[p_1]} \otimes M_{n_2}^{[p_2]} - n_1 n_2 D_{n_1,n_2}^{[p_1,p_2]}(\kappa_{12}+\kappa_{21}) H_{n_1}^{[p_1]} \otimes H_{n_2}^{[p_2]} \\ &\quad + n_1 D_{n_1,n_2}^{[p_1,p_2]}(\beta_1) M_{n_1}^{[p_1]} \otimes H_{n_2}^{[p_2]} + n_2 D_{n_1,n_2}^{[p_1,p_2]}(\beta_2) H_{n_1}^{[p_1]} \otimes M_{n_2}^{[p_2]} + D_{n_1,n_2}^{[p_1,p_2]}(\gamma) M_{n_1}^{[p_1]} \otimes M_{n_2}^{[p_2]}, \end{aligned} \tag{5.23}$$

with $D_{n_1,n_2}^{[p_1,p_2]}(a) = \mathrm{diag}_{i_1=1,\dots,n_1+p_1-2}\left(\mathrm{diag}_{i_2=2,\dots,n_2+p_2-2}(a(\boldsymbol{\xi}_{i_1+1,i_2+1,[p_1,p_2]}))\right)$.

In the general case where Ω and $\mathbf{G}$ are nontrivial, let us consider, for any $u : \Omega \to \mathbb{R}$, the corresponding function

$$\hat{u} : \hat{\Omega} \to \mathbb{R}, \qquad \hat{u}(\hat{\mathbf{x}}) := u(\mathbf{x}), \qquad \mathbf{x} = \mathbf{G}(\hat{\mathbf{x}}).$$

In other words, $\hat{u} := u(\mathbf{G})$. Then, u satisfies (5.2) if and only if $\hat{u}$ satisfies the corresponding transformed problem

$$\begin{cases} -\mathbf{1}(K_{\mathbf{G}} \circ P\hat{u})\mathbf{1}^T + \boldsymbol{\beta}_{\mathbf{G}} \cdot \nabla \hat{u} + \gamma_{\mathbf{G}} \hat{u} = f_{\mathbf{G}} & \text{in } \hat{\Omega}, \\ \hat{u} = 0 & \text{on } \partial\hat{\Omega}, \end{cases} \tag{5.24}$$

where $P\hat{u}$ is the Hessian of $\hat{u}$, $\gamma_{\mathbf{G}} := \gamma(\mathbf{G})$, $f_{\mathbf{G}} := f(\mathbf{G})$, and $K_{\mathbf{G}} := [\kappa_{\mathbf{G},ij}]_{i,j=1}^d$, $\boldsymbol{\beta}_{\mathbf{G}} := [\beta_{\mathbf{G},i}]_{i=1}^d$ are the transformed coefficients of the PDE. The expression of $\boldsymbol{\beta}_{\mathbf{G}}$ in terms of K, $\boldsymbol{\beta}$, $\mathbf{G}$ is complicated and hence not reported here, while for $K_{\mathbf{G}}$ we have

$$K_{\mathbf{G}} = (J_{\mathbf{G}})^{-1} K(\mathbf{G}) (J_{\mathbf{G}})^{-T}, \tag{5.25}$$

where $J_{\mathbf{G}}$ is the Jacobian matrix of $\mathbf{G}$,

$$J_{\mathbf{G}} := \left[\frac{\partial G_i}{\partial \hat{x}_j}\right]_{i,j=1}^d.$$

In this case, the collocation matrix (5.7), with φ_i, $\boldsymbol{\tau}_i$ as in (5.10), (5.12) and $\hat{\varphi}_i$, $\hat{\boldsymbol{\tau}}_i$ as in (5.14)–(5.15), is

$$A_{\mathbf{G},\boldsymbol{n}}^{[\boldsymbol{p}]} = A_{\mathbf{G},\boldsymbol{n},D}^{[\boldsymbol{p}]} + A_{\mathbf{G},\boldsymbol{n},A}^{[\boldsymbol{p}]} + A_{\mathbf{G},\boldsymbol{n},R}^{[\boldsymbol{p}]},$$

where $A_{\mathbf{G},\boldsymbol{n},D}^{[\boldsymbol{p}]}$, $A_{\mathbf{G},\boldsymbol{n},A}^{[\boldsymbol{p}]}$, $A_{\mathbf{G},\boldsymbol{n},R}^{[\boldsymbol{p}]}$ are given again by (5.17)–(5.19), in which κ_{rs}, β_r, γ are replaced by $\kappa_{\mathbf{G},rs}$, $\beta_{\mathbf{G},r}$, $\gamma_{\mathbf{G}}$, respectively.

For example, let us consider problem (5.2) in the one-dimensional case $d = 1$ with $\Omega = (a, b)$:

$$\begin{cases} -\kappa(x)u''(x) + \beta(x)u'(x) + \gamma(x)u(x) = f(x) & a < x < b, \\ u(a) = u(b) = 0. \end{cases}$$

Given any geometry function $G : [0,1] \to [a,b]$, the transformed problem reads as

$$\begin{cases} -\dfrac{\kappa(G(\hat{x}))}{(G'(\hat{x}))^2}\hat{u}''(\hat{x}) + \left(\dfrac{\kappa(G(\hat{x}))G''(\hat{x})}{(G'(\hat{x}))^3} + \dfrac{\beta(G(\hat{x}))}{G'(\hat{x})}\right)\hat{u}'(\hat{x}) + \gamma(G(\hat{x}))\hat{u}(\hat{x}) = f(G(\hat{x})) & 0 < \hat{x} < 1, \\ \hat{u}(0) = \hat{u}(1) = 0, \end{cases}$$

and the resulting collocation matrix $A_{G,n}^{[p]}$ is given by (5.22) in which κ, β, γ are replaced by

$$\kappa_G := \frac{\kappa(G)}{(G')^2}, \qquad \beta_G := \frac{\kappa(G)G''}{(G')^3} + \frac{\beta(G)}{G'}, \qquad \gamma_G := \gamma(G), \tag{5.26}$$

respectively. Note that κ_G is given precisely by (5.25), because $J_G = G'$.

5.2.1 Construction of $K_n^{[p]}$, $H_n^{[p]}$, $M_n^{[p]}$

We now provide, for $p, n \geq 2$, the constructions of the matrices $K_n^{[p]}$, $H_n^{[p]}$, $M_n^{[p]}$ in (5.21). Note that the elements of these matrices can be computed by using the recurrence relation (4.4) and by iterating the derivative formula (4.74). As mentioned in Chapter 4 (see Subsection 4.3.1), the 'central' basis functions $N_{j,[p]}(x)$, $j = p+1, \ldots, n$, are 'uniformly shifted and scaled versions' of the cardinal B-spline $\phi_{[p]}$ introduced in Section 4.2. Indeed, by (4.42) we have

$$N_{j,[p]}(x) = \phi_{[p]}(nx - j + p + 1), \quad j = p+1, \ldots, n, \tag{5.27}$$

and, consequently,

$$N'_{j,[p]}(x) = n\,\dot{\phi}_{[p]}(nx - j + p + 1), \quad j = p+1, \ldots, n,$$
$$N''_{j,[p]}(x) = n^2\,\ddot{\phi}_{[p]}(nx - j + p + 1), \quad j = p+1, \ldots, n.$$

In addition, the 'interior' Greville abscissae, given by (5.13) for $i = p+1, \ldots, n$, simplify to

$$\xi_{i,[p]} = \frac{i}{n} - \frac{p+1}{2n}, \quad i = p+1, \ldots, n, \tag{5.28}$$

or, equivalently,

$$n\xi_{i,[p]} + p + 1 = i + \frac{p+1}{2}, \quad i = p+1, \ldots, n.$$

We now focus on the 'central part' of $K_n^{[p]}$, $H_n^{[p]}$, $M_n^{[p]}$, which is determined only by the central basis functions in (5.27) and by the interior Greville abscissae (5.28). In other words, we focus on the submatrices

$$[(K_n^{[p]})_{ij}]_{i,j=p}^{n-1}, \qquad [(H_n^{[p]})_{ij}]_{i,j=p}^{n-1}, \qquad [(M_n^{[p]})_{ij}]_{i,j=p}^{n-1},$$

which are nonempty for $n \geq p+1$. The entries of these submatrices are given explicitly by

$$(K_n^{[p]})_{ij} = -\ddot{\phi}_{[p]}\left(\frac{p+1}{2} + i - j\right) = -\ddot{\phi}_{[p]}\left(\frac{p+1}{2} - i + j\right),$$
$$(H_n^{[p]})_{ij} = \dot{\phi}_{[p]}\left(\frac{p+1}{2} + i - j\right) = -\dot{\phi}_{[p]}\left(\frac{p+1}{2} - i + j\right),$$
$$(M_n^{[p]})_{ij} = \phi_{[p]}\left(\frac{p+1}{2} + i - j\right) = \phi_{[p]}\left(\frac{p+1}{2} - i + j\right),$$

for $i, j = p, \ldots, n-1$, where in the last equalities we have invoked Lemma 4.1. It follows that the above central submatrices of $K_n^{[p]}$ and $M_n^{[p]}$ are symmetric, whereas the above central submatrix of $H_n^{[p]}$ is skew-symmetric. We note that the coefficients depend only on the difference $i-j$, and so all the above submatrices are Toeplitz matrices. In fact, recalling (1.29) and the properties $\mathrm{supp}(\phi_{[p]}) = [0, p+1]$ and $\dot{\phi}_{[p]}\left(\frac{p+1}{2}\right) = 0$, we have

$$[(K_n^{[p]})_{ij}]_{i,j=p}^{n-1} = T_{n-p}(f_p), \tag{5.29}$$
$$[(H_n^{[p]})_{ij}]_{i,j=p}^{n-1} = \mathrm{i}T_{n-p}(g_p), \tag{5.30}$$
$$[(M_n^{[p]})_{ij}]_{i,j=p}^{n-1} = T_{n-p}(h_p), \tag{5.31}$$

where the functions h_p, g_p, $f_p : [-\pi, \pi] \to \mathbb{R}$ are defined by

$$h_p(\theta) := \sum_{k\in\mathbb{Z}} \phi_{[p]}\left(\frac{p+1}{2} - k\right) \mathrm{e}^{\mathrm{i}k\theta} = \phi_{[p]}\left(\frac{p+1}{2}\right) + 2\sum_{k=1}^{\lfloor p/2\rfloor} \phi_{[p]}\left(\frac{p+1}{2} - k\right)\cos(k\theta), \qquad p \geq 0, \tag{5.32}$$

$$g_p(\theta) := \sum_{k\in\mathbb{Z}} -\dot{\phi}_{[p]}\left(\frac{p+1}{2} - k\right) \mathrm{e}^{\mathrm{i}k\theta} = -2\sum_{k=1}^{\lfloor p/2\rfloor} \dot{\phi}_{[p]}\left(\frac{p+1}{2} - k\right)\sin(k\theta), \qquad p \geq 2, \tag{5.33}$$

$$f_p(\theta) := \sum_{k\in\mathbb{Z}} -\ddot{\phi}_{[p]}\left(\frac{p+1}{2} - k\right) \mathrm{e}^{\mathrm{i}k\theta} = -\ddot{\phi}_{[p]}\left(\frac{p+1}{2}\right) - 2\sum_{k=1}^{\lfloor p/2\rfloor} \ddot{\phi}_{[p]}\left(\frac{p+1}{2} - k\right)\cos(k\theta), \qquad p \geq 2, \tag{5.34}$$

with the usual assumption that an empty sum is zero.[1] Using (4.13)–(4.14) and (4.21), it can be easily checked that

$$h_0(\theta) = h_1(\theta) = 1, \qquad g_2(\theta) = g_3(\theta) = -\sin\theta, \qquad f_2(\theta) = f_3(\theta) = 2 - 2\cos\theta. \tag{5.35}$$

Remark 5.1. The functions (5.32) and (5.34) have already been analyzed in Chapter 4 for odd degrees $p = 2q+1$, $q \geq 1$. Indeed, f_q (resp. h_q) in Chapter 4 coincides with f_{2q+1} (resp. h_{2q+1}) here.

We conclude this subsection by giving a formal definition of what we call 'central rows' of $K_n^{[p]}$, $H_n^{[p]}$, $M_n^{[p]}$. They are defined as the rows of $K_n^{[p]}$, $H_n^{[p]}$, $M_n^{[p]}$ corresponding to indices $i \in \{1, \dots, n+p-2\}$ satisfying the following conditions:

$$(K_n^{[p]})_{ij} = -\ddot{\phi}_{[p]}\left(\frac{p+1}{2} + i - j\right), \qquad j = 1, \dots, n+p-2, \tag{5.36}$$

$$(H_n^{[p]})_{ij} = \dot{\phi}_{[p]}\left(\frac{p+1}{2} + i - j\right), \qquad j = 1, \dots, n+p-2, \tag{5.37}$$

$$(M_n^{[p]})_{ij} = \phi_{[p]}\left(\frac{p+1}{2} + i - j\right), \qquad j = 1, \dots, n+p-2. \tag{5.38}$$

The central rows of $K_n^{[p]}$, $H_n^{[p]}$, $M_n^{[p]}$ coincide with the corresponding rows of $T_{n+p-2}(f_p)$, $\mathrm{i}T_{n+p-2}(g_p)$, $T_{n+p-2}(h_p)$, respectively. Using the properties $\mathrm{supp}(N_{2,[p]}) \subseteq \cdots \subseteq \mathrm{supp}(N_{p,[p]}) = [t_p, t_{2p+1}] = [0, \frac{p}{n}]$ and $[1 - \frac{p}{n}, 1] = [t_{n+1}, t_{n+p+2}] = \mathrm{supp}(N_{n+1,[p]}) \supseteq \cdots \supseteq \mathrm{supp}(N_{n+p-1,[p]})$, the fact that $\xi_{i+1,[p]} = \frac{i+1}{n} - \frac{p+1}{2n}$ for $i = p, \dots, n-1$, and the equality $\mathrm{supp}(\phi_{[p]}) = [0, p+1]$, it can be shown that every $i \in \{\lfloor 3p/2\rfloor, \dots, n+p-1-\lfloor 3p/2\rfloor\}$ satisfies (5.36)–(5.38). Consequently, a condition to ensure that $K_n^{[p]}$, $H_n^{[p]}$, $M_n^{[p]}$ have at least one central row is $n+p-1-\lfloor 3p/2\rfloor \geq \lfloor 3p/2\rfloor$, i.e., $n \geq p^*$, where

$$p^* := \begin{cases} 2p & \text{if } p \text{ is odd,} \\ 2p+1 & \text{if } p \text{ is even.} \end{cases}$$

5.3 Properties of $f_p(\theta)$, $g_p(\theta)$, $h_p(\theta)$

In this section we provide some properties of the functions $f_p(\theta)$, $g_p(\theta)$, $h_p(\theta)$ defined in (5.32)–(5.34). They extend to any degree p the results obtained in Chapter 4 for odd degree $p = 2q+1$ (cf. Remark 5.1). We shall see later that these functions are involved in the expression of the spectral symbol characterizing the asymptotic spectral distribution of the B-spline IgA collocation matrices. The next lemma gives an alternative expression for h_p, g_p and f_p.

[1] An empty sum is a sum where the upper index is less than the lower one, such as $\sum_{k=1}^{0} k^2$.

Lemma 5.1. *Let $p \geq 2$, and let h_p, g_p and f_p be the functions defined in* (5.32)–(5.34)*. Then the following properties hold.*

a) $\forall \theta \in [-\pi, \pi]$,

$$h_p(\theta) = \sum_{k\in\mathbb{Z}} \left(\widehat{\phi^*_{[0]}}(\theta + 2k\pi)\right)^{p+1} = \sum_{k\in\mathbb{Z}} \frac{(2\sin(\theta/2 + k\pi))^{p+1}}{(\theta + 2k\pi)^{p+1}}. \tag{5.39}$$

b) $\forall \theta \in [-\pi, \pi]$,

$$g_p(\theta) = -\sum_{k\in\mathbb{Z}} \frac{(2\sin(\theta/2 + k\pi))^{p+1}}{(\theta + 2k\pi)^{p}}. \tag{5.40}$$

c) $\forall \theta \in [-\pi, \pi]$,

$$f_p(\theta) = (2 - 2\cos\theta) h_{p-2}(\theta) \tag{5.41}$$

and, for $p \geq 4$,

$$f_p(\theta) = \sum_{k\in\mathbb{Z}} \frac{(2\sin(\theta/2 + k\pi))^{p+1}}{(\theta + 2k\pi)^{p-1}}. \tag{5.42}$$

Proof. We first recall the Parseval identity for Fourier transforms, i.e.,

$$\int_{\mathbb{R}} \varphi(t)\overline{\psi(t)}\, dt = \frac{1}{2\pi} \int_{\mathbb{R}} \widehat{\varphi}(\theta)\overline{\widehat{\psi}(\theta)}\, d\theta, \qquad \varphi,\, \psi \in L^2(\mathbb{R}), \tag{5.43}$$

and the translation property of the Fourier transform, i.e.,

$$\widehat{\psi(\cdot + x)}(\theta) = \widehat{\psi}(\theta)\, \mathrm{e}^{\mathrm{i}x\theta}, \qquad \psi \in L^1(\mathbb{R}),\ x \in \mathbb{R}. \tag{5.44}$$

We differentiate the cases of odd and even degree p. We start with proving the relation (5.39) for $p = 2q$. From Lemma 4.2, see in particular (4.24), we know that, for all $k \in \mathbb{Z}$,

$$\phi_{[p]}\left(\frac{p+1}{2} - k\right) = \phi_{[2q]}\left(q + \frac{1}{2} - k\right) = \int_{\mathbb{R}} \phi_{[q]}(t)\phi_{[q-1]}\left(t + k - \frac{1}{2}\right) dt. \tag{5.45}$$

In view of (5.43)–(5.44) and (4.32), for any $k \in \mathbb{Z}$ the expression in (5.45) is equal to

$$\begin{aligned}
\int_{\mathbb{R}} \phi_{[q]}(t)\phi_{[q-1]}\left(t + k - \frac{1}{2}\right) dt &= \frac{1}{2\pi} \int_{\mathbb{R}} \widehat{\phi_{[q]}}(\theta)\, \overline{\widehat{\phi_{[q-1]}}(\theta)}\, \mathrm{e}^{-\mathrm{i}(k-1/2)\theta}\, d\theta = \frac{1}{2\pi} \int_{\mathbb{R}} \left|\widehat{\phi_{[q-1]}}(\theta)\right|^2 \widehat{\phi_{[0]}}(\theta)\, \mathrm{e}^{-\mathrm{i}(k-1/2)\theta}\, d\theta \\
&= \frac{1}{2\pi} \sum_{l\in\mathbb{Z}} \int_{-\pi}^{\pi} \left|\widehat{\phi_{[q-1]}}(\theta + 2l\pi)\right|^2 \widehat{\phi_{[0]}}(\theta + 2l\pi)\, (-1)^l \mathrm{e}^{-\mathrm{i}(k-1/2)\theta}\, d\theta \\
&= \frac{1}{2\pi} \int_{-\pi}^{\pi} \left[\sum_{l\in\mathbb{Z}} \left|\widehat{\phi_{[q-1]}}(\theta + 2l\pi)\right|^2 \widehat{\phi_{[0]}}(\theta + 2l\pi)\, (-1)^l \mathrm{e}^{\mathrm{i}\theta/2}\right] \mathrm{e}^{-\mathrm{i}k\theta}\, d\theta.
\end{aligned}$$

Note that the last equality follows from the uniform convergence of the series. Indeed, since $q \geq 1$, from (4.33) we obtain that, for all $\theta \in [-\pi, \pi]$,

$$\left|\left|\widehat{\phi_{[q-1]}}(\theta + 2l\pi)\right|^2 \widehat{\phi_{[0]}}(\theta + 2l\pi)(-1)^l \mathrm{e}^{\mathrm{i}\theta/2} \mathrm{e}^{-\mathrm{i}k\theta}\right| = \left(\frac{2 - 2\cos\theta}{(\theta + 2l\pi)^2}\right)^{q+1/2} \leq \begin{cases} 1 & \text{if } l = 0 \\ \dfrac{4^{q+1/2}}{(2|l|\pi - \pi)^{2q+1}} & \text{if } l \neq 0 \end{cases}$$

We conclude that the values (5.45), i.e. the Fourier coefficients of h_{2q} in (5.32), are also the Fourier coefficients of the function

$$\sum_{l\in\mathbb{Z}} \left|\widehat{\phi_{[q-1]}}(\theta + 2l\pi)\right|^2 \widehat{\phi_{[0]}}(\theta + 2l\pi)\, (-1)^l \mathrm{e}^{\mathrm{i}\theta/2}, \tag{5.46}$$

which therefore coincides with h_{2q}. Moreover, by using (4.32), (4.34) we get

$$\widehat{\phi_{[0]}}(\theta + 2l\pi)(-1)^l \mathrm{e}^{\mathrm{i}\theta/2} = \frac{1 - \mathrm{e}^{-\mathrm{i}\,(\theta+2l\pi)}}{\mathrm{i}\,(\theta + 2l\pi)} \mathrm{e}^{\mathrm{i}(\theta+2l\pi)/2} = \frac{\sin(\theta/2 + l\pi)}{\theta/2 + l\pi} = \widehat{\phi^*_{[0]}}(\theta + 2l\pi),$$

and it follows that h_{2q} is given by (5.39) for $q \geq 1$.

To prove the expression (5.40) of g_p for $p = 2q$, we follow an argument similar to the one in the proof of (5.39). By Lemma 4.2, for all $k \in \mathbb{Z}$ we have

$$-\dot{\phi}_{[p]}\left(\frac{p+1}{2} - k\right)\frac{1}{\mathrm{i}} = -\dot{\phi}_{[2q]}\left(q + \frac{1}{2} - k\right)\frac{1}{\mathrm{i}} = \frac{1}{\mathrm{i}}\int_{\mathbb{R}} \phi_{[q]}(t)\dot{\phi}_{[q-1]}\left(t + k - \frac{1}{2}\right)dt. \tag{5.47}$$

In view of (5.43)–(5.44) and (4.32), (4.35), for any $k \in \mathbb{Z}$ the expression in (5.47) is equal to

$$\begin{aligned}
\frac{1}{\mathrm{i}}\int_{\mathbb{R}} \phi_{[q]}(t)\dot{\phi}_{[q-1]}\left(t + k - \frac{1}{2}\right)dt &= \frac{1}{2\mathrm{i}\pi}\int_{\mathbb{R}} \widehat{\phi_{[q]}}(\theta)\,\overline{\widehat{\dot{\phi}_{[q-1]}}(\theta)}\,\mathrm{e}^{-\mathrm{i}(k-1/2)\theta}\,d\theta = -\frac{1}{2\pi}\int_{\mathbb{R}} \left|\widehat{\phi_{[q-1]}}(\theta)\right|^2 2\sin(\theta/2)\mathrm{e}^{-\mathrm{i}k\theta}\,d\theta \\
&= -\frac{1}{2\pi}\int_{-\pi}^{\pi}\left[\sum_{l\in\mathbb{Z}} \left|\widehat{\phi_{[q-1]}}(\theta + 2l\pi)\right|^2 2\sin(\theta/2 + l\pi)\right]\mathrm{e}^{-\mathrm{i}k\theta}\,d\theta \\
&= \frac{1}{2\pi}\int_{-\pi}^{\pi}\left[-\sum_{l\in\mathbb{Z}} \frac{2(\sin(\theta/2 + l\pi))^{p+1}}{(\theta/2 + l\pi)^p}\right]\mathrm{e}^{-\mathrm{i}k\theta}\,d\theta.
\end{aligned}$$

We conclude that the values (5.47), i.e. the Fourier coefficients of g_{2q} in (5.33), are also the Fourier coefficients of the function

$$-\sum_{l\in\mathbb{Z}} \frac{2(\sin(\theta/2 + l\pi))^{p+1}}{(\theta/2 + l\pi)^p} = -\sum_{l\in\mathbb{Z}} \frac{(2\sin(\theta/2 + l\pi))^{p+1}}{(\theta + 2l\pi)^p}.$$

To prove the expression (5.41) of f_p for $p = 2q$, we follow again a similar argument as the one to prove (5.39). By Lemma 4.2, for all $k \in \mathbb{Z}$ we have

$$-\ddot{\phi}_{[p]}\left(\frac{p+1}{2} - k\right) = -\ddot{\phi}_{[2q]}\left(q + \frac{1}{2} - k\right) = \int_{\mathbb{R}} \dot{\phi}_{[q]}(t)\dot{\phi}_{[q-1]}\left(t + k - \frac{1}{2}\right)dt. \tag{5.48}$$

In view of (5.43)–(5.44) and (4.32), (4.35), for $q \geq 2$ and for any $k \in \mathbb{Z}$, the expression in (5.48) is equal to

$$\begin{aligned}
\int_{\mathbb{R}} \dot{\phi}_{[q]}(t)\dot{\phi}_{[q-1]}\left(t + k - \frac{1}{2}\right)dt &= \frac{1}{2\pi}\int_{\mathbb{R}} \widehat{\dot{\phi}_{[q]}}(\theta)\,\overline{\widehat{\dot{\phi}_{[q-1]}}(\theta)}\,\mathrm{e}^{-\mathrm{i}(k-1/2)\theta}\,d\theta \\
&= \frac{1}{2\pi}\int_{\mathbb{R}} \left|\widehat{\phi_{[q-2]}}(\theta)\right|^2 \widehat{\phi_{[0]}}(\theta)\,(2 - 2\cos\theta)\mathrm{e}^{-\mathrm{i}(k-1/2)\theta}\,d\theta \\
&= \frac{1}{2\pi}\int_{-\pi}^{\pi}\left[\sum_{l\in\mathbb{Z}} \left|\widehat{\phi_{[q-2]}}(\theta + 2l\pi)\right|^2 \widehat{\phi_{[0]}}(\theta + 2l\pi)\,(-1)^l(2 - 2\cos\theta)\mathrm{e}^{\mathrm{i}\theta/2}\right]\mathrm{e}^{-\mathrm{i}k\theta}\,d\theta.
\end{aligned}$$

We conclude that the values (5.48), i.e. the Fourier coefficients of f_{2q} in (5.34), are also the Fourier coefficients of the function

$$\sum_{l\in\mathbb{Z}} \left|\widehat{\phi_{[q-2]}}(\theta + 2l\pi)\right|^2 \widehat{\phi_{[0]}}(\theta + 2l\pi)\,(-1)^l(2 - 2\cos\theta)\mathrm{e}^{\mathrm{i}\theta/2}.$$

Hence, recalling that (5.46) is an alternative expression for $h_{2q}(\theta)$, we obtain that $f_{2q}(\theta) = (2 - 2\cos\theta)h_{2(q-1)}(\theta)$ for $q \geq 2$. From (5.35) we see that the equality (5.41) holds for $q = 1$ as well. Moreover, (5.42) immediately follows from (5.39) and (5.41).

For odd degree $p = 2q+1$, keeping in mind Remark 5.1, we recall from Lemma 4.4 that

$$h_{2q+1}(\theta) = \sum_{k\in\mathbb{Z}} \left|\widehat{\phi_{[q]}}(\theta + 2k\pi)\right|^2, \quad q \geq 1.$$

In view of (4.33), (4.34), we immediately obtain the relation (5.39). The equality (5.41) follows from Lemma 4.5 for $q \geq 2$ and from (5.35) for $q = 1$. Moreover, (5.42) is obtained by combining (5.39) and (5.41). The expression of g_{2q+1} can be derived by applying the same arguments as in the case of even degree. □

To establish lower and upper bounds for h_p, g_p and f_p we need the following technical lemma.

Lemma 5.2. *Let $p \geq 2$, and let us consider the functions*

$$r_p(\theta) := \sum_{k\neq 0} \frac{(-1)^{k(p+1)}}{(\theta + 2k\pi)^{p+1}}, \quad \tilde{r}_p(\theta) := -\sum_{k\neq 0} \frac{(-1)^{k(p+1)}}{(\theta + 2k\pi)^{p}}, \quad \theta \in [-\pi, \pi]. \tag{5.49}$$

Then, r_p and $\tilde{r}_p$ are continuous functions over $[-\pi, \pi]$, and

$$0 < r_p(\theta) \leq r_p(\pi) \leq \left(\frac{\pi^4}{48} - 1\right)\frac{1}{\pi^{p+1}}, \quad \theta \in (0, \pi]; \tag{5.50}$$

$$0 < \tilde{r}_p(\theta) \leq \tilde{r}_p(\pi) = \frac{1}{\pi^p}, \quad \theta \in (0, \pi]. \tag{5.51}$$

Proof. The functions r_p and $\tilde{r}_p$ are continuous over $[-\pi, \pi]$ because the two series in (5.49) converge uniformly. We now derive an upper and lower bound for $r_p(\theta)$, $\theta \in [0, \pi]$. From (5.49) we obtain

$$r_p(\theta) = \sum_{k=1}^{\infty} \left[\frac{(-1)^{k(p+1)}}{(2k\pi + \theta)^{p+1}} + \frac{(-1)^{k(p+1)}}{(-2k\pi + \theta)^{p+1}}\right]. \tag{5.52}$$

We differentiate the cases of odd and even degree. We first focus on the odd case $p = 2q+1$. From (5.52),

$$r_{2q+1}(\theta) = \sum_{k=1}^{\infty} \left[\frac{1}{(2k\pi + \theta)^{2q+2}} + \frac{1}{(2k\pi - \theta)^{2q+2}}\right].$$

It is clear that $r_{2q+1}(\theta) > 0$ for $\theta \in [0, \pi]$. For $\rho > 1$, $k \geq 1$ and $\theta \in [0, \pi]$, one can check that

$$\frac{1}{(2k\pi + \theta)^{\rho}} + \frac{1}{(2k\pi - \theta)^{\rho}} \leq \frac{1}{(2k\pi + \pi)^{\rho}} + \frac{1}{(2k\pi - \pi)^{\rho}},$$

and then we obtain, for $q \geq 1$,

$$r_{2q+1}(\theta) \leq \sum_{k=1}^{\infty} \left[\frac{1}{(2k\pi + \pi)^{2q+2}} + \frac{1}{(2k\pi - \pi)^{2q+2}}\right] \leq \frac{1}{\pi^{2q+2}} \sum_{k=1}^{\infty} \left[\frac{1}{(2k+1)^4} + \frac{1}{(2k-1)^4}\right] = \frac{1}{\pi^{2q+2}}\left(\frac{\pi^4}{48} - 1\right).$$

We follow a similar argument for the even case $p = 2q$. In this case, from (5.52) we have

$$r_{2q}(\theta) = \sum_{k=1}^{\infty} (-1)^k \left[\frac{1}{(2k\pi + \theta)^{2q+1}} - \frac{1}{(2k\pi - \theta)^{2q+1}}\right] = \sum_{l=1}^{\infty} \left[\frac{1}{(4l\pi + \theta)^{2q+1}} - \frac{1}{(4l\pi - \theta)^{2q+1}} - \frac{1}{((4l-2)\pi + \theta)^{2q+1}} + \frac{1}{((4l-2)\pi - \theta)^{2q+1}}\right].$$

Let us define

$$s_\rho(\theta) := \frac{1}{(b+\theta)^\rho} - \frac{1}{(b-\theta)^\rho} + \frac{1}{(a-\theta)^\rho} - \frac{1}{(a+\theta)^\rho}.$$

If $\rho > 1$ and $0 \le \theta \le \pi < a < b$, then we have

$$s'_\rho(\theta) = -\frac{\rho}{(b+\theta)^{\rho+1}} - \frac{\rho}{(b-\theta)^{\rho+1}} + \frac{\rho}{(a-\theta)^{\rho+1}} + \frac{\rho}{(a+\theta)^{\rho+1}} > 0,$$

and thus s_ρ is a strictly increasing function, which implies that $s_\rho(\pi) \ge s_\rho(\theta) > s_\rho(0) = 0$ for $\theta \in (0, \pi]$. As a consequence, we have $r_{2q}(\pi) \ge r_{2q}(\theta) > 0$ for $\theta \in (0, \pi]$. Moreover, for $q \ge 1$ and $\theta \in [0, \pi]$,

$$r_{2q}(\theta) \le \sum_{l=1}^{\infty} \frac{1}{((4l-2)\pi - \theta)^{2q+1}} \le \sum_{l=1}^{\infty} \frac{1}{((4l-2)\pi - \pi)^{2q+1}} \le \frac{1}{\pi^{2q+1}} \sum_{l=1}^{\infty} \frac{1}{(4l-3)^3},$$

where

$$\sum_{l=1}^{\infty} \frac{1}{(4l-3)^3} < 1.02 < \frac{\pi^4}{48} - 1.$$

Hence, both in the odd and even case we obtain the bounds in (5.50).

We now derive an upper and lower bound for $\tilde{r}_p(\theta)$, $\theta \in [0, \pi]$. From (5.49) we have

$$\tilde{r}_p(\theta) = -\sum_{k=1}^{\infty} \left[\frac{(-1)^{k(p+1)}}{(2k\pi + \theta)^p} + \frac{(-1)^{k(p+1)}}{(-2k\pi + \theta)^p} \right].$$

We differentiate the cases of odd and even degree. We first focus on the odd case $p = 2q + 1$. Note that

$$\tilde{r}_{2q+1}(\theta) = \sum_{k=1}^{\infty} \left[\frac{1}{(2k\pi - \theta)^{2q+1}} - \frac{1}{(2k\pi + \theta)^{2q+1}} \right].$$

The function

$$\frac{1}{(a-\theta)^\rho} - \frac{1}{(a+\theta)^\rho}, \quad 0 \le \theta \le \pi < a,\ \rho > 1,$$

is nonnegative and increasing. Then, for all $\theta \in (0, \pi]$ we have

$$0 < \tilde{r}_{2q+1}(\theta) \le \tilde{r}_{2q+1}(\pi) = \frac{1}{\pi^{2q+1}} \sum_{k=1}^{\infty} \left[\frac{1}{(2k-1)^{2q+1}} - \frac{1}{(2k+1)^{2q+1}} \right] = \frac{1}{\pi^{2q+1}},$$

which immediately gives (5.51).

Let us now consider the case $p = 2q$. We have

$$\tilde{r}_{2q}(\theta) = \sum_{k=1}^{\infty} \left[\frac{1}{((4k-2)\pi + \theta)^{2q}} - \frac{1}{(4k\pi + \theta)^{2q}} + \frac{1}{((4k-2)\pi - \theta)^{2q}} - \frac{1}{(4k\pi - \theta)^{2q}} \right].$$

The function

$$\tilde{s}_\rho(\theta) := \frac{1}{(a+\theta)^\rho} - \frac{1}{(b+\theta)^\rho} + \frac{1}{(a-\theta)^\rho} - \frac{1}{(b-\theta)^\rho}, \quad 0 \le \theta \le \pi < a < b,\ \rho > 1,$$

is positive, and $\tilde{s}'_\rho(\theta) = \rho s_{\rho+1}(\theta) > \tilde{s}'_\rho(0) = 0$ for $\theta \in (0, \pi]$. Therefore, $\tilde{s}_\rho$ is increasing in $[0, \pi]$. As a consequence, for $\theta \in (0, \pi]$,

$$\begin{aligned} 0 < \tilde{r}_{2q}(\theta) \le \tilde{r}_{2q}(\pi) &= \frac{1}{\pi^{2q}} \sum_{k=1}^{\infty} \left[\frac{1}{(4k-1)^{2q}} - \frac{1}{(4k+1)^{2q}} + \frac{1}{(4k-3)^{2q}} - \frac{1}{(4k-1)^{2q}} \right] \\ &= \frac{1}{\pi^{2q}} \sum_{k=1}^{\infty} \left[\frac{1}{(4k-3)^{2q}} - \frac{1}{(4k+1)^{2q}} \right] = \frac{1}{\pi^{2q}}. \end{aligned}$$

Thus we obtain (5.51) for the even case as well. □

We now provide lower and upper bounds for h_p.

Lemma 5.3. *Let $p \geq 2$, and let h_p be the function defined in* (5.32)*. Then the following properties hold.*

a) $\forall \theta \in [-\pi, \pi]$,

$$L_p(\theta) \leq h_p(\theta) \leq \min(1, U_p(\theta)), \tag{5.53}$$

where

$$L_p(\theta) := \left(\frac{2 - 2\cos\theta}{\theta^2}\right)^{\frac{p+1}{2}}, \tag{5.54}$$

$$U_p(\theta) := \left(\frac{2 - 2\cos\theta}{\theta^2}\right)^{\frac{p+1}{2}} + \left(\frac{\pi^4}{48} - 1\right)\left(\frac{2 - 2\cos\theta}{\pi^2}\right)^{\frac{p+1}{2}}. \tag{5.55}$$

b) $\max\limits_{\theta\in[-\pi,\pi]} h_p(\theta) = h_p(0) = 1.$

c) Let $m_{h_p} := \min\limits_{\theta\in[-\pi,\pi]} h_p(\theta)$*, then*

$$m_{h_p} \geq \left(\frac{2}{\pi}\right)^{p+1} > 0. \tag{5.56}$$

d) We have

$$h_p(\pi) \leq \frac{h_p(\pi)}{h_p(\frac{\pi}{2})} \leq 2^{\frac{1-p}{2}}. \tag{5.57}$$

In particular, the value $h_p(\pi)$ converges to 0 exponentially as $p \to \infty$.

Proof. First of all, we remark that h_p, L_p and U_p are symmetric around $\theta = 0$. Hence, it is sufficient to prove the various statements of the lemma for $\theta \in [0, \pi]$. We also recall that

$$\frac{\sin(\theta/2)}{\theta/2} = \left(\frac{2 - 2\cos\theta}{\theta^2}\right)^{1/2}, \quad \theta \in [-\pi, \pi].$$

Let us consider the first statement of the lemma. From (5.39) we obtain

$$h_p(\theta) = (\sin(\theta/2))^{p+1} \sum_{k\in\mathbb{Z}} \frac{(-1)^{k(p+1)}}{(\theta/2 + k\pi)^{p+1}} = \left(\frac{\sin(\theta/2)}{\theta/2}\right)^{p+1} + (2\sin(\theta/2))^{p+1}\, r_p(\theta),$$

where r_p is defined in (5.49). Hence, from (5.50) we get

$$L_p(\theta) \leq h_p(\theta) \leq U_p(\theta).$$

We now focus on the second statement of the lemma. By using the positivity (4.16), the local support property (4.17) and the partition of unity property (4.19) of cardinal B-splines, from (5.32) we obtain

$$h_p(\theta) = \sum_{k\in\mathbb{Z}} \phi_{[p]}\left(\frac{p+1}{2} - k\right) e^{ik\theta} \leq \sum_{k\in\mathbb{Z}} \phi_{[p]}\left(\frac{p+1}{2} - k\right) |e^{ik\theta}| = 1.$$

In addition, it can be easily checked that $h_p(0) = 1$. This also completes the proof of the upper bound in (5.53).

To address the lower bound (5.56), we observe that $(2-2\cos\theta)/\theta^2$ is monotonically decreasing in $[0,\pi]$. As a consequence,

$$L_p(\theta) \geq \left(\frac{2}{\pi}\right)^{p+1} > 0, \qquad \theta \in [-\pi,\pi]. \tag{5.58}$$

Finally, we focus on (5.57). Since $h_p(\theta) \leq 1$, it is sufficient to prove the second inequality in (5.57). From (5.39) we have

$$h_p\left(\frac{\pi}{2}\right) = \frac{2^{\frac{3(p+1)}{2}}}{\pi^{p+1}} \sum_{k\in\mathbb{Z}} \frac{(-1)^{k(p+1)}}{(4k+1)^{p+1}}, \quad h_p(\pi) = \frac{2^{p+1}}{\pi^{p+1}} \sum_{k\in\mathbb{Z}} \frac{(-1)^{k(p+1)}}{(2k+1)^{p+1}}.$$

We differentiate the case of even and odd degree. We start with the even case $p = 2q$. Then,

$$h_{2q}\left(\frac{\pi}{2}\right) = \frac{2^{\frac{3(2q+1)}{2}}}{\pi^{2q+1}} \sum_{k\in\mathbb{Z}} \frac{(-1)^{k}}{(4k+1)^{2q+1}}, \quad h_{2q}(\pi) = \frac{2^{2q+1}}{\pi^{2q+1}} \sum_{k\in\mathbb{Z}} \frac{(-1)^{k}}{(2k+1)^{2q+1}}.$$

By splitting the latter sum into a sum over the even integers and a sum over the odd integers, we get

$$\begin{aligned}\sum_{k\in\mathbb{Z}} \frac{(-1)^k}{(2k+1)^{2q+1}} &= \sum_{l\in\mathbb{Z}} \frac{1}{(4l+1)^{2q+1}} - \sum_{l\in\mathbb{Z}} \frac{1}{(4l+3)^{2q+1}} = \sum_{l\in\mathbb{Z}} \frac{1}{(4l+1)^{2q+1}} + \sum_{m\in\mathbb{Z}} \frac{1}{(4m+1)^{2q+1}} = \sum_{l\in\mathbb{Z}} \frac{2}{(4l+1)^{2q+1}} \\ &= 2(a_{2q}+b_{2q}),\end{aligned}$$

where

$$a_{2q} := \sum_{l\in\mathbb{Z}} \frac{1}{(8l+1)^{2q+1}}, \quad b_{2q} := \sum_{l\in\mathbb{Z}} \frac{1}{(8l+5)^{2q+1}}.$$

Hence,

$$h_{2q}\left(\frac{\pi}{2}\right) = \frac{2^{\frac{3(2q+1)}{2}}}{\pi^{2q+1}}(a_{2q}-b_{2q}), \quad h_{2q}(\pi) = \frac{2^{2q+1}}{\pi^{2q+1}} 2(a_{2q}+b_{2q}).$$

It is easy to see that $b_{2q} < 0$. In addition, from (5.56) we know that $h_p(\theta) > 0$, so that $a_{2q} + b_{2q} > 0$, $a_{2q} - b_{2q} > 0$. Therefore, we obtain

$$\frac{h_{2q}(\pi)}{h_{2q}(\frac{\pi}{2})} = \frac{2^{2q+2}(a_{2q}+b_{2q})}{2^{\frac{3(2q+1)}{2}}(a_{2q}-b_{2q})} \leq 2^{\frac{1-2q}{2}} = 2^{\frac{1-p}{2}}.$$

For odd degree $p = 2q+1$, by using a completely similar manipulation (or by applying Lemma 4.4 with a look at Remark 5.1) we obtain the exact equality

$$\frac{h_{2q+1}(\pi)}{h_{2q+1}(\frac{\pi}{2})} = 2^{\frac{1-p}{2}},$$

and it follows that (5.57) holds even in this case. □

The next lemma is devoted to lower and upper bounds for g_p.

Lemma 5.4. *Let $p \geq 2$, and let g_p be the function defined in* (5.33)*. Then the following properties hold.*

a) $\forall \theta \in [-\pi,\pi]$,

$$|2\sin(\theta/2)|^{p+1}\left(\frac{1}{|\theta|^p} - \frac{1}{\pi^p}\right) \leq |g_p(\theta)| \leq |2\sin(\theta/2)|^{p+1}\frac{1}{|\theta|^p}. \tag{5.59}$$

b) The zeros of g_p are given by

$$g_p(-\pi) = g_p(0) = g_p(\pi) = 0. \tag{5.60}$$

Proof. We first remark that from (5.33) it follows that g_p is antisymmetric around $\theta = 0$. Hence, it is sufficient to study it on the interval $[0, \pi]$. From (5.40) and (5.49) we have

$$g_p(\theta) = -\sum_{k\in\mathbb{Z}} \frac{(2\sin(\theta/2 + k\pi))^{p+1}}{(\theta + 2k\pi)^p} = -(2\sin(\theta/2))^{p+1}\left[\frac{1}{\theta^p} - \tilde{r}_p(\theta)\right]. \tag{5.61}$$

Then, (5.51) immediately gives (5.59) and (5.60). □

In the following lemma we provide lower and upper bounds for f_p.

Lemma 5.5. *Let $p \geq 2$, and let f_p be the function defined in* (5.34)*. Then the following properties hold.*

a) $\forall \theta \in [-\pi, \pi]$,

$$f_p(\theta) = 2 - 2\cos\theta, \quad p = 2, 3, \tag{5.62}$$

and

$$(2 - 2\cos\theta)L_{p-2}(\theta) \leq f_p(\theta) \leq (2 - 2\cos\theta)\min(1, U_{p-2}(\theta)), \quad p \geq 4, \tag{5.63}$$

where L_p and U_p are defined in (5.54) *and* (5.55) *respectively.*

b) $\min_{\theta\in[-\pi,\pi]} f_p(\theta) = f_p(0) = 0$*, and $\theta = 0$ is the unique zero of f_p in $[-\pi, \pi]$; the order of this zero is two.*

c) let $M_{f_p} := \max_{\theta\in[-\pi,\pi]} f_p(\theta)$*, then*

$$M_{f_p} \leq \min\left(4, \frac{17}{p+1} + \left(\frac{\pi^4}{12} - 4\right)\left(\frac{2}{\pi}\right)^{p-1}\right). \tag{5.64}$$

In particular, $M_{f_p} \to 0$ as $p \to \infty$.

d) We have

$$\frac{f_p(\pi)}{M_{f_p}} \leq \frac{f_p(\pi)}{f_p(\frac{\pi}{2})} \leq 2^{\frac{5-p}{2}}. \tag{5.65}$$

In particular, the ratio $f_p(\pi)/M_{f_p}$ converges to 0 exponentially as $p \to \infty$.

Proof. The first statement of the lemma immediately follows from (5.35), (5.41) and (5.53).

The relations (5.62)–(5.63) and the lower bound (5.58) imply that $f_p(\theta) \geq 0$ in $[-\pi, \pi]$ and that it has a unique zero at $\theta = 0$ in $[-\pi, \pi]$. Moreover, from (5.41) we obtain

$$\begin{aligned} f_p'(\theta) &= 2(\sin\theta)h_{p-2}(\theta) + (2 - 2\cos\theta)h_{p-2}'(\theta), \\ f_p''(\theta) &= 2(\cos\theta)h_{p-2}(\theta) + 4(\sin\theta)h_{p-2}'(\theta) + (2 - 2\cos\theta)h_{p-2}''(\theta). \end{aligned}$$

By using the equality $h_p(0) = 1$ (see (5.35) and Lemma 5.3), we get $f_p'(0) = 0$ and $f_p''(0) = 2$. This proves that f_p has a zero of order two at $\theta = 0$ and completes the proof of the second statement of the lemma.

From (5.62)–(5.63) it is also easy to see that $M_{f_p} \leq 4$. Now we derive the upper bound (5.64) for M_{f_p} in the third statement of the lemma. To this end, we use the inequalities

$$2 - 2\cos\theta \leq \theta^2 - \frac{\theta^4}{18} \leq \theta^2, \quad \forall \theta \in [-\pi, \pi].$$

It follows

$$(2 - 2\cos\theta)\left(\frac{2 - 2\cos\theta}{\theta^2}\right)^{\frac{p-1}{2}} \leq \theta^2\left(1 - \frac{\theta^2}{18}\right)^{\frac{p-1}{2}}, \quad \forall \theta \in [-\pi, \pi].$$

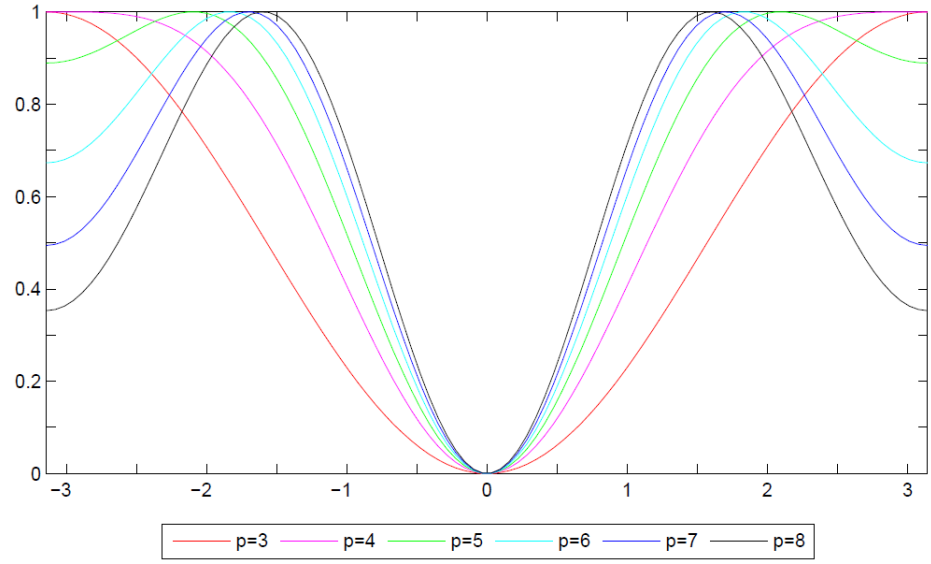


Figure 5.1: graph of f_p/M_{f_p} for $p = 3, \dots, 8$.

p	2	3	4	5	6	7	8	9	10	11	12	13	14
$f_p(\pi)/M_{f_p}$	1.000	1.000	1.000	0.889	0.673	0.494	0.353	0.249	0.174	0.121	0.083	0.057	0.039

Table 5.1: values of the ratio $f_p(\pi)/M_{f_p}$ for $p = 2, \dots, 14$.

Let ρ be a positive real number. If $\frac{18}{\rho+1} \le \pi^2$, then the maximum of $\theta^2\left(1-\frac{\theta^2}{18}\right)^\rho$ over $[-\pi,\pi]$ is located at $\theta^2 = \frac{18}{\rho+1}$ and its value is given by

$$\frac{18}{\rho+1}\left(1-\frac{1}{\rho+1}\right)^\rho.$$

Therefore, when $p \ge 3$, we have

$$(2-2\cos\theta)\left(\frac{2-2\cos\theta}{\theta^2}\right)^{\frac{p-1}{2}} \le \frac{36}{p+1}\left(1-\frac{2}{p+1}\right)^{\frac{p-1}{2}}, \quad \forall\theta\in[-\pi,\pi]. \tag{5.66}$$

Moreover, $\forall\theta \in [-\pi,\pi]$,

$$(2-2\cos\theta)\left(\frac{\pi^4}{48}-1\right)\left(\frac{2-2\cos\theta}{\pi^2}\right)^{\frac{p-1}{2}} \le 4\left(\frac{\pi^4}{48}-1\right)\left(\frac{2}{\pi}\right)^{p-1}. \tag{5.67}$$

From (5.55) and (5.63), the inequalities (5.66)–(5.67) imply that, for $p \ge 4$,

$$M_{f_p} \le \frac{36}{p+1}\left(1-\frac{2}{p+1}\right)^{\frac{p-1}{2}} + \left(\frac{\pi^4}{12}-4\right)\left(\frac{2}{\pi}\right)^{p-1} \le \frac{17}{p+1} + \left(\frac{\pi^4}{12}-4\right)\left(\frac{2}{\pi}\right)^{p-1}. \tag{5.68}$$

In addition, (5.68) holds for $p = 2$ and $p = 3$ too, because we see from (5.62) that $M_{f_2} = M_{f_3} = 4$.

To conclude the proof, we notice that the inequalities in (5.57) are satisfied also for $p = 0, 1$ (see (5.35)). The inequalities in (5.65) follow from (5.57) taking into account that $f_p(\theta) = (2-2\cos\theta)h_{p-2}(\theta)$. □

Figure 5.1 shows the graph of f_p normalized by its maximum M_{f_p} for $p = 3, \dots, 8$. We see from the figure and from Table 5.1 that the ratio $f_p(\pi)/M_{f_p}$ decreases exponentially to zero as $p \to \infty$, in accordance with the last statement in Lemma 5.5. From a numerical viewpoint, we can say that, for large p, the function f_p/M_{f_p} possesses two zeros over $[0,\pi]$: one in $\theta = 0$ and the other in $\theta = \pi$.

In the last lemma, we provide an important relation between the functions h_p, g_p and f_p.

Lemma 5.6. *For all $\theta \in [-\pi,\pi]\backslash\{0\}$, we have*

$$f_p(\theta)h_p(\theta) - [g_p(\theta)]^2 > 0. \tag{5.69}$$

Proof. From (5.41), (5.53), (5.54) and (5.59) we have

$$f_p(\theta)h_p(\theta) - [g_p(\theta)]^2 \geq \frac{(2\sin(\theta/2))^{2p+2}}{(\theta^p)^2} - [g_p(\theta)]^2 \geq 0.$$

Moreover, since $g_p(\theta)$ is antisymmetric (see (5.33)) and $\tilde{r}_p(\theta)$ is strictly positive if $\theta \in (0,\pi]$ (see (5.51)), from (5.61) we obtain the complete statement of the lemma. □

5.4 Spectral distribution and spectral symbol of the normalized sequences $\{\frac{1}{n^2}A_{\boldsymbol{n}}^{[\boldsymbol{p}]}\}_n$ and $\{\frac{1}{n^2}A_{\mathbf{G},\boldsymbol{n}}^{[\boldsymbol{p}]}\}_n$

In this section, we assume that $\boldsymbol{n} = \boldsymbol{\nu}n = (\nu_1 n, \ldots, \nu_d n)$, where $\boldsymbol{\nu} \in \mathbb{Q}_+^d$ is fixed and n varies in the set of natural numbers such that $\boldsymbol{n} = \boldsymbol{\nu}n \in \mathbb{N}^d$. In Theorem 5.1 we prove that the sequence of matrices $\{\frac{1}{n^2}A_{\boldsymbol{n}}^{[\boldsymbol{p}]}\}_n$ is distributed, in the sense of the eigenvalues, like the real function $f_{\boldsymbol{p}}^{(\boldsymbol{\nu})} : [0,1]^d \times [-\pi,\pi]^d \to \mathbb{R}$,

$$\begin{aligned} f_{\boldsymbol{p}}^{(\boldsymbol{\nu})}(\mathbf{x},\boldsymbol{\theta}) &:= \sum_{r=1}^{d} \nu_r^2 (\kappa_{rr} \otimes h_{p_1} \otimes \cdots \otimes h_{p_{r-1}} \otimes f_{p_r} \otimes h_{p_{r+1}} \otimes \cdots \otimes h_{p_d})(\mathbf{x},\boldsymbol{\theta}) \\ &+ \sum_{\substack{r,s=1 \\ r<s}}^{d} \nu_r \nu_s ((\kappa_{rs} + \kappa_{sr}) \otimes h_{p_1} \otimes \cdots \otimes h_{p_{r-1}} \otimes g_{p_r} \otimes h_{p_{r+1}} \otimes \cdots \otimes h_{p_{s-1}} \otimes g_{p_s} \otimes h_{p_{s+1}} \otimes \cdots \otimes h_{p_d})(\mathbf{x},\boldsymbol{\theta}), \end{aligned} \tag{5.70}$$

where $\mathbf{x} = (x_1, \ldots, x_d) \in [0,1]^d$ and $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d) \in [-\pi,\pi]^d$. Therefore, $f_{\boldsymbol{p}}^{(\boldsymbol{\nu})}$ is the symbol of the sequence $\{\frac{1}{n^2}A_{\boldsymbol{n}}^{[\boldsymbol{p}]}\}_n$ (compare the expression of the symbol (5.70) with the expression of $A_{\boldsymbol{n}}^{[\boldsymbol{p}]}$ and $A_{\boldsymbol{n},D}^{[\boldsymbol{p}]}$ in (5.16)–(5.17)). We note that $\{\frac{1}{n^2}A_{\boldsymbol{n}}^{[\boldsymbol{p}]}\}_n$ is really a sequence of matrices, due to the assumption $\boldsymbol{n} = \boldsymbol{\nu}n$. This assumption must be kept in mind while reading this section.

Recalling that $A_{\mathbf{G},\boldsymbol{n}}^{[\boldsymbol{p}]}$ coincides with $A_{\boldsymbol{n}}^{[\boldsymbol{p}]}$ given in (5.16)–(5.19), with the only difference that $\kappa_{rs}, \beta_r, \gamma$ are replaced by $\kappa_{\mathbf{G},rs}, \beta_{\mathbf{G},r}, \gamma_{\mathbf{G}}$ (see Section 5.2), from Theorem 5.1 it follows that $\{\frac{1}{n^2}A_{\mathbf{G},\boldsymbol{n}[\boldsymbol{p}]}\}_n \sim_\lambda f_{\mathbf{G},\boldsymbol{p}}^{(\boldsymbol{\nu})}$, where $f_{\mathbf{G},\boldsymbol{p}}^{(\boldsymbol{\nu})} : [0,1]^d \times [-\pi,\pi]^d \to \mathbb{R}$ is given by

$$\begin{aligned} f_{\mathbf{G},\boldsymbol{p}}^{(\boldsymbol{\nu})}(\mathbf{x},\boldsymbol{\theta}) &:= \sum_{r=1}^{d} \nu_r^2 (\kappa_{\mathbf{G},rr} \otimes h_{p_1} \otimes \cdots \otimes h_{p_{r-1}} \otimes f_{p_r} \otimes h_{p_{r+1}} \otimes \cdots \otimes h_{p_d})(\mathbf{x},\boldsymbol{\theta}) \\ &+ \sum_{\substack{r,s=1 \\ r<s}}^{d} \nu_r \nu_s ((\kappa_{\mathbf{G},rs} + \kappa_{\mathbf{G},sr}) \otimes h_{p_1} \otimes \cdots \otimes h_{p_{r-1}} \otimes g_{p_r} \otimes h_{p_{r+1}} \otimes \cdots \otimes h_{p_{s-1}} \otimes g_{p_s} \otimes h_{p_{s+1}} \otimes \cdots \otimes h_{p_d})(\mathbf{x},\boldsymbol{\theta}). \end{aligned} \tag{5.71}$$

In order to prove Theorem 5.1, some preliminary work is needed. Let us decompose the matrix $K_n^{[p]}$ into

$$K_n^{[p]} = T_{n+p-2}(f_p) + R_n^{[p]}, \tag{5.72}$$

where $T_{n+p-2}(f_p)$, the $(n+p-2)$-th Toeplitz matrix associated with f_p, is nothing else than the symmetric $(2\lfloor p/2 \rfloor + 1)$-band matrix whose generic central row is given by (5.36), while $R_n^{[p]} := K_n^{[p]} - T_{n+p-2}(f_p)$ is a low-rank correction term. Indeed, we know from Subsection 5.2.1 that $R_n^{[p]}$ has at most $2\lfloor 3p/2 \rfloor - 2$ nonzero rows, hence

$$\operatorname{rank}(R_n^{[p]}) \leq 2\lfloor 3p/2 \rfloor - 2 \leq 3p. \tag{5.73}$$

Similarly, we decompose the matrices $H_n^{[p]}$, $M_n^{[p]}$ into

$$H_n^{[p]} = \mathrm{i}T_{n+p-2}(g_p) + Q_n^{[p]}, \tag{5.74}$$

$$M_n^{[p]} = T_{n+p-2}(h_p) + S_n^{[p]}, \tag{5.75}$$

where $\mathrm{i}T_{n+p-2}(g_p) = T_{n+p-2}(\mathrm{i}g_p)$ and $T_{n+p-2}(h_p)$ are just the $(2\lfloor p/2 \rfloor + 1)$-band matrices whose generic central rows are given by (5.37) and (5.38), respectively, while $Q_n^{[p]} := H_n^{[p]} - \mathrm{i}T_{n+p-2}(g_p)$ and $S_n^{[p]} := M_n^{[p]} - T_{n+p-2}(h_p)$ are low-rank correction terms analogous to $R_n^{[p]}$:

$$\mathrm{rank}(Q_n^{[p]}) \leq 2\lfloor 3p/2 \rfloor - 2 \leq 3p. \tag{5.76}$$

$$\mathrm{rank}(S_n^{[p]}) \leq 2\lfloor 3p/2 \rfloor - 2 \leq 3p. \tag{5.77}$$

The next lemma provides upper bounds for the 2-norm of the matrices $K_n^{[p]}$, $H_n^{[p]}$, $M_n^{[p]}$.

Lemma 5.7. *For every $p \geq 2$ and every $n \geq 2$, we have*

$$\|M_n^{[p]}\| \leq \sqrt{\frac{3p}{2}}, \qquad \|H_n^{[p]}\| \leq p\sqrt{3p}, \qquad \|K_n^{[p]}\| \leq 2p(p-1)\sqrt{3p}.$$

Proof. By (1.3), the 2-norm of any square matrix X can be bounded as

$$\|X\| \leq \sqrt{\|X\|_\infty \|X^T\|_\infty}.$$

Hence, we now look for bounds of the infinity norm of the matrices $K_n^{[p]}$, $H_n^{[p]}$, $M_n^{[p]}$ and their transposes.

We first bound the infinity norm of $K_n^{[p]}$, $H_n^{[p]}$, $M_n^{[p]}$. From (5.21), the positivity property and the partition of unity property of B-splines, we obtain

$$\|M_n^{[p]}\|_\infty = \max_{i=1,\ldots,n+p-2} \sum_{j=1}^{n+p-2} N_{j+1,[p]}(\xi_{i+1,[p]}) \leq 1.$$

Similarly, from (5.21), (4.74), the partition of unity property of B-splines, and by taking into account that the sequence of knots (4.2)–(4.3) implies that $t_{i+p+1} - t_{i+1} \geq \frac{1}{n}$ for all $i = 1, \ldots, n+p-1$, we have

$$\|nH_n^{[p]}\|_\infty = \max_{i=1,\ldots,n+p-2} \sum_{j=1}^{n+p-2} \left| N'_{j+1,[p]}(\xi_{i+1,[p]}) \right| \leq \max_{i=1,\ldots,n+p-2} p \sum_{j=1}^{n+p-2} \left(\frac{N_{j+1,[p-1]}(\xi_{i+1,[p]})}{t_{j+p+1} - t_{j+1}} + \frac{N_{j+2,[p-1]}(\xi_{i+1,[p]})}{t_{j+p+2} - t_{j+2}} \right) \leq 2pn.$$

By using similar arguments and by iterating (4.74), from (5.21) we obtain

$$\|n^2 K_n^{[p]}\|_\infty = \max_{i=1,\ldots,n+p-2} \sum_{j=1}^{n+p-2} \left| N''_{j+1,[p]}(\xi_{i+1,[p]}) \right| \leq 4p(p-1)n^2.$$

We now bound the infinity norm of the transposes of $K_n^{[p]}$, $H_n^{[p]}$, $M_n^{[p]}$. The number of Greville abscissae in the interior of the support of any B-spline is at most $\frac{3p}{2}$. In combination with the positivity property and the partition of unity property of B-splines, we obtain

$$\|(M_n^{[p]})^T\|_\infty = \max_{j=1,\ldots,n+p-2} \sum_{i=1}^{n+p-2} N_{j+1,[p]}(\xi_{i+1,[p]}) \leq \frac{3p}{2}.$$

From (4.74), and by exploiting again the properties of the B-splines, we have

$$|N'_{j+1,[p]}(\xi_{i+1,[p]})| \le p\left(\frac{N_{j+1,[p-1]}(\xi_{i+1,[p]})}{t_{j+p+1}-t_{j+1}} + \frac{N_{j+2,[p-1]}(\xi_{i+1,[p]})}{t_{j+p+2}-t_{j+2}}\right) \le pn,$$

and so

$$\|(nH_n^{[p]})^T\|_\infty = \max_{j=1,\dots,n+p-2} \sum_{i=1}^{n+p-2} \left|N'_{j+1,[p]}(\xi_{i+1,[p]})\right| \le pn\frac{3p}{2}.$$

Similarly, by iterating (4.74), we obtain

$$\|(n^2K_n^{[p]})^T\|_\infty = \max_{i=1,\dots,n+p-2} \sum_{j=1}^{n+p-2} \left|N''_{j+1,[p]}(\xi_{i+1,[p]})\right| \le 2p(p-1)n^2\frac{3p}{2}.$$

The proof is completed by combining the above bounds with (1.3). □

The following lemma plays an important role in the proof of Theorem 5.1. It shows that the Greville abscissae (5.13) are somehow 'equivalent' to the uniform knots in $[0,1]$.

Lemma 5.8. *Let $p, n \ge 2$ and let $\xi_{i,[p]}$, $i = 2,\dots,n+p-1$, be the Greville abscissae* (5.13)*. Then, for every $i = 2,\dots,n+p-1$ and every j such that $|i-j| \le c$, with c a constant independent of n, we have*

$$\left|\xi_{i,[p]} - \frac{j}{n+p-2}\right| = O\left(\frac{1}{n}\right).$$

Proof. If $i \in \{p+1,\dots,n\}$, then $\xi_{i,[p]} = \frac{i}{n} - \frac{p+1}{2n}$ by (5.28) and

$$\begin{aligned}\left|\xi_{i,[p]} - \frac{j}{n+p-2}\right| &= \left|\frac{i}{n} - \frac{p+1}{2n} - \frac{j}{n+p-2}\right| = \left|\frac{n(i-j)+(p-2)i}{n(n+p-2)} - \frac{p+1}{2n}\right| \\ &\le \frac{c}{n+p-2} + \frac{(p-2)n}{n(n+p-2)} + \frac{p+1}{2n} = O\left(\frac{1}{n}\right).\end{aligned}$$

If $i \in \{2,\dots,p\}$, then

$$\begin{aligned}\left|\xi_{i,[p]} - \frac{j}{n+p-2}\right| &\le \xi_{i,[p]} + \frac{|j|}{n+p-2} \le \xi_{p+1,[p]} + \frac{|j-i|}{n+p-2} + \frac{i}{n+p-2} \\ &\le \frac{p+1}{2n} + \frac{c}{n+p-2} + \frac{p}{n+p-2} = O\left(\frac{1}{n}\right).\end{aligned}$$

If $i \in \{n+1,\dots,n+p-1\}$, then

$$\begin{aligned}\left|\xi_{i,[p]} - \frac{j}{n+p-2}\right| &\le |\xi_{i,[p]} - 1| + \left|1 - \frac{j-i}{n+p-2} - \frac{i}{n+p-2}\right| \le |\xi_{n,[p]} - 1| + \left|1 - \frac{i}{n+p-2}\right| + \frac{|j-i|}{n+p-2} \\ &\le \frac{p+1}{2n} + \max\left(\left|1 - \frac{n+1}{n+p-2}\right|, \left|1 - \frac{n+p-1}{n+p-2}\right|\right) + \frac{c}{n+p-2} = O\left(\frac{1}{n}\right).\end{aligned}$$

□

We are almost ready for proving Theorem 5.1, but we still need to recall the concept of modulus of continuity, which is used in the proof. For any function $\psi : [0,1]^d \to \mathbb{R}$, the modulus of continuity of ψ is defined as the function $\omega(\psi,\cdot) : (0,\infty) \to [0,\infty]$,

$$\omega(\psi,\delta) := \sup_{\substack{\mathbf{x},\mathbf{y}\in[0,1]^d \\ \|\mathbf{x}-\mathbf{y}\|_\infty \le \delta}} |\psi(\mathbf{x}) - \psi(\mathbf{y})|.$$

If ψ is continuous over $[0,1]^d$, and hence uniformly continuous by the Heine–Cantor theorem, then

$$\lim_{\delta\to 0}\omega(\psi,\delta)=0.$$

Theorem 5.1 (**spectral distribution without geometry map**). *Let $\boldsymbol{p}\geq\mathbf{2}$, $\boldsymbol{\nu}\in\mathbb{Q}_+^d$ and $\boldsymbol{n}=\boldsymbol{\nu}n$. Then, $\{\frac{1}{n^2}A_{\boldsymbol{n}}^{[\boldsymbol{p}]}\}_n\sim_\lambda f_{\boldsymbol{p}}^{(\boldsymbol{\nu})}$, with $f_{\boldsymbol{p}}^{(\boldsymbol{\nu})}$ defined in* (5.70)*. In particular, $\{\frac{1}{n^2}A_{\boldsymbol{n}}^{[\boldsymbol{p}]}\}_n$ is weakly clustered at the range $[0,M_{f_{\boldsymbol{p}}^{(\boldsymbol{\nu})}}]$ of $f_{\boldsymbol{p}}^{(\boldsymbol{\nu})}$, where $M_{f_{\boldsymbol{p}}^{(\boldsymbol{\nu})}}:=\max_{(\mathbf{x},\boldsymbol{\theta})\in[0,1]^d\times[-\pi,\pi]^d}f_{\boldsymbol{p}}^{(\boldsymbol{\nu})}(\mathbf{x},\boldsymbol{\theta})$, and every point of $[0,M_{f_{\boldsymbol{p}}^{(\boldsymbol{\nu})}}]$ strongly attracts $\Lambda(\frac{1}{n^2}A_{\boldsymbol{n}}^{[\boldsymbol{p}]})$ with infinite order (cf. Theorem 1.5).*

Proof. Throughout this proof, the letter C will denote a generic constant independent of n. Recalling (5.16), we have

$$\frac{1}{n^2}A_{\boldsymbol{n}}^{[\boldsymbol{p}]}=\frac{1}{n^2}A_{\boldsymbol{n},D}^{[\boldsymbol{p}]}+\frac{1}{n^2}A_{\boldsymbol{n},A}^{[\boldsymbol{p}]}+\frac{1}{n^2}A_{\boldsymbol{n},R}^{[\boldsymbol{p}]}=\frac{1}{n^2}\tilde{A}_{\boldsymbol{n},D}^{[\boldsymbol{p}]}+\left(\frac{1}{n^2}A_{\boldsymbol{n},D}^{[\boldsymbol{p}]}-\frac{1}{n^2}\tilde{A}_{\boldsymbol{n},D}^{[\boldsymbol{p}]}\right)+\frac{1}{n^2}A_{\boldsymbol{n},A}^{[\boldsymbol{p}]}+\frac{1}{n^2}A_{\boldsymbol{n},R}^{[\boldsymbol{p}]},$$

where

$$\begin{aligned}\tilde{A}_{\boldsymbol{n},D}^{[\boldsymbol{p}]}&=\sum_{r=1}^{d}n_r^2\tilde{D}_{\boldsymbol{n}+\boldsymbol{p}-\mathbf{2}}(\kappa_{rr})\circ T_{\boldsymbol{n}+\boldsymbol{p}-\mathbf{2}}(h_{p_1}\otimes\cdots\otimes h_{p_{r-1}}\otimes f_{p_r}\otimes h_{p_{r+1}}\otimes\cdots\otimes h_{p_d})\\&-\sum_{\substack{r,s=1\\r<s}}^{d}n_rn_s\tilde{D}_{\boldsymbol{n}+\boldsymbol{p}-\mathbf{2}}(\kappa_{rs}+\kappa_{sr})\circ T_{\boldsymbol{n}+\boldsymbol{p}-\mathbf{2}}(h_{p_1}\otimes\cdots\otimes h_{p_{r-1}}\otimes \mathrm{i}g_{p_r}\otimes h_{p_{r+1}}\otimes\cdots\otimes h_{p_{s-1}}\otimes \mathrm{i}g_{p_s}\otimes h_{p_{s+1}}\otimes\cdots\otimes h_{p_d}),\end{aligned}\tag{5.78}$$

and $\tilde{D}_{\boldsymbol{m}}(a)$ is defined in (1.50)–(1.51). Before proceeding further, we suggest comparing the expression of $\tilde{A}_{\boldsymbol{n},D}^{[\boldsymbol{p}]}$ with the expression of $A_{\boldsymbol{n},D}^{[\boldsymbol{p}]}$ in (5.17), taking into account the relations (5.72), (5.74), (5.75) and noting that Lemma 1.8 can be applied here to express the d-level Toeplitz matrices involved in (5.78) as tensor products of unilevel Toeplitz matrices. We will show that the hypotheses of Theorem 1.6 are satisfied with

$$Z_n:=\frac{1}{n^2}A_{\boldsymbol{n}}^{[\boldsymbol{p}]},\qquad X_n:=\frac{1}{n^2}\tilde{A}_{\boldsymbol{n},D}^{[\boldsymbol{p}]},\qquad Y_n:=\left(\frac{1}{n^2}A_{\boldsymbol{n},D}^{[\boldsymbol{p}]}-\frac{1}{n^2}\tilde{A}_{\boldsymbol{n},D}^{[\boldsymbol{p}]}\right)+\frac{1}{n^2}A_{\boldsymbol{n},A}^{[\boldsymbol{p}]}+\frac{1}{n^2}A_{\boldsymbol{n},R}^{[\boldsymbol{p}]}.$$

X_n is Hermitian, due to the properties of the Hadamard product (see Subsection 1.2.2) and to the fact that the generating functions of the d-level Toeplitz matrices in (5.78) are real-valued. Moreover, $\|X_n\|$ is uniformly bounded with respect to n, due to Lemma 1.6 (item 1), to the fact that κ_{rs}, $r,s=1,\dots,d$, and the generating functions of the d-level Toeplitz matrices in (5.78) are continuous, and to the inequality (1.37). Finally, we have

$$\{X_n\}\sim_\lambda f_{\boldsymbol{p}}^{(\boldsymbol{\nu})},$$

by Corollary 1.1 and by the assumption $\boldsymbol{n}=\boldsymbol{\nu}n$. In addition, $\|Z_n\|$ is uniformly bounded with respect to n, due to the inequality (1.13), to the fact that κ_{rs}, $r,s=1,\dots,d$, are continuous, and to Lemma 5.7.

Now we turn to Y_n. We will analyze separately the three summands that compose Y_n and show that each of them has a $o(n^d)$ trace-norm. By using the expression (5.18) of $A_{\boldsymbol{n},A}^{[\boldsymbol{p}]}$, the inequality (1.13), Lemma 5.7 and the continuity of β_r, $r=1,\dots,d$, we have

$$\begin{aligned}&\left\|\frac{1}{n^2}A_{\boldsymbol{n},A}^{[\boldsymbol{p}]}\right\|=\frac{1}{n^2}\left\|\sum_{r=1}^{d}n_rD_{\boldsymbol{n}}^{[\boldsymbol{p}]}(\beta_r)(M_{n_1}^{[p_1]}\otimes\cdots\otimes M_{n_{r-1}}^{[p_{r-1}]}\otimes H_{n_r}^{[p_r]}\otimes M_{n_{r+1}}^{[p_{r+1}]}\otimes\cdots\otimes M_{n_d}^{[p_d]})\right\|=O\left(\frac{1}{n}\right)\\&\Rightarrow\ \left|\!\left|\!\left|\frac{1}{n^2}A_{\boldsymbol{n},A}^{[\boldsymbol{p}]}\right|\!\right|\!\right|_1\leq N(\boldsymbol{n}+\boldsymbol{p}-\mathbf{2})\left\|\frac{1}{n^2}A_{\boldsymbol{n},A}^{[\boldsymbol{p}]}\right\|=O(n^{d-1}).\end{aligned}\tag{5.79}$$

By using the expression (5.19) of $A_{\boldsymbol{n},R}^{[\boldsymbol{p}]}$, the inequality (1.13), Lemma 5.7 and the continuity of γ, we have

$$\left\|\frac{1}{n^2}A_{\boldsymbol{n},R}^{[\boldsymbol{p}]}\right\|=\frac{1}{n^2}\left\|D_{\boldsymbol{n}}^{[\boldsymbol{p}]}(\gamma)(M_{n_1}^{[p_1]}\otimes\cdots\otimes M_{n_d}^{[p_d]})\right\|=O\left(\frac{1}{n^2}\right)\ \Rightarrow\ \left|\!\left|\!\left|\frac{1}{n^2}A_{\boldsymbol{n},R}^{[\boldsymbol{p}]}\right|\!\right|\!\right|_1=O(n^{d-2}).\tag{5.80}$$

Now we consider the term $\frac{1}{n^2}A^{[\boldsymbol{p}]}_{\boldsymbol{n},D} - \frac{1}{n^2}\tilde{A}^{[\boldsymbol{p}]}_{\boldsymbol{n},D}$. Keeping in mind that $\boldsymbol{n} = \boldsymbol{\nu} n$, we decompose this term as follows:

$$
\begin{aligned}
&\frac{1}{n^2}A^{[\boldsymbol{p}]}_{\boldsymbol{n},D} - \frac{1}{n^2}\tilde{A}^{[\boldsymbol{p}]}_{\boldsymbol{n},D} \\
&= \sum_{r=1}^{d} \nu_r^2 \Big[D^{[\boldsymbol{p}]}_{\boldsymbol{n}}(\kappa_{rr})(M^{[p_1]}_{n_1} \otimes \cdots \otimes M^{[p_{r-1}]}_{n_{r-1}} \otimes K^{[p_r]}_{n_r} \otimes M^{[p_{r+1}]}_{n_{r+1}} \otimes \cdots \otimes M^{[p_d]}_{n_d}) \\
&\qquad\qquad - \tilde{D}_{\boldsymbol{n}+\boldsymbol{p}-\boldsymbol{2}}(\kappa_{rr}) \circ T_{\boldsymbol{n}+\boldsymbol{p}-\boldsymbol{2}}(h_{p_1} \otimes \cdots \otimes h_{p_{r-1}} \otimes f_{p_r} \otimes h_{p_{r+1}} \otimes \cdots \otimes h_{p_d}) \Big] \\
&- \sum_{\substack{r,s=1 \\ r<s}}^{d} \nu_r \nu_s \Big[D^{[\boldsymbol{p}]}_{\boldsymbol{n}}(\kappa_{rs}+\kappa_{sr})(M^{[p_1]}_{n_1} \otimes \cdots \otimes M^{[p_{r-1}]}_{n_{r-1}} \otimes H^{[p_r]}_{n_r} \otimes M^{[p_{r+1}]}_{n_{r+1}} \otimes \cdots \otimes M^{[p_{s-1}]}_{n_{s-1}} \otimes H^{[p_s]}_{n_s} \otimes M^{[p_{s+1}]}_{n_{s+1}} \otimes \cdots \otimes M^{[p_d]}_{n_d}) \\
&\qquad\qquad - \tilde{D}_{\boldsymbol{n}+\boldsymbol{p}-\boldsymbol{2}}(\kappa_{rs}+\kappa_{sr}) \circ T_{\boldsymbol{n}+\boldsymbol{p}-\boldsymbol{2}}(h_{p_1} \otimes \cdots \otimes h_{p_{r-1}} \otimes \mathrm{i}g_{p_r} \otimes h_{p_{r+1}} \otimes \cdots \otimes h_{p_{s-1}} \otimes \mathrm{i}g_{p_s} \otimes h_{p_{s+1}} \otimes \cdots \otimes h_{p_d}) \Big] \\
&= \sum_{r=1}^{d} \nu_r^2 \Big[D^{[\boldsymbol{p}]}_{\boldsymbol{n}}(\kappa_{rr})(M^{[p_1]}_{n_1} \otimes \cdots \otimes M^{[p_{r-1}]}_{n_{r-1}} \otimes K^{[p_r]}_{n_r} \otimes M^{[p_{r+1}]}_{n_{r+1}} \otimes \cdots \otimes M^{[p_d]}_{n_d}) \\
&\qquad\qquad - D^{[\boldsymbol{p}]}_{\boldsymbol{n}}(\kappa_{rr}) T_{\boldsymbol{n}+\boldsymbol{p}-\boldsymbol{2}}(h_{p_1} \otimes \cdots \otimes h_{p_{r-1}} \otimes f_{p_r} \otimes h_{p_{r+1}} \otimes \cdots \otimes h_{p_d}) \Big] \qquad (5.81) \\
&+ \sum_{r=1}^{d} \nu_r^2 \Big[D^{[\boldsymbol{p}]}_{\boldsymbol{n}}(\kappa_{rr}) T_{\boldsymbol{n}+\boldsymbol{p}-\boldsymbol{2}}(h_{p_1} \otimes \cdots \otimes h_{p_{r-1}} \otimes f_{p_r} \otimes h_{p_{r+1}} \otimes \cdots \otimes h_{p_d}) \\
&\qquad\qquad - \tilde{D}_{\boldsymbol{n}+\boldsymbol{p}-\boldsymbol{2}}(\kappa_{rr}) \circ T_{\boldsymbol{n}+\boldsymbol{p}-\boldsymbol{2}}(h_{p_1} \otimes \cdots \otimes h_{p_{r-1}} \otimes f_{p_r} \otimes h_{p_{r+1}} \otimes \cdots \otimes h_{p_d}) \Big] \qquad (5.82) \\
&- \sum_{\substack{r,s=1 \\ r<s}}^{d} \nu_r \nu_s \Big[D^{[\boldsymbol{p}]}_{\boldsymbol{n}}(\kappa_{rs}+\kappa_{sr})(M^{[p_1]}_{n_1} \otimes \cdots \otimes M^{[p_{r-1}]}_{n_{r-1}} \otimes H^{[p_r]}_{n_r} \otimes M^{[p_{r+1}]}_{n_{r+1}} \otimes \cdots \otimes M^{[p_{s-1}]}_{n_{s-1}} \otimes H^{[p_s]}_{n_s} \otimes M^{[p_{s+1}]}_{n_{s+1}} \otimes \cdots \otimes M^{[p_d]}_{n_d}) \\
&\qquad\qquad - D^{[\boldsymbol{p}]}_{\boldsymbol{n}}(\kappa_{rs}+\kappa_{sr}) T_{\boldsymbol{n}+\boldsymbol{p}-\boldsymbol{2}}(h_{p_1} \otimes \cdots \otimes h_{p_{r-1}} \otimes \mathrm{i}g_{p_r} \otimes h_{p_{r+1}} \otimes \cdots \otimes h_{p_{s-1}} \otimes \mathrm{i}g_{p_s} \otimes h_{p_{s+1}} \otimes \cdots \otimes h_{p_d}) \Big] \qquad (5.83) \\
&- \sum_{\substack{r,s=1 \\ r<s}}^{d} \nu_r \nu_s \Big[D^{[\boldsymbol{p}]}_{\boldsymbol{n}}(\kappa_{rs}+\kappa_{sr}) T_{\boldsymbol{n}+\boldsymbol{p}-\boldsymbol{2}}(h_{p_1} \otimes \cdots \otimes h_{p_{r-1}} \otimes \mathrm{i}g_{p_r} \otimes h_{p_{r+1}} \otimes \cdots \otimes h_{p_{s-1}} \otimes \mathrm{i}g_{p_s} \otimes h_{p_{s+1}} \otimes \cdots \otimes h_{p_d}) \\
&\qquad\qquad - \tilde{D}_{\boldsymbol{n}+\boldsymbol{p}-\boldsymbol{2}}(\kappa_{rs}+\kappa_{sr}) \circ T_{\boldsymbol{n}+\boldsymbol{p}-\boldsymbol{2}}(h_{p_1} \otimes \cdots \otimes h_{p_{r-1}} \otimes \mathrm{i}g_{p_r} \otimes h_{p_{r+1}} \otimes \cdots \otimes h_{p_{s-1}} \otimes \mathrm{i}g_{p_s} \otimes h_{p_{s+1}} \otimes \cdots \otimes h_{p_d}) \Big]. \qquad (5.84)
\end{aligned}
$$

Taking into account that

$$
\begin{aligned}
&T_{\boldsymbol{n}+\boldsymbol{p}-\boldsymbol{2}}(h_{p_1} \otimes \cdots \otimes h_{p_{r-1}} \otimes f_{p_r} \otimes h_{p_{r+1}} \otimes \cdots \otimes h_{p_d}) \\
&= T_{n_1+p_1-2}(h_{p_1}) \otimes \cdots \otimes T_{n_{r-1}+p_{r-1}-2}(h_{p_{r-1}}) \otimes T_{n_r+p_r-2}(f_{p_r}) \otimes T_{n_{r+1}+p_{r+1}-2}(h_{p_{r+1}}) \otimes \cdots \otimes T_{n_d+p_d-2}(h_{p_d}) \qquad (5.85)
\end{aligned}
$$

(see Lemma 1.8), the trace-norm of the r-th term in the first summation (5.81) can be bounded using (1.18). Recalling the inequalities (5.73), (5.77) and (1.5), we have

$$
\begin{aligned}
&\Big\| D^{[\boldsymbol{p}]}_{\boldsymbol{n}}(\kappa_{rr})(M^{[p_1]}_{n_1} \otimes \cdots \otimes M^{[p_{r-1}]}_{n_{r-1}} \otimes K^{[p_r]}_{n_r} \otimes M^{[p_{r+1}]}_{n_{r+1}} \otimes \cdots \otimes M^{[p_d]}_{n_d} - T_{\boldsymbol{n}+\boldsymbol{p}-\boldsymbol{2}}(h_{p_1} \otimes \cdots \otimes h_{p_{r-1}} \otimes f_{p_r} \otimes h_{p_{r+1}} \otimes \cdots \otimes h_{p_d})) \Big\|_1 \\
&\le C\, N(\boldsymbol{n}+\boldsymbol{p}-\boldsymbol{2}) \sum_{i=1}^{d} \frac{3p_i}{n_i+p_i-2}, \qquad (5.86)
\end{aligned}
$$

where C is some constant independent of n that provides an upper bound for the spectral norm of the matrix in the left-hand side of (5.86). It follows that the trace-norm in (5.86) is $o(n^d)$ and, consequently, the trace-norm of the first summation (5.81) is $o(n^d)$. With the same argument, one can show that the trace-norm of the (r,s)-th term in the third summation (5.83) is $o(n^d)$, implying that the trace-norm of the

third summation itself is $o(n^d)$. Concerning the second summation (5.82), for every $\boldsymbol{i},\boldsymbol{j} = \mathbf{1},\dots,\boldsymbol{n}+\boldsymbol{p}-\mathbf{2}$, the $(\boldsymbol{i},\boldsymbol{j})$ entry in the r-th term of (5.82) is given by

$$\left(\kappa_{rr}(\boldsymbol{\xi}_{\boldsymbol{i}+\mathbf{1},[\boldsymbol{p}]}) - \kappa_{rr}\left(\frac{\boldsymbol{i}\wedge\boldsymbol{j}-\mathbf{1}}{\boldsymbol{n}+\boldsymbol{p}-\mathbf{2}}\right)\right)\left(T_{\boldsymbol{n}+\boldsymbol{p}-\mathbf{2}}(h_{p_1}\otimes\cdots\otimes h_{p_{r-1}}\otimes f_{p_r}\otimes h_{p_{r+1}}\otimes\cdots\otimes h_{p_d})\right)_{\boldsymbol{i},\boldsymbol{j}}; \tag{5.87}$$

to see this, recall the definitions of $D_{\boldsymbol{n}}^{[\boldsymbol{p}]}(a)$ in (5.20), the definition of $D_{\boldsymbol{n}+\boldsymbol{p}-\mathbf{2}}(a)$ in (1.49), and the definition of $\tilde{D}_{\boldsymbol{n}+\boldsymbol{p}-\mathbf{2}}(a)$ in (1.51). Using (5.85) together with the fundamental property (1.12), the $(\boldsymbol{i},\boldsymbol{j})$ entry (5.87) is equal to

$$\left(\kappa_{rr}(\boldsymbol{\xi}_{\boldsymbol{i}+\mathbf{1},[\boldsymbol{p}]}) - \kappa_{rr}\left(\frac{\boldsymbol{i}\wedge\boldsymbol{j}-\mathbf{1}}{\boldsymbol{n}+\boldsymbol{p}-\mathbf{2}}\right)\right)(T_{n_r+p_r-2}(f_{p_r}))_{i_r j_r}\prod_{\substack{k=1\\k\neq r}}^{d}(T_{n_k+p_k-2}(h_{p_k}))_{i_k j_k}$$

and is zero for $\|\boldsymbol{i}-\boldsymbol{j}\|_\infty > \|\boldsymbol{p}/2\|_\infty$, because $T_{n+p-2}(h_p)$, $T_{n+p-2}(g_p)$, $T_{n+p-2}(f_p)$ have a $(2\lfloor p/2\rfloor+1)$-band structure. Therefore, the only nonzero entries (5.87) are obtained for $\|\boldsymbol{i}-\boldsymbol{j}\|_\infty \le \|\boldsymbol{p}/2\|_\infty$. For any multi-indices $\boldsymbol{i},\boldsymbol{j}$ satisfying this condition, we have $0 \le |i_k - j_k| \le \|\boldsymbol{p}/2\|_\infty$ for all $k = 1,\dots,d$, and so, by Lemma 5.8,

$$\left|\xi_{i_k+1,[p_k]} - \frac{(\boldsymbol{i}\wedge\boldsymbol{j})_k - 1}{n_k + p_k - 2}\right| = O\left(\frac{1}{n_k}\right) = O\left(\frac{1}{n}\right), \qquad k = 1,\dots,d.$$

It follows that, for all $\boldsymbol{i},\boldsymbol{j} = \mathbf{1},\dots,\boldsymbol{n}+\boldsymbol{p}-\mathbf{2}$ such that $\|\boldsymbol{i}-\boldsymbol{j}\|_\infty \le \|\boldsymbol{p}/2\|_\infty$, we have

$$\left\|\boldsymbol{\xi}_{\boldsymbol{i}+\mathbf{1},[\boldsymbol{p}]} - \frac{\boldsymbol{i}\wedge\boldsymbol{j}-\mathbf{1}}{\boldsymbol{n}+\boldsymbol{p}-\mathbf{2}}\right\|_\infty \le \frac{C}{n}, \tag{5.88}$$

and

$$\left|\kappa_{rr}(\boldsymbol{\xi}_{\boldsymbol{i}+\mathbf{1},[\boldsymbol{p}]}) - \kappa_{rr}\left(\frac{\boldsymbol{i}\wedge\boldsymbol{j}-\mathbf{1}}{\boldsymbol{n}+\boldsymbol{p}-\mathbf{2}}\right)\right| \le C\,\omega\left(\kappa_{rr},\frac{1}{n}\right) \tag{5.89}$$

(the constant C in (5.89) is not necessarily the same as the constant C in (5.88); recall that in this proof the letter C denotes a generic constant independent of n). The inequalities (5.89) and (1.37) imply that

$$\left|\left(\kappa_{rr}(\boldsymbol{\xi}_{\boldsymbol{i}+\mathbf{1},[\boldsymbol{p}]}) - \kappa_{rr}\left(\frac{\boldsymbol{i}\wedge\boldsymbol{j}-\mathbf{1}}{\boldsymbol{n}+\boldsymbol{p}-\mathbf{2}}\right)\right)\left(T_{\boldsymbol{n}+\boldsymbol{p}-\mathbf{2}}(h_{p_1}\otimes\cdots\otimes h_{p_{r-1}}\otimes f_{p_r}\otimes h_{p_{r+1}}\otimes\cdots\otimes h_{p_d})\right)_{\boldsymbol{i},\boldsymbol{j}}\right| \le C\,\omega\left(\kappa_{rr},\frac{1}{n}\right).$$

Recalling that the number of nonzero entries (5.87) for a fixed row or column is independent of n, we have proved that the (r,s)-th matrix in (5.82) and its transpose have infinity norms bounded from above by

$$C\,\omega\left(\kappa_{rr},\frac{1}{n}\right). \tag{5.90}$$

Hence, by (1.3), also the spectral norms of the (r,s)-th matrix in (5.82) and of its transpose are bounded from above by (5.90), and so their trace-norms are bounded from above by

$$C\,\omega\left(\kappa_{rr},\frac{1}{n}\right)N(\boldsymbol{n}+\boldsymbol{p}-\mathbf{2}) = o(n^d), \tag{5.91}$$

implying that the trace-norm of the whole summation (5.82) is $o(n^d)$. With the same argument, one can show that the trace-norm of the last summation (5.84) is $o(n^d)$. Hence, $\left|\left|\left|\frac{1}{n^2}A_{\boldsymbol{n},D}^{[\boldsymbol{p}]} - \frac{1}{n^2}\tilde{A}_{\boldsymbol{n},D}^{[\boldsymbol{p}]}\right|\right|\right|_1 = o(n^d)$ and, by recalling (5.79)–(5.80), we conclude that $|||Y_n|||_1 = o(n^d)$. □

Remark 5.2. Following the steps of the previous proof in the case $d = 1$, with $\boldsymbol{\nu} = 1$ and $\boldsymbol{p} = p$ (see in particular (5.79), (5.80) and (5.91)), it turns out that, if

$$\omega\left(\kappa, \frac{1}{n}\right) = O\left(\frac{1}{n}\right), \tag{5.92}$$

then

$$|||Y_n|||_1 = O(1).$$

Therefore, all the hypotheses of Theorem 1.7 are satisfied and the sequence of normalized univariate collocation matrices $\{\frac{1}{n^2}A_n^{[p]}\}_n$ is strongly clustered at $[0, M_{\kappa\otimes f_p}]$; note that $A_n^{[p]}$ is given explicitly by (5.22), while $\kappa \otimes f_p$ (with f_p as in (5.34)) coincides precisely with $f_{\boldsymbol{p}}^{(\boldsymbol{\nu})}$ in the case $\boldsymbol{p} = p$ and $\boldsymbol{\nu} = 1$. In particular, if $\kappa \in C^1([0,1])$, then (5.92) is satisfied and $\{\frac{1}{n^2}A_n^{[p]}\}_n$ is strongly clustered at $[0, M_{\kappa\otimes f_p}]$.

Remark 5.3. Suppose that the diffusion coefficient K in (5.2) is just the identity matrix: $K = I$. In this case, the symbol of the normalized IgA collocation matrices $\{\frac{1}{n^2}A_{\boldsymbol{n}}^{[\boldsymbol{p}]}\}_n$ ($\boldsymbol{n} = \boldsymbol{\nu}n$) is given by, cf. (5.70),

$$\begin{aligned} f_{\boldsymbol{p}}^{(\boldsymbol{\nu})} : [0,1]^d \times [-\pi,\pi]^d \to \mathbb{R}, \qquad f_{\boldsymbol{p}}^{(\boldsymbol{\nu})}(\mathbf{x},\boldsymbol{\theta}) &= \sum_{r=1}^{d} \nu_r^2 (1 \otimes h_{p_1} \otimes \cdots \otimes h_{p_{r-1}} \otimes f_{p_r} \otimes h_{p_{r+1}} \otimes \cdots \otimes h_{p_d})(\mathbf{x},\boldsymbol{\theta}) \\ &= \sum_{r=1}^{d} \nu_r^2 (h_{p_1} \otimes \cdots \otimes h_{p_{r-1}} \otimes f_{p_r} \otimes h_{p_{r+1}} \otimes \cdots \otimes h_{p_d})(\boldsymbol{\theta}). \end{aligned}$$

Since $f_{\boldsymbol{p}}^{(\boldsymbol{\nu})}$ is independent of $\mathbf{x}$, it follows from Definition 1.1 that $f_{\boldsymbol{p}}^{(\boldsymbol{\nu})}$ regarded as a function from $[-\pi,\pi]^d$ to $\mathbb{R}$, i.e.

$$f_{\boldsymbol{p}}^{(\boldsymbol{\nu})} : [-\pi,\pi]^d \to \mathbb{R}, \qquad f_{\boldsymbol{p}}^{(\boldsymbol{\nu})}(\boldsymbol{\theta}) = \sum_{r=1}^{d} \nu_r^2 (h_{p_1} \otimes \cdots \otimes h_{p_{r-1}} \otimes f_{p_r} \otimes h_{p_{r+1}} \otimes \cdots \otimes h_{p_d})(\boldsymbol{\theta}), \tag{5.93}$$

is still a symbol for $\{\frac{1}{n^2}A_{\boldsymbol{n}}^{[\boldsymbol{p}]}\}_n$. In particular, in the case where $\boldsymbol{\nu} = \mathbf{1}$ and $\boldsymbol{p} = 2\boldsymbol{q} + \mathbf{1}$ for some $\boldsymbol{q} \in \mathbb{N}^d$, the symbol (5.93) of the normalized IgA collocation matrices $\{\frac{1}{n^2}A_{\boldsymbol{n}}^{[2\boldsymbol{q}+1]}\}_n$ is the same as the symbol of the normalized IgA Galerkin matrices $\{n^{d-2}A_{\boldsymbol{n}}^{[\boldsymbol{q}]}\}_n$ considered in Chapter 4; see Theorem 4.7, Remark 5.1, and the expression (4.65) for the symbol of normalized IgA Galerkin matrices.

Recalling the discussion at the beginning of this section, from Theorem 5.1 and Remark 5.2 we obtain the following result for the sequence of normalized IgA collocation matrices $\{A_{\mathbf{G},\boldsymbol{n}}^{[\boldsymbol{p}]}\}_n$, $\boldsymbol{n} = \boldsymbol{\nu}n$.

Theorem 5.2 (**spectral distribution with a geometry map**). *Let $\mathbf{G} : \hat{\Omega} = [0,1]^d \to \overline{\Omega}$ be a geometry map such that $\mathbf{G} \in C^2(\hat{\Omega})$ and $\mathbf{G}$ is invertible in $\hat{\Omega}$ with $\mathbf{G}(\partial\hat{\Omega}) = \partial\overline{\Omega}$. Let $\boldsymbol{p} \geq \mathbf{2}$, $\boldsymbol{\nu} \in \mathbb{Q}_+^d$ and $\boldsymbol{n} = \boldsymbol{\nu}n$. Then, $\{\frac{1}{n^2}A_{\mathbf{G},\boldsymbol{n}}^{[\boldsymbol{p}]}\}_n \sim_\lambda f_{\mathbf{G},\boldsymbol{p}}^{(\boldsymbol{\nu})}$, with $f_{\mathbf{G},\boldsymbol{p}}^{(\boldsymbol{\nu})}$ defined in (5.71). In particular, $\{\frac{1}{n^2}A_{\mathbf{G},\boldsymbol{n}}^{[\boldsymbol{p}]}\}_n$ is weakly clustered at the range $[0, M_{f_{\mathbf{G},\boldsymbol{p}}^{(\boldsymbol{\nu})}}]$ of $f_{\mathbf{G},\boldsymbol{p}}^{(\boldsymbol{\nu})}$, where $M_{f_{\mathbf{G},\boldsymbol{p}}^{(\boldsymbol{\nu})}} := \max_{\mathbf{x}\in[0,1]^d,\,\boldsymbol{\theta}\in[-\pi,\pi]^d} f_{\mathbf{G},\boldsymbol{p}}^{(\boldsymbol{\nu})}(\mathbf{x},\boldsymbol{\theta})$, and every point of $[0, M_{f_{\mathbf{G},\boldsymbol{p}}^{(\boldsymbol{\nu})}}]$ strongly attracts $\Lambda(\frac{1}{n^2}A_{\mathbf{G},\boldsymbol{n}}^{[\boldsymbol{p}]})$ with infinite order. Moreover, in the case $d = 1$, if the function κ_G defined in (5.26) satisfies*

$$\omega\left(\kappa_G, \frac{1}{n}\right) = O\left(\frac{1}{n}\right),$$

then the sequence of univariate collocation matrices $\{\frac{1}{n^2}A_{G,n}^{[p]}\}_n$ is strongly clustered at the range $[0, M_{f_{G,p}}]$ of $f_{G,p} = \kappa_G \otimes f_p$.

Remark 5.4. Note that the spectral distribution results obtained in Theorems 5.1 and 5.2 hold without any assumption on the coefficient matrix K except continuity. However, in order to ensure that (5.2) is an elliptic problem, this matrix has to be SPD. Moreover, the geometry map $\mathbf{G}$ in Theorem 5.2 can be given in any representation and is not confined to the B-spline form (5.9) as prescribed by the IgA paradigm.

We conclude with the observation that the structure of the symbol $f^{(\nu)}_{\mathbf{G},\boldsymbol{p}}$ incorporates:

(a) the approximation technique, which is identified by a trigonometric polynomial in the Fourier variables $\boldsymbol{\theta} = (\theta_1, \dots, \theta_d) \in [-\pi, \pi]^d$;

(b) the geometry, which is identified by the map $\mathbf{G}$ in the variables $\hat{\mathbf{x}} = (\hat{x}_1, \dots, \hat{x}_d)$ defined on the parametric domain $\hat{\Omega} := [0,1]^d$;

(c) the coefficients of the principal terms of the PDE, namely K, in the physical variables $\mathbf{x} := (x_1, \dots, x_d)$ defined on the physical domain Ω.

In reality, the above picture is intrinsic to the approximation of PDE by any local method, such as Finite Differences and Finite Elements. In fact, formally, the structure of the symbol is substantially the same when considering different techniques to approximate the same problem; see [6, 63, 64] and references therein. The only difference is due to the polynomial in the Fourier variables $\boldsymbol{\theta}$, and this is no surprise, since this part specifically depends on the chosen approximation technique (in this case, the IgA Collocation Method).

5.4.1 Properties of the spectral symbol

The symbols in Theorems 5.1 and 5.2, given by (5.70)–(5.71), can be compactly written in matrix form as

$$f^{(\nu)}_{\boldsymbol{p}} = [\nu_1 \ \cdots \ \nu_d] \left(K \circ P_{p_1,\dots,p_d} \right) [\nu_1 \ \cdots \ \nu_d]^T = \boldsymbol{\nu} (K \circ P_{p_1,\dots,p_d}) \boldsymbol{\nu}^T, \tag{5.94}$$

$$f^{(\nu)}_{\mathbf{G},\boldsymbol{p}} = [\nu_1 \ \cdots \ \nu_d] \left(K_{\mathbf{G}} \circ P_{p_1,\dots,p_d} \right) [\nu_1 \ \cdots \ \nu_d]^T = \boldsymbol{\nu} (K_{\mathbf{G}} \circ P_{p_1,\dots,p_d}) \boldsymbol{\nu}^T, \tag{5.95}$$

where K is the coefficient matrix of our problem (5.2), $K_{\mathbf{G}}$ is the transformed coefficient matrix given in (5.25) and

$$\left(P_{p_1,\dots,p_d} \right)_{rs} := \begin{cases} h_{p_1} \otimes \cdots \otimes h_{p_{r-1}} \otimes f_{p_r} \otimes h_{p_{r+1}} \otimes \cdots \otimes h_{p_d}, & \text{if } r = s, \\ h_{p_1} \otimes \cdots \otimes h_{p_{r-1}} \otimes g_{p_r} \otimes h_{p_{r+1}} \otimes \cdots \otimes h_{p_{s-1}} \otimes g_{p_s} \otimes h_{p_{s+1}} \otimes \cdots \otimes h_{p_d}, & \text{if } r < s, \\ h_{p_1} \otimes \cdots \otimes h_{p_{s-1}} \otimes g_{p_s} \otimes h_{p_{s+1}} \otimes \cdots \otimes h_{p_{r-1}} \otimes g_{p_r} \otimes h_{p_{r+1}} \otimes \cdots \otimes h_{p_d}, & \text{if } r > s. \end{cases} \tag{5.96}$$

In (5.94)–(5.95), it is understood that $K = K(\mathbf{x})$ and $K_{\mathbf{G}} = K_{\mathbf{G}}(\mathbf{x})$ are functions of $\mathbf{x}$, while $P_{p_1,\dots,p_d} = P_{p_1,\dots,p_d}(\boldsymbol{\theta})$ is a function of $\boldsymbol{\theta}$. For instance, if we want to specify the variables in (5.94), we must write $f^{(\nu)}_{\boldsymbol{p}}(\mathbf{x}, \boldsymbol{\theta}) = \boldsymbol{\nu}(K(\mathbf{x}) \circ P_{p_1,\dots,p_d}(\boldsymbol{\theta}))\boldsymbol{\nu}^T$. Note that the expressions of the symbols (5.94), (5.95) have a completely similar structure as the expressions of the differential problems (5.2), (5.24). This motivates the reformulation of (5.1) into the less common form (5.2).

In the case $d = 2$ we have

$$f^{(\nu)}_{\boldsymbol{p}} = [\nu_1 \ \nu_2] \left(K \circ P_{p_1,p_2} \right) [\nu_1 \ \nu_2]^T, \tag{5.97}$$

$$f^{(\nu)}_{\mathbf{G},\boldsymbol{p}} = [\nu_1 \ \nu_2] \left(K_{\mathbf{G}} \circ P_{p_1,p_2} \right) [\nu_1 \ \nu_2]^T,$$

where

$$P_{p_1,p_2} := \begin{bmatrix} f_{p_1} \otimes h_{p_2} & g_{p_1} \otimes g_{p_2} \\ g_{p_1} \otimes g_{p_2} & h_{p_1} \otimes f_{p_2} \end{bmatrix}. \tag{5.98}$$

Theorem 5.3. *The matrix P_{p_1,p_2} in (5.98) is SPSD over $[-\pi,\pi]^2$ and SPD for all (θ_1, θ_2) such that $\theta_1\theta_2 \neq 0$. Moreover, if K is SPD over $[0,1]^2$ then $K \circ P_{p_1,p_2}$ is SPSD over $[0,1]^2 \times [-\pi,\pi]^2$ and SPD if $\theta_1\theta_2 \neq 0$.*

Proof. It is clear that P_{p_1,p_2} is symmetric. Moreover, Lemmas 5.3 and 5.5 imply that

$$f_{p_1} \otimes h_{p_2} \geq 0, \quad h_{p_1} \otimes f_{p_2} \geq 0,$$

and if $\theta_1\theta_2 \neq 0$ then

$$f_{p_1} \otimes h_{p_2} > 0, \quad h_{p_1} \otimes f_{p_2} > 0.$$

In addition, Lemma 5.6 ensures that

$$\det(P_{p_1,p_2}) = f_{p_1}(\theta_1)h_{p_1}(\theta_1)h_{p_2}(\theta_2)f_{p_2}(\theta_2) - (g_{p_1}(\theta_1))^2(g_{p_2}(\theta_2))^2 \geq 0,$$

and if $\theta_1\theta_2 \neq 0$ then $\det(P_{p_1,p_2}) > 0$. Thus, P_{p_1,p_2} is SPSD over $[-\pi,\pi]^2$ and SPD if $\theta_1\theta_2 \neq 0$. Finally, by Lemma 1.6, if K is SPD over $[0,1]^2$ then $K \circ P_{p_1,p_2}$ is SPSD over $[0,1]^2 \times [-\pi,\pi]^2$ and SPD if $\theta_1\theta_2 \neq 0$. □

From (5.41) and (5.59) we also obtain the following factorization of the matrix P_{p_1,p_2}.

Theorem 5.4. *Let P_{p_1,p_2} be defined as in* (5.98)*, then*

$$P_{p_1,p_2} = S\,\hat{P}_{p_1,p_2}S,$$

with

$$S := \begin{bmatrix} 2\sin(\theta_1/2) & 0 \\ 0 & 2\sin(\theta_2/2) \end{bmatrix}, \qquad \hat{P}_{p_1,p_2} := \begin{bmatrix} h_{p_1-2}(\theta_1)h_{p_2}(\theta_2) & \hat{g}_{p_1}(\theta_1)\hat{g}_{p_2}(\theta_2) \\ \hat{g}_{p_1}(\theta_1)\hat{g}_{p_2}(\theta_2) & h_{p_1}(\theta_1)h_{p_2-2}(\theta_2) \end{bmatrix},$$

and $\hat{g}_p(\theta) := \dfrac{g_p(\theta)}{2\sin(\theta/2)}$*. Moreover,*

$$|2\sin(\theta/2)|^p \left(\frac{1}{|\theta|^p} - \frac{1}{\pi^p}\right) \leq |\hat{g}_p(\theta)| \leq |2\sin(\theta/2)|^p \frac{1}{|\theta|^p}.$$

The next theorem analyzes the zeros of the symbol (5.97).

Theorem 5.5. *If K is SPD over $[0,1]^2$, the symbol $f_{\boldsymbol{p}}^{(\nu)}$ in* (5.97) *is nonnegative over $[0,1]^2 \times [-\pi,\pi]^2$. Moreover, for any fixed $\mathbf{x} \in [0,1]^2$, $f_{\boldsymbol{p}}^{(\nu)}(\mathbf{x},\cdot)$ has a unique zero of order two at $(\theta_1,\theta_2) = (0,0)$ over $[-\pi,\pi]^2$.*

Proof. By Theorem 5.3, the matrix $K \circ P_{p_1,p_2}$ is SPSD over $[0,1]^2 \times [-\pi,\pi]^2$, so that the symbol $f_{\boldsymbol{p}}^{(\nu)}$ in (5.97) is nonnegative over $[0,1]^2 \times [-\pi,\pi]^2$ and it can vanish only if $[\nu_1\ \nu_2]^T$ is an eigenvector associated with a zero eigenvalue of $K \circ P_{p_1,p_2}$. From Theorem 5.3 we know that this can occur only if $\theta_1\theta_2 = 0$.

From Theorem 5.4 and the properties derived in Section 5.3, we may conclude that $\boldsymbol{\theta} := (\theta_1,\theta_2) = (0,0)$ is a zero of order two for $f_{\boldsymbol{p}}^{(\nu)}(\mathbf{x},\cdot)$, for any fixed $\mathbf{x} \in [0,1]^2$. Indeed, by taking the Taylor expansion around $\boldsymbol{\theta} = (0,0)$ of the matrix $\hat{P}_{p_1,p_2}$ in Theorem 5.4 we have

$$\hat{P}_{p_1,p_2}(\boldsymbol{\theta}) = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} + O(\|\boldsymbol{\theta}\|),$$

where $O(\|\boldsymbol{\theta}\|)$ is a 2×2 matrix whose components are bounded from above (in absolute value) by $C\|\boldsymbol{\theta}\|$ for some constant C independent of $\boldsymbol{\theta}$:

$$\left|[O(\|\boldsymbol{\theta}\|)]_{ij}\right| \leq C\|\boldsymbol{\theta}\|, \qquad i,j = 1,2.$$

Thus, we obtain

$$f_{\boldsymbol{p}}^{(\nu)}(\mathbf{x},\boldsymbol{\theta}) = [\nu_1\ \nu_2]\,S(\boldsymbol{\theta})K(\mathbf{x})S(\boldsymbol{\theta})\,[\nu_1\ \nu_2]^T + O(\|\boldsymbol{\theta}\|^3).$$

Since $K(\mathbf{x})$ is assumed to be SPD, we have $mI \leq K(\mathbf{x}) \leq MI$ for some constants $m, M > 0$,

$$m[\nu_1\ \nu_2][S(\boldsymbol{\theta})]^2[\nu_1\ \nu_2]^T + O(\|\boldsymbol{\theta}\|^3) \leq f_{\boldsymbol{p}}^{(\nu)}(\mathbf{x},\boldsymbol{\theta}) \leq M[\nu_1\ \nu_2][S(\boldsymbol{\theta})]^2[\nu_1\ \nu_2]^T + O(\|\boldsymbol{\theta}\|^3),$$

and it follows that the function $f_{\boldsymbol{p}}^{(\nu)}(\mathbf{x},\cdot)$ has a zero of order two at $\boldsymbol{\theta} = (0,0)$, like the function

$$[1 \;\; 1][S(\boldsymbol{\theta})]^2[1 \;\; 1]^T = 4\sin^2(\theta_1/2) + 4\sin^2(\theta_2/2) = (2 - 2\cos\theta_1) + (2 - 2\cos\theta_2).$$

Moreover, it is easy to check that the matrix $K(\mathbf{x}) \circ P_{p_1,p_2}(\boldsymbol{\theta})$ has a zero eigenvalue of multiplicity one if $\theta_1 = 0$ or $\theta_2 = 0$ but $(\theta_1,\theta_2) \neq (0,0)$. In the first case, the second component of the corresponding eigenvector is zero. In the second case, the first component of the corresponding eigenvector is zero. Since $\nu_1\nu_2 \neq 0$, it follows that, in both cases, $[\nu_1 \;\; \nu_2]^T$ cannot be an eigenvector associated with the zero eigenvalue of $K(\mathbf{x}) \circ P_{p_1,p_2}(\boldsymbol{\theta})$. Hence, $(\theta_1,\theta_2) = (0,0)$ is the unique zero of $f_{\boldsymbol{p}}^{(\nu)}(\mathbf{x},\cdot)$ over $[-\pi,\pi]^2$. □

Remark 5.5. Theorem 5.5 states that the symbol $f_{\boldsymbol{p}}^{(\nu)}$ in (5.97) has a unique (theoretical) zero at $(\theta_1,\theta_2) = (0,0)$, for any fixed $\mathbf{x} \in [0,1]^2$. However, other numerical zeros occur elsewhere for large $\boldsymbol{p} := (p_1,p_2)$. Indeed, from Lemmas 5.3–5.5 we see that, for large values of $\boldsymbol{p}$, all entries in the matrix P_{p_1,p_2} vanish numerically when $\theta_1 = \pi$ or $\theta_2 = \pi$. Therefore, for large $\boldsymbol{p}$, the symbol $f_{\boldsymbol{p}}^{(\nu)}$ in (5.97) has numerical zeros at the points $(\mathbf{x},\boldsymbol{\theta})$ such that $\theta_1 = \pi$ or $\theta_2 = \pi$.

We conclude by observing that the results in Theorem 5.5 for the symbol $f_{\boldsymbol{p}}^{(\nu)}$ immediately provide analogous results for the symbol $f_{\mathbf{G},\boldsymbol{p}}^{(\nu)}$, since the difference between these two symbols is only in the fact that K is replaced by $K_{\mathbf{G}}$. We also remark that Theorems 5.3 and 5.5 have been proved for the case $d = 2$, but, on the basis of our experience with 'spectral distributions and symbols', we are pretty sure that they can be extended to any dimensionality d.

Conjecture 5.1. *The matrix $P_{p_1,\ldots,p_d}$ in* (5.96) *is SPSD over $[-\pi,\pi]^d$ and SPD for all $(\theta_1,\ldots,\theta_d)$ such that $\theta_1\cdots\theta_d \neq 0$. Moreover, if K is SPD over $[0,1]^d$, the symbol $f_{\boldsymbol{p}}^{(\nu)}$ in* (5.94) *is nonnegative over $[0,1]^d \times [-\pi,\pi]^d$ and, for any fixed $\mathbf{x} \in [0,1]^d$, $f_{\boldsymbol{p}}^{(\nu)}(\mathbf{x},\cdot)$ has a unique zero of order two at $\boldsymbol{\theta} = \mathbf{0}$ over $[-\pi,\pi]^d$, like the function $\sum_{k=1}^{d}(2 - 2\cos\theta_k)$.*

Chapter 6

Fast iterative solvers for Galerkin B-spline IgA linear systems

In Chapter 4, we studied the spectral properties of the stiffness matrices $A_{\boldsymbol{n}}^{[\boldsymbol{p}]}$ coming from the Galerkin B-spline IgA approximation of the second-order elliptic problem

$$\begin{cases} -\Delta u + \boldsymbol{\beta} \cdot \nabla u + \gamma u = f & \text{in } \Omega := (0,1)^d, \\ u = 0 & \text{on } \partial\Omega, \end{cases} \tag{6.1}$$

where $f \in L^2(\Omega)$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_d) \in \mathbb{R}^d$ and $\gamma \geq 0$. In particular, we have computed the corresponding spectral symbol $f_{\boldsymbol{p}}^{(\boldsymbol{\nu})}$ (see Theorem 4.7). We will now exploit the properties of the symbol in order to design fast iterative solvers for linear systems with coefficient matrix $A_{\boldsymbol{n}}^{[\boldsymbol{p}]}$. Our ultimate goal is to design iterative algorithms with the following two properties. First, their computational cost is optimal, that is linear with respect to the matrix size; this property is essentially equivalent to requiring that their convergence rate (number of iterations for reaching a preassigned accuracy) is independent of the fineness parameters $\boldsymbol{n}$. Second, they are robust, i.e., their convergence rate is substantially independent of the spline degrees $\boldsymbol{p}$ associated with the IgA approximation order. Using carefully the properties of $f_{\boldsymbol{p}}^{(\boldsymbol{\nu})}$, we will succeed in designing a multi-iterative multigrid method, whose convergence rate will prove to be optimal and robust at the same time. This multi-iterative solver involves the PCG/PGMRES as a smoother at the finest level, where the related preconditioner is chosen as the Toeplitz matrix generated by a specific function coming from a certain factorization of the symbol. The properties of the symbol will be used also to explain the behavior of classical multigrid methods, whose convergence rate is optimal (independent of the fineness parameters $\boldsymbol{n}$) but not robust, because it deteriorates when the spline degrees $\boldsymbol{p}$ increase.

Before starting, let us recall from Chapter 4 that the stiffness matrix $A_{\boldsymbol{n}}^{[\boldsymbol{p}]}$ coming from the Galerkin B-spline IgA approximation of the second-order elliptic problem (6.1) is given explicitly by

$$\begin{aligned} A_{\boldsymbol{n}}^{[\boldsymbol{p}]} := & \sum_{k=1}^{d} \frac{1}{n_1} M_{n_1}^{[p_1]} \otimes \cdots \otimes \frac{1}{n_{k-1}} M_{n_{k-1}}^{[p_{k-1}]} \otimes n_k K_{n_k}^{[p_k]} \otimes \frac{1}{n_{k+1}} M_{n_{k+1}}^{[p_{k+1}]} \otimes \cdots \otimes \frac{1}{n_d} M_{n_d}^{[p_d]} \\ & + \sum_{k=1}^{d} \beta_k \frac{1}{n_1} M_{n_1}^{[p_1]} \otimes \cdots \otimes \frac{1}{n_{k-1}} M_{n_{k-1}}^{[p_{k-1}]} \otimes H_{n_k}^{[p_k]} \otimes \frac{1}{n_{k+1}} M_{n_{k+1}}^{[p_{k+1}]} \otimes \cdots \otimes \frac{1}{n_d} M_{n_d}^{[p_d]} \\ & + \gamma \frac{1}{n_1} M_{n_1}^{[p_1]} \otimes \cdots \otimes \frac{1}{n_d} M_{n_d}^{[p_d]}, \end{aligned} \tag{6.2}$$

where $K_n^{[p]}$, $H_n^{[p]}$, $M_n^{[p]}$ are defined in (4.41). Moreover, the spectral symbol $f_{\boldsymbol{p}}^{(\boldsymbol{\nu})} : [-\pi, \pi]^d \to \mathbb{R}$ of the normalized matrix-sequence $\{n^{d-2} A_{\boldsymbol{n}}^{(\boldsymbol{\nu})}\}_n$, with $\boldsymbol{n} = \boldsymbol{\nu} n$ and $\boldsymbol{\nu} \in \mathbb{Q}_+^d$, is

$$f_{\boldsymbol{p}}^{(\boldsymbol{\nu})}(\boldsymbol{\theta}) = \sum_{k=1}^{d} c_k(\boldsymbol{\nu}) \left(h_{p_1} \otimes \cdots \otimes h_{p_{k-1}} \otimes f_{p_k} \otimes h_{p_{k+1}} \otimes \cdots \otimes h_{p_d} \right)(\boldsymbol{\theta}) = \sum_{k=1}^{d} c_k(\boldsymbol{\nu}) f_{p_k}(\theta_k) \prod_{\substack{j=1 \\ j \neq k}}^{d} h_{p_j}(\theta_j), \tag{6.3}$$

where $c_k(\boldsymbol{\nu}) := \frac{\nu_k}{\nu_1 \cdots \nu_{k-1}\nu_{k+1}\cdots\nu_d}$; see (4.46)–(4.47) for the definition of f_p and h_p. We note that $f_{\boldsymbol{p}}^{(\boldsymbol{\nu})}(\boldsymbol{\theta})$ is symmetric in each variable θ_j, because both f_p and h_p are even functions. Hence, $f_{\boldsymbol{p}}^{(\boldsymbol{\nu})}$ restricted to $[0,\pi]^d$ is also a symbol for $\{n^{d-2}A_{\boldsymbol{n}}^{[\boldsymbol{p}]}\}_n$ (this follows directly from Definition 1.1). The symbol $f_{\boldsymbol{p}}^{(\boldsymbol{\nu})}$ is independent of $\boldsymbol{\beta}$ and γ, and possesses the following properties, which are consequences of Lemmas 4.4–4.5.

Lemma 6.1. *We have*

$$\left(\frac{4}{\pi^2}\right)^{\sum_{j=1}^d p_j+d-1} \min_{j=1,\ldots,d} c_j(\boldsymbol{\nu}) \sum_{k=1}^d (2-2\cos\theta_k) \le f_{\boldsymbol{p}}^{(\boldsymbol{\nu})}(\boldsymbol{\theta}) \le \max_{j=1,\ldots,d} c_j(\boldsymbol{\nu}) \sum_{k=1}^d (2-2\cos\theta_k).$$

Moreover, setting $M_{f_{\boldsymbol{p}}^{(\boldsymbol{\nu})}} := \max_{\boldsymbol{\theta}\in[0,\pi]^d} f_{\boldsymbol{p}}^{(\boldsymbol{\nu})}(\boldsymbol{\theta})$*, for all* $j = 1,\ldots,d$ *we have*

$$f_{\boldsymbol{p}}^{(\boldsymbol{\nu})}(\theta_1,\ldots,\theta_{j-1},\pi,\theta_{j+1},\ldots,\theta_d) \le \frac{1}{2^{p_j-2}} f_{\boldsymbol{p}}^{(\boldsymbol{\nu})}(\theta_1,\ldots,\theta_{j-1},\tfrac{\pi}{2},\theta_{j+1},\ldots,\theta_d) \le \frac{1}{2^{p_j-2}} M_{f_{\boldsymbol{p}}^{(\boldsymbol{\nu})}}.$$

By Lemma 6.1, the normalized symbol $f_{\boldsymbol{p}}^{(\boldsymbol{\nu})}/M_{f_{\boldsymbol{p}}^{(\boldsymbol{\nu})}}$ has only one actual zero of order two at $\boldsymbol{\theta} = \mathbf{0}$, like the function $\sum_{k=1}^d (2-2\cos\theta_k)$. However, when the spline degrees p_j are large, it also has infinitely many 'numerical zeros' over $[0,\pi]^d$, located at the 'π-edge points'

$$\{\boldsymbol{\theta}\in[0,\pi]^d : \exists j \in \{1,\ldots,d\} \text{ with } \theta_j = \pi\}. \tag{6.4}$$

Because of this unpleasant property, the classical multigrid schemes for the matrix $n^{d-2}A_{\boldsymbol{n}}^{[\boldsymbol{p}]}$ that we shall see in later sections show a bad (though optimal) convergence rate when one of the p_j is large. In practice, their convergence rate is optimal, because it is independent of the fineness parameters $\boldsymbol{n}$ and so it does not increase when the matrix size grows, but it is also non-robust, because it rapidly worsens when the spline degrees $\boldsymbol{p}$ grow. We will see that this lack of robustness in classical multigrid methods:

(a) is due to the fact that they ignore the numerical zeros of the normalized symbol $f_{\boldsymbol{p}}^{(\boldsymbol{\nu})}/M_{f_{\boldsymbol{p}}^{(\boldsymbol{\nu})}}$ located at the π-edge points (6.4);

(b) can be bypassed by adopting a multi-iterative multigrid strategy that involves the PCG/PGMRES as a smoother at the finest level, with a properly chosen preconditioner which takes into account the numerical zeros (6.4).

6.1 How to use the symbol? A basic guide to the user

In order to provide explanations for the non-robustness of classical multigrid methods, as well as to design the winning multi-iterative multigrid solver mentioned above, this section is of fundamental importance. What we are going to see in this section is the heuristic information that can be extracted from the symbol f of a given matrix-sequence $\{Z_n\}$ and that provides a guideline in understanding/predicting the convergence features of the various iterative solvers applied to Z_n. We mainly focus our attention on a perturbed Toeplitz setting (i.e., on the case where Z_n is a 'small' perturbation of a Toeplitz matrix), because our IgA matrices $n^{d-2}A_{\boldsymbol{n}}^{[\boldsymbol{p}]}$ are indeed 'small' perturbations of the Toeplitz matrices $T_{\boldsymbol{n}+\boldsymbol{p}-\mathbf{2}}(f_{\boldsymbol{p}}^{(\boldsymbol{\nu})})$ associated with the symbol $f_{\boldsymbol{p}}^{(\boldsymbol{\nu})}$. To see this, we recall from the proof of Theorem 4.7 that, fixed $\boldsymbol{n} = \boldsymbol{\nu}n$, $n^{d-2}A_{\boldsymbol{n}}^{[\boldsymbol{p}]}$ is equal to its 'Toeplitz part' $T_{\boldsymbol{n}+\boldsymbol{p}-\mathbf{2}}(f_{\boldsymbol{p}}^{(\boldsymbol{\nu})})$ plus a correction Y_n whose trace-norm $|||Y_n|||_1$ is $o(N(\boldsymbol{n}+\boldsymbol{p}-\mathbf{2}))$ when $n\to\infty$, where $N(\boldsymbol{n}+\boldsymbol{p}-\mathbf{2})$ is the dimension of $A_{\boldsymbol{n}}^{[\boldsymbol{p}]}$. This allows us to conclude that $n^{d-2}A_{\boldsymbol{n}}^{[\boldsymbol{p}]}$ coincides with a d-level Toeplitz matrix, namely $T_{\boldsymbol{n}+\boldsymbol{p}-\mathbf{2}}(f_{\boldsymbol{p}}^{(\boldsymbol{\nu})})$, up to a 'small' correction Y_n whose trace-norm $|||Y_n|||_1$ is negligible with respect to the matrix size $N(\boldsymbol{n}+\boldsymbol{p}-\mathbf{2})$. This result was actually the key to prove that $\{n^{d-2}A_{\boldsymbol{n}}^{[\boldsymbol{p}]}\}_n$ has the same symbol $f_{\boldsymbol{p}}^{(\boldsymbol{\nu})}$ of the Toeplitz sequence $\{T_{\boldsymbol{n}+\boldsymbol{p}-\mathbf{2}}(f_{\boldsymbol{p}}^{(\boldsymbol{\nu})})\}_n$.

6.1.1 Counting the eigenvalues belonging to a given interval

The starting point of our reasoning is Definition 1.1 and, especially, the subsequent Remark 1.2. Let $a < b$, let $\{Z_n\}$ be a sequence of Hermitian matrices, with Z_n of size d_n tending to infinity, and assume that $\{Z_n\} \sim_\lambda f$, where $f : D \subseteq \mathbb{R}^d \to \mathbb{R}$. Moreover, let $E_n([a,b])$ be the number of eigenvalues of Z_n belonging to the interval $[a,b]$. Then, relation (1.20), or, more precisely, its 'scalar version' (1.22), implies that

$$E_n([a,b]) = I[a,b]d_n + o(d_n) \tag{6.5}$$

with

$$I[a,b] := \frac{m_d(\{\boldsymbol{\theta} \in D : \ f(\boldsymbol{\theta}) \in [a,b]\})}{m_d(D)},$$

if

$$0 = m_d(\{\boldsymbol{\theta} \in D : \ f(\boldsymbol{\theta}) = a\}) = m_d(\{\boldsymbol{\theta} \in D : \ f(\boldsymbol{\theta}) = b\}). \tag{6.6}$$

Regarding the hypothesis (6.6), we observe that it is never violated when f is a non-constant trigonometric polynomial. Moreover, it can be shown that, for a general measurable function f, it can be violated only for countably many values of a and b.

The expression of the error term $o(d_n)$ can be better estimated under specific circumstances. For example, if $d = 1$, $Z_n = T_{d_n}(f)$, and $f : [-\pi,\pi] \to \mathbb{R}$ is a real-valued trigonometric polynomial, the error term $o(d_n)$ can be replaced by a constant which depends linearly on the degree of f (this can be deduced by using Cauchy interlacing arguments; see [57]). The same holds for the univariate IgA matrices $\frac{1}{n}A_n^{[p]}$ obtained from (6.2) for $d = 1$ and $\beta = \gamma = 0$, because they are constant rank corrections of the Toeplitz matrices $T_{n+p-2}(f_p)$, where the rank of the correction is proportional to p; see (4.83)–(4.84).

Formula (6.5) is of interest, e.g., when $a = 0$ and $b = \epsilon \ll 1$, for having a good guess of the size of the eigenspace related to small positive eigenvalues $\lambda \leq \epsilon \ll 1$. In fact, if Z_n is HPD, this eigenspace is the so-called ill-conditioned subspace, which is responsible for the ill-conditioning of the matrix and for the slow convergence of general purpose iterative solvers.

6.1.2 Eigenvectors vs. frequencies in a perturbed Toeplitz setting

This subsection is the most interesting from the viewpoint of designing fast iterative solvers for matrices Z_n such that the sequence $\{Z_n\}$ is distributed like a certain symbol $f : D \subseteq \mathbb{R}^d \to \mathbb{R}$. For the sake of simplicity, and also for the purposes of this chapter, we can restrict our attention to the case where $D = [-\pi,\pi]^d$ and $\{Z_n\} \in \mathcal{T}$. Here and in the following, $\mathcal{T}$ is the set of matrix-sequences $\{Z_n\}$ of the form

$$Z_n = \sum_{i=1}^{r} \prod_{j=1}^{q_r} T_{\boldsymbol{m}(n)}(g_{ij}) + Y_n, \tag{6.7}$$

where $r, q_1, \ldots, q_r \in \mathbb{N}$, $\{Y_n\} \sim_\lambda 0$ is a zero-distributed sequence (see Definition 1.1), and every $\{T_{\boldsymbol{m}(n)}(g_{ij})\}_n$, with $\boldsymbol{m}(n) = (m_1(n), \ldots, m_d(n))$, is a sequence of d-level Toeplitz matrices such that $\boldsymbol{m}(n) \to \infty$ as $n \to \infty$. Note that a matrix-sequence $\{Z_n\} \in \mathcal{T}$ is the sum of a sequence $\{\sum_{i=1}^{r} \prod_{j=1}^{q_r} T_{\boldsymbol{m}(n)}(g_{ij})\}$ belonging to the algebra generated by Toeplitz sequences and of a zero-distributed sequence $\{Y_n\}$. In particular, the sequence of IgA matrices $\{n^{d-2}A_{\boldsymbol{n}}^{[\boldsymbol{p}]}\}_n$ ($\boldsymbol{n} = \boldsymbol{\nu}n$) is of the form (6.7) with $r = 1$, $q_1 = 1$, $\boldsymbol{m}(n) = \boldsymbol{n} + \boldsymbol{p} - \boldsymbol{2}$, $T_{\boldsymbol{m}(n)}(g_{11}) = T_{\boldsymbol{n}+\boldsymbol{p}-\boldsymbol{2}}(f_{\boldsymbol{p}}^{(\boldsymbol{\nu})})$ and $Y_n = n^{d-2}A_{\boldsymbol{n}}^{[\boldsymbol{p}]} - T_{\boldsymbol{n}+\boldsymbol{p}-\boldsymbol{2}}(f_{\boldsymbol{p}}^{(\boldsymbol{\nu})})$. Note that, as mentioned at the beginning of this section, the sequence of corrections $\{Y_n\}$ satisfy $|||Y_n|||_1 = o(N(\boldsymbol{n} + \boldsymbol{p} - \boldsymbol{2}))$, and so in particular it is zero-distributed. Indeed, by Weyl's majorant theorem [7, Theorem II.3.6], it holds in general that, if $\{Y_n\}$ is a matrix-sequence, with Y_n of size d_n tending to infinity, and if $|||Y_n|||_1 = o(d_n)$, then $\{Y_n\} \sim_\lambda 0$.

For matrix-sequences $\{Z_n\}$ of the form (6.7), a lot can be said, in terms of (Fourier) frequences, concerning the approximate structure of the eigenspaces of Z_n. Roughly speaking, let $d = 1$ and let $f : [-\pi,\pi] \to \mathbb{R}$ be a

continuous even function, such that it is the symbol of a sequence $\{Z_n\} \in \mathcal{T}$, with Z_n of size d_n. Note that f restricted to $[0,\pi]$ is also a symbol for $\{Z_n\}$ (because f is even). Then, the eigenvalues $\lambda_j(Z_n)$, $j = 1,\dots,d_n$, behave like the uniform sampling of f over $[0,\pi]$ given by

$$f\left(\frac{j\pi}{d_n+1}\right), \quad j = 1,\dots,d_n,$$

and the related eigenvectors behave like the following set of frequency vectors:

$$\mathbf{v}_j^{(d_n)} := \left(\sin\left(\frac{jk\pi}{d_n+1}\right)\right)_{k=1}^{d_n}, \quad j = 1,\dots,d_n.$$

The statement above is quite vague, but it can be made more precise without using technicalities (see [10, 72] for a rigorous analysis). In any case, what the reader should keep in mind is the following: if $\{Z_n\} \in \mathcal{T}$, with Z_n of size d_n, and if $\{Z_n\} \sim_\lambda f : [-\pi,\pi] \to \mathbb{R}$, with f a continuous even function, then we may think about the matrix Z_n as if it were the matrix

$$\tau_{d_n}(f) := \mathcal{S}_{d_n}\left[\operatorname*{diag}_{j=1,\dots,d_n} f\left(\frac{j\pi}{d_n+1}\right)\right]\mathcal{S}_{d_n}, \tag{6.8}$$

where

$$\mathcal{S}_{d_n} := \sqrt{\frac{2}{d_n+1}}\left[\sin\left(\frac{jk\pi}{d_n+1}\right)\right]_{j,k=1}^{d_n} = \sqrt{\frac{2}{d_n+1}}\left[\mathbf{v}_1^{(d_n)}|\mathbf{v}_2^{(d_n)}|\cdots|\mathbf{v}_{d_n}^{(d_n)}\right] \tag{6.9}$$

is a real symmetric unitary matrix, the so-called sine transform. The matrix in (6.8) is called the (unilevel) τ-matrix of order d_n associated with the function f; see, e.g., [25, Definition 2.1] and the references reported in [25]. If Z_n is Hermitian and we are interested in the eigenvectors of Z_n associated with the eigenvalues in the interval $[a,b]$, then we know from Subsection 6.1.1 that the subspace generated by these eigenvectors has dimension $I[a,b]d_n + o(d_n)$ and it is approximately described by

$$\operatorname{span}\left\{\mathbf{v}_j^{(d_n)} : \ f\left(\frac{j\pi}{d_n+1}\right) \in [a,b]\right\}. \tag{6.10}$$

From the relation above, and taking into account that the vectors $\mathbf{v}_j^{(d_n)}$ corresponding to large (small) indices j are referred to as high (low) frequences, it can be seen that a zero of the symbol f at $\theta = 0$ implies that the ill-conditioned subspace of Z_n is related to low frequencies, while a zero of the symbol at π implies that the ill-conditioned subspace is related to high frequencies.

If $d > 1$, proper tensor-like arguments show that the same conclusions hold. To be a little more precise, assume that:

- $\{Z_n\} \in \mathcal{T}$ is a sequence of d-level matrices as in (6.7), with Z_n of dimension $d_n = N(\boldsymbol{m}(n))$ and partial dimensions $m_1(n),\dots,m_d(n)$ tending to infinity;
- $\{Z_n\} \sim_\lambda f$, where $f : [-\pi,\pi]^d \to \mathbb{R}$ is continuous and symmetric in each variable, in the sense that

  $$f(\pm\theta_1,\pm\theta_2,\dots,\pm\theta_d) = f(\theta_1,\theta_2,\dots,\theta_d)$$

  for all $(\theta_1,\theta_2,\dots,\theta_d) \in [-\pi,\pi]^d$.

Then, f restricted to $[0,\pi]^d$ is also a symbol for $\{Z_n\}$ and we may think about the matrix Z_n as if it were the d-level τ-matrix

$$\tau_{\boldsymbol{m}(n)}(f) := \mathcal{S}_{\boldsymbol{m}(n)}\left[\operatorname*{diag}_{\boldsymbol{j}=1,\dots,\boldsymbol{m}(n)} f\left(\frac{\boldsymbol{j}\pi}{\boldsymbol{m}(n)+\mathbf{1}}\right)\right]\mathcal{S}_{\boldsymbol{m}(n)},$$

where

$$\mathcal{S}_{\boldsymbol{m}(n)} := \mathcal{S}_{m_1(n)} \otimes \cdots \otimes \mathcal{S}_{m_d(n)}$$

is the d-level sine transform; refer again to [25] for the definitions. For instance, if $d = 2$, Z_n is Hermitian and we are interested in the eigenvectors of Z_n associated with the eigenvalues in the interval $[a, b]$, then we know that the subspace generated by these eigenvectors has dimension $I[a, b]d_n + o(d_n)$ and it is approximately described by

$$\operatorname{span}\left\{ \mathbf{v}_{j_1}^{(m_1(n))} \otimes \mathbf{v}_{j_2}^{(m_2(n))} : \ f\left(\frac{j_1\pi}{m_1(n)+1}, \frac{j_2\pi}{m_2(n)+1}\right) \in [a, b] \right\}. \tag{6.11}$$

6.2 Iterative solvers and the multi-iterative approach

In this section, we review some basic iterative methods that we will use in order to build up a fast iterative solver for our IgA stiffness matrices $A_{\boldsymbol{n}}^{[\boldsymbol{p}]}$ in (6.2). In particular, we consider:

1. classical stationary iterative methods (Richardson, Gauss-Seidel, the weighted versions, etc. [71]);
2. the PCG method [4];
3. two-grid, V-cycle, W-cycle methods [51, 69];
4. multi-iterative techniques [56].

We will present them in view of the multi-iterative approach [56], which is a way of combining different (basic) iterative methods having complementary spectral behavior. We anticipate that the optimal and robust multi-iterative multigrid solver for the IgA stiffness matrices $A_{\boldsymbol{n}}^{[\boldsymbol{p}]}$, that has been mentioned at the beginning of this chapter and that we are going to design in the following, is just a combination of basic iterative methods in a unique multigrid algorithm, in the spirit of the multi-iterative idea. We first explain and discuss the main idea of the multi-iterative approach in Subsections 6.2.1–6.2.3. Then, in Subsections 6.2.4–6.2.5, we focus on two-grid and multigrid methods, as well as on the PCG method, in our IgA context, and we shall see how to combine them in a unique optimal and robust multi-iterative multigrid solver for $A_{\boldsymbol{n}}^{[\boldsymbol{p}]}$.

6.2.1 Unity makes strength: the multi-iterative approach

Stationary iterative methods for solving a linear system $A\mathbf{u} = \mathbf{b}$ (with $A \in \mathbb{R}^{m\times m}$) can be written in the general form

$$\mathbf{u}^{(k+1)} = \mathbf{u}^{(k)} + M^{-1}(\mathbf{b} - A\mathbf{u}^{(k)}), \qquad k = 0, 1, \ldots, \tag{6.12}$$

where M is chosen as an approximation of A such that a linear system with matrix M is easily solvable. In this way, M^{-1} in (6.12) can be regarded as an approximation of A^{-1} and the vector $M^{-1}(\mathbf{b} - A\mathbf{u}^{(k)})$ can be easily computed. By defining the iteration matrix $S := I - M^{-1}A$, we can reformulate the stationary iteration (6.12) as

$$\mathbf{u}^{(k+1)} = \mathcal{S}(\mathbf{u}^{(k)}) := S\,\mathbf{u}^{(k)} + (I - S)A^{-1}\mathbf{b}. \tag{6.13}$$

The error $\mathbf{e}^{(k+1)} := A^{-1}\mathbf{b} - \mathbf{u}^{(k+1)}$ is then given by $\mathbf{e}^{(k+1)} = S\,\mathbf{e}^{(k)} = S^{k+1}\mathbf{e}^{(0)}$, and its norm is quickly reduced if $\|S\|$ is much smaller than one.

Two specific examples of stationary iterations (6.13), of interest later on, are the relaxed Richardson method and the relaxed Gauss-Seidel method, whose corresponding iteration matrices are, respectively,

$$\overline{S} := I - \omega A, \tag{6.14}$$

$$\hat{S} := I - \left(\frac{1}{\omega}D - L\right)^{-1} A. \tag{6.15}$$

In both cases, $\omega \in \mathbb{R}$ is the relaxation parameter, while D and L are the matrices coming from the splitting of A associated with the Gauss-Seidel method: D is the diagonal part of A and $-L$ is the lower triangular part of A, excluding the diagonal elements.

Let us now consider l different (invertible) approximations of A, say M_i, $i = 1, \dots, l$, and then l iterative methods with iteration matrices $S_i := I - M_i^{-1}A$, $i = 1, \dots, l$. The following multi-iterative scheme can then be defined [56]:

$$\begin{aligned} \mathbf{u}^{(k,1)} &= S_1\mathbf{u}^{(k)} + \mathbf{b}_1, \\ \mathbf{u}^{(k,2)} &= S_2\mathbf{u}^{(k,1)} + \mathbf{b}_2, \\ &\vdots \\ \mathbf{u}^{(k+1)} &= S_l\mathbf{u}^{(k,l-1)} + \mathbf{b}_l, \end{aligned} \tag{6.16}$$

where $\mathbf{b}_i := M_i^{-1}\mathbf{b}$. Hence,

$$\mathbf{u}^{(k+1)} = S_l S_{l-1} \cdots S_2 S_1 \mathbf{u}^{(k)} + \mathbf{c}, \tag{6.17}$$

where

$$\mathbf{c} = \mathbf{b}_l + S_l\mathbf{b}_{l-1} + S_l S_{l-1}\mathbf{b}_{l-2} + \cdots + S_l S_{l-1} \cdots S_2\mathbf{b}_1.$$

The errors $\mathbf{e}^{(k)} := A^{-1}\mathbf{b} - \mathbf{u}^{(k)}$ and $\mathbf{e}^{(k,i)} := A^{-1}\mathbf{b} - \mathbf{u}^{(k,i)}$, $k \geq 0$, $i = 1, \dots, l-1$, are such that

$$\begin{aligned} \mathbf{e}^{(k,i)} &= S_i \cdots S_2 S_1 \mathbf{e}^{(k)}, \\ \mathbf{e}^{(k+1)} &= S_l \cdots S_2 S_1 \mathbf{e}^{(k)}. \end{aligned}$$

If S_i is highly contractive in a subspace $\mathscr{H}_i$ and if $S_{i-1}(\mathscr{L}_{i-1}) \subset \mathscr{H}_i$, where $\mathscr{L}_{i-1}$ is another subspace where S_{i-1} reduces slowly the norm of the error, then $\|S_i S_{i-1}\|$ can be much smaller than $\|S_i\|\,\|S_{i-1}\|$. This implies that multi-iterative methods can be fast, even when the basic iteration matrices have norms close to one, or even when the basic iterations are non-convergent.

The multi-iterative idea, briefly outlined above, can be extended to include non-stationary iterations. For instance, we may replace the last iteration in (6.16) with a single step (or a few steps) by the PCG method. In this case, the overall iteration (6.17) is not stationary, but the scheme (6.16) is still called a multi-iterative method. As we shall see later, our optimal and robust multi-iterative multigrid solver for the IgA stiffness matrices $A_{\boldsymbol{n}}^{[\boldsymbol{p}]}$ will involve the PCG (or the PGMRES) inside a multi-iterative scheme of multigrid type.

6.2.2 Two-grid and multigrid methods in a multi-iterative perspective

Consider again the linear system $A\mathbf{u} = \mathbf{b}$, $A \in \mathbb{R}^{m \times m}$. Suppose we have two stationary iterative methods (the smoothers) as in (6.13) for the solution of the linear system, and a full-rank matrix (the projector) $P \in \mathbb{R}^{l \times m}$, $l \leq m$. Then, the corresponding two-grid method for solving the linear system is given by the following algorithm.

Algorithm 6.1. Given an approximation $\mathbf{u}^{(k)}$ to the solution $\mathbf{u} = A^{-1}\mathbf{b}$, the new approximation $\mathbf{u}^{(k+1)}$ is obtained by applying ν_{pre} steps of pre-smoothing as in (6.13) with iteration matrix S_{pre}, a coarse-grid correction, and ν_{post} steps of post-smoothing as in (6.13) with iteration matrix S_{post}, as follows:

1. apply ν_{pre} steps of pre-smoothing: $\mathbf{u}^{(k,1)} = \mathcal{S}_{\text{pre}}^{\nu_{\text{pre}}}(\mathbf{u}^{(k)}) = S_{\text{pre}}^{\nu_{\text{pre}}}\mathbf{u}^{(k)} + (I - S_{\text{pre}}^{\nu_{\text{pre}}})A^{-1}\mathbf{b}$;
2. compute the residual: $\mathbf{r} = \mathbf{b} - A\mathbf{u}^{(k,1)}$;
3. project the residual: $\mathbf{r}^{(c)} = P\mathbf{r}$;
4. compute the correction: $\mathbf{e}^{(c)} = (PAP^T)^{-1}\mathbf{r}^{(c)}$;

5. extend the correction: $\mathbf{e} = P^T \mathbf{e}^{(c)}$;

6. correct the initial approximation: $\mathbf{u}^{(k,2)} = \mathbf{u}^{(k,1)} + \mathbf{e}$;

7. apply ν_{post} steps of post-smoothing: $\mathbf{u}^{(k+1)} = \mathcal{S}_{\text{post}}^{\nu_{\text{post}}}(\mathbf{u}^{(k,2)}) = S_{\text{post}}^{\nu_{\text{post}}}\mathbf{u}^{(k)} + (I - S_{\text{post}}^{\nu_{\text{post}}})A^{-1}\mathbf{b}$.

Steps 2–6 in Algorithm 6.1 define the so-called coarse-grid correction, which is a standard non-convergent iterative method with iteration matrix

$$CGC := I - P^T(PAP^T)^{-1}PA. \tag{6.18}$$

The iteration matrix of the two-grid scheme is denoted by $TG(S_{\text{pre}}^{\nu_{\text{pre}}}, S_{\text{post}}^{\nu_{\text{post}}}, P)$ and is explicitly given by

$$TG(S_{\text{pre}}^{\nu_{\text{pre}}}, S_{\text{post}}^{\nu_{\text{post}}}, P) = S_{\text{post}}^{\nu_{\text{post}}} \cdot CGC \cdot S_{\text{pre}}^{\nu_{\text{pre}}}.$$

When the pre-smoothing is not present, i.e., $\nu_{\text{pre}} = 0$, the two-grid iteration matrix is denoted by $TG(S_{\text{post}}^{\nu_{\text{post}}}, P)$.

We point out that two-grid (and multigrid) methods can be written in the general multi-iterative form (6.16), in which $l = 2$ or $l = 3$. In this case, S_1 is the pre-smoothing operator, S_2 is the coarse-grid operator, and S_3 is the post-smoothing operator. Interestingly enough, we observe that $\|S_2\| \geq 1$ because the spectral radius of S_2 is equal to 1 (see [51]), while S_1 and S_3 are usually weakly contractive. However, as we will see later in Subsection 6.3.1, there are examples in which the best contraction factor of the whole multi-iterative two-grid scheme is achieved by choosing a non-convergent smoother. Therefore, it may happen that a very fast multi-iterative method is obtained by combining basic iterations that are all slowly convergent or even non-convergent.

6.2.3 Multi-iterative solvers vs. spectral distributions

The main idea of the multi-iterative approach is to choose the different iteration matrices S_i, $i = 1, \ldots, l$, in the scheme (6.16) such that they have a complementary spectral behavior. Let us assume that S_i is highly contractive in a subspace $\mathscr{H}_i$, and weakly (or not) contractive in the complementary subspace $\mathscr{L}_i$. Then, the recipe for designing fast multi-iterative solvers is to choose the iteration matrices S_i such that

$$\bigoplus_{i=1}^{l} \mathscr{H}_i = \mathbb{C}^m.$$

This recipe is aesthetically beautiful and appealing, but totally unpractical if we are unable to identify l pairs of subspaces $(\mathscr{H}_i, \mathscr{L}_i)$, $i = 1, \ldots, l$, with the properties described above and such that $\mathscr{H}_i \oplus \mathscr{L}_i = \mathbb{C}^m$. However, our IgA stiffness matrices $n^{d-2}A_{\boldsymbol{n}}^{[\boldsymbol{p}]}$ can be considered as 'small' perturbations of Toeplitz matrices (see the discussion at the beginning of Section 6.1), and so Subsections 6.1.1–6.1.2 can provide an heuristic guide in identifying such subspaces in terms of frequencies and estimating their dimensions.

Let us now illustrate this concept in the case where the d-dimensional Laplacian problem $-\Delta u = f$ over $[0,1]^d$ is approximated by standard centered Finite Differences (FD). The resulting discretization matrices have a pure d-level Toeplitz structure with corresponding generating function (symmetric in each variable) given by

$$f(\boldsymbol{\theta}) = f_{\text{FD}}(\boldsymbol{\theta}) := \sum_{j=1}^{d}(2 - 2\cos\theta_j), \quad \boldsymbol{\theta} \in [-\pi, \pi]^d. \tag{6.19}$$

More precisely, if the discretization step in each direction x_i, $i = 1, \ldots, d$, is $1/n$, the resulting discretization matrix is $T_{\boldsymbol{m}}(f)$, where $\boldsymbol{m} = (n-1, \ldots, n-1)$.

Now consider the following multigrid method in the framework of multi-iterative solvers. It is composed of three iterations ($l = 3$): a pre-smoothing given by the Richardson method (6.13)–(6.14) with parameter ω_{pre}

(iteration matrix S_1), a coarse-grid correction with iteration matrix S_2 defined in (6.18), and a post-smoothing given by the Richardson method with parameter ω_{post} (iteration matrix S_3). The coarse-grid iteration S_2, which uses as projector P the traditional full-weighting restriction (6.20)–(6.21), is designed in such a way that the related iteration is not convergent globally, but strongly reduces the error in low frequencies. Now, $S_1 = I - \omega_{\text{pre}} T_{\boldsymbol{m}}(f) = T_{\boldsymbol{m}}(1 - \omega_{\text{pre}} f)$ and $S_3 = I - \omega_{\text{post}} T_{\boldsymbol{m}}(f) = T_{\boldsymbol{m}}(1 - \omega_{\text{post}} f)$. If we choose $\omega_{\text{pre}} = \|f\|_\infty^{-1} = (4d)^{-1}$, the symbol of the iteration matrix S_1 is equal to $1 - f/\|f\|_\infty$, which is maximal at $\boldsymbol{\theta} = (0, \ldots, 0)$ and attains its minimum at $\boldsymbol{\theta} = (\pi, \ldots, \pi)$. As a consequence, the pre-smoothing iteration is highly convergent (contractive) in the high frequencies and slowly convergent in the low frequencies. In fact, if we consider the two-grid (and also the related V-cycle multigrid) with the latter coarse-grid correction operator and the latter pre-smoother, then we already obtain an optimal method (see [62, 65]), even though the two basic iterations S_1 and S_2 are very slow or non-convergent. However, at this point, we have understood the machinery, and hence, if we desire to accelerate further the global multi-iterative method, then we can consider a post-smoothing iteration which may be slowly convergent both in the very low and very high frequencies but very fast in a space of 'intermediate' frequencies. The choice is obtained by setting $\omega_{\text{post}} = 2\|f\|_\infty^{-1}$ so that $S_3 = T_{\boldsymbol{m}}(1 - 2f/\|f\|_\infty)$. It is interesting to remark that the symbol $|1 - 2f(\boldsymbol{\theta})/\|f\|_\infty|$ evaluated at $\boldsymbol{\theta} = (0, \ldots, 0)$ and $\boldsymbol{\theta} = (\pi, \ldots, \pi)$ is equal to 1. Therefore, the method is slowly convergent both in high and low frequencies, since the moduli of the eigenvalues of S_3 are close to 1. However, the symbol is very small in absolute value in regions of $[0, \pi]^d$ associated with intermediate frequencies, corresponding to values of $\boldsymbol{\theta}$ near $\boldsymbol{\theta} = (\frac{\pi}{2}, \ldots, \frac{\pi}{2})$. Hence, S_3 is highly convergent in the subspace generated by these frequencies. The resulting multi-iterative method is indeed extremely fast, as shown in [65]. We will use these guiding ideas in our choice of the solvers for the IgA matrices.

6.2.4 Choice of the projector in our two-grid and multigrid methods

We now look for an appropriate projector P in the coarse-grid correction (6.18) in order to address our specific IgA linear systems. Since our IgA stiffness matrices $n^{d-2} A_{\boldsymbol{n}}^{[\boldsymbol{p}]}$ can be considered as 'small' perturbations of d-level Toeplitz matrices $T_{\boldsymbol{n}+\boldsymbol{p}-\boldsymbol{2}}(f_{\boldsymbol{p}}^{(\nu)})$ (see the discussion at the beginning of Section 6.1), we follow the approach in [62] and focus on a particular projector $P = P_{\boldsymbol{n}+\boldsymbol{p}-\boldsymbol{2}}$ which is appropriate for $T_{\boldsymbol{n}+\boldsymbol{p}-\boldsymbol{2}}(f_{\boldsymbol{p}}^{(\nu)})$. More specifically, for any odd $m \geq 3$, denote by U_m the cutting matrix of size $\frac{m-1}{2} \times m$ given by

$$U_m := \begin{bmatrix} 0 & 1 & & & & & 0 \\ & & 0 & 1 & & & 0 \\ & & & & \ddots & & \vdots \\ & & & & & 0 \; 1 & 0 \end{bmatrix} \in \mathbb{R}^{\frac{m-1}{2} \times m}.$$

Then, for any $\boldsymbol{m} \in \mathbb{N}^d$ with odd $m_1, \ldots, m_d \geq 3$, we define $U_{\boldsymbol{m}} := U_{m_1} \otimes \cdots \otimes U_{m_d}$ and we set

$$P_{\boldsymbol{m}} := U_{\boldsymbol{m}} \cdot T_{\boldsymbol{m}}(q_d), \qquad q_d(\theta_1, \ldots, \theta_d) := \prod_{j=1}^{d} (1 + \cos\theta_j). \tag{6.20}$$

It can be shown that $P_{\boldsymbol{m}}$ admits a 'recursive expression'

$$P_{\boldsymbol{m}} = \bigotimes_{j=1}^{d} P_{m_j}, \qquad P_{m_j} = \bigotimes_{j=1}^{d} U_{m_j} \cdot T_{m_j}(q) = \bigotimes_{j=1}^{d} \frac{1}{2} \underbrace{\begin{bmatrix} 1 & 2 & 1 & & & \\ & & 1 & 2 & 1 & \\ & & & & \ddots & \\ & & & & & 1 \; 2 \; 1 \end{bmatrix}}_{m_j}, \qquad q(\theta) = 1 + \cos\theta. \tag{6.21}$$

From (6.21), we see that $P_{\boldsymbol{m}}$ is the traditional full-weighting restriction, which has full rank $\prod_{j=1}^{d} \frac{m_j-1}{2}$, being the Kronecker product of full-rank matrices. The projector $P_{\boldsymbol{m}}$ leads to a coarse-grid correction (6.18) which is highly contractive in the subspace of low frequencies.

Let us now consider $d = 1$ and our specific linear systems, with coefficient matrix $\frac{1}{n}A_n^{[p]}$. The symbol associated to the sequence of univariate IgA stiffness matrices

$$\frac{1}{n}A_n^{[p]} = K_n^{[p]} + \frac{\beta}{n}H_n^{[p]} + \frac{\gamma}{n^2}M_n^{[p]}, \qquad n = 1, 2, \ldots \tag{6.22}$$

is $f_p(\theta)$, as defined in (4.46) (see Remark 4.5). Since $\theta = 0$ is the only zero of the symbol, we expect (see e.g. [62, 1, 23]) that the classical full-weighting projector P_m combined with any classical smoother (Richardson, Gauss-Seidel, Conjugate Gradient, GMRES) leads to two-grid, V-cycle, and W-cycle algorithms with an optimal convergence rate, independent of the matrix size and of the fineness parameter n. However, for large p, a numerical zero occurs at $\theta = \pi$ for the normalized symbol $f_p(\theta)/M_{f_p}$; see the discussion after Lemma 4.5. The projector P_m, as well as the aforementioned classical smoothers, are not designed for coping with this numerical zero, which represents a source of ill-conditioning in high frequencies of our matrices $\frac{1}{n}A_n^{[p]}$. Therefore, we can predict that the traditional projector P_m combined with any classical smoother will lead to two-grid (and multigrid) algorithms with convergence rate that, despite being independent of n, worsens with p. These forecasts are numerically confirmed in Section 6.3 and theoretically motivated in [25, Section 4], where it is shown that the p-worsening of the convergence rate is actually expected to be exponential in p, due to the fact that $f_p(\pi)/M_{f_p} \to 0$ exponentially (see Lemma 4.5).

If $d \geq 2$ and we consider our specific linear systems, with coefficient matrix $n^{d-2}A_{\boldsymbol{n}}^{[\boldsymbol{p}]}$ ($\boldsymbol{n} = \boldsymbol{\nu}n$), the situation is even worse than in the case $d = 1$, because of the specific analytic features of the symbol $f_{\boldsymbol{p}}^{(\boldsymbol{\nu})}(\boldsymbol{\theta})$ associated with the sequence $\{n^{d-2}A_{\boldsymbol{n}}^{[\boldsymbol{p}]}\}_n$; see Lemma 6.1 and the discussion following it. Since $\boldsymbol{\theta} = \mathbf{0}$ is the only zero of the symbol, we know (see e.g. [1, 23]) that the projector $P_{\boldsymbol{m}}$ combined with any classical smoother (Richardson, Gauss-Seidel, Conjugate Gradient) will lead to two-grid, V-cycle and W-cycle algorithms with an optimal convergence rate, independent of the fineness parameters $\boldsymbol{n}$. However, for large $\boldsymbol{p}$, infinitely (sic!) many numerical zeros of $f_{\boldsymbol{p}}^{(\boldsymbol{\nu})}/M_{f_{\boldsymbol{p}}^{(\boldsymbol{\nu})}}$ occur at the π-edge points (6.4). Thus, as in the one-dimensional setting, the traditional projector $P_{\boldsymbol{m}}$ (with any classical smoother) leads to multigrid algorithms having a convergence rate that deteriorates with $\boldsymbol{p}$. This means that, when combining the standard full-weighting projector $P_{\boldsymbol{m}}$ with any classical smoother such as Richardson, Gauss-Seidel, Conjugate Gradient, GMRES and so on, the resulting two-grid and multigrid algorithms will have a convergence rate that, despite being optimal ($\boldsymbol{n}$-independent), is not robust in $\boldsymbol{p}$.

In order to overcome this problem of classical two-grid and multigrid schemes, the suggestion coming from the multi-iterative idea is in this case the following: keep the full-weighting projector $P_{\boldsymbol{m}}$ for dealing with the actual zero of the symbol $f_{\boldsymbol{p}}^{(\boldsymbol{\nu})}$ located at the origin $\boldsymbol{\theta} = \mathbf{0}$, and replace all classical smoothers with a PCG or a PGMRES method, whose preconditioner takes care of the numerical zeros of the normalized symbol $f_{\boldsymbol{p}}^{(\boldsymbol{\nu})}/M_{f_{\boldsymbol{p}}^{(\boldsymbol{\nu})}}$ located at the π-edge points (6.4). Of course, this is only a vague idea, that should be made more clear. We will do it in the next subsection.

6.2.5 PCG with p-independent convergence rate

Let us start with recalling the PCG method for solving the linear system $A\mathbf{u} = \mathbf{b}$ with A a real SPD matrix. Since we consider the preconditioned version of the CG method, we assume to have an SPD matrix M such that M is an approximation of A and such that a linear system with matrix M is easily solvable.

Algorithm 6.2. Let $\mathbf{u}^{(k)}$ be a given approximation of the solution $\mathbf{u} = A^{-1}\mathbf{b}$, with A a real SPD matrix, and let M be a SPD approximation of A. Then, the new approximation $\mathbf{u}^{(k+1)}$ is obtained as follows:

1. compute the approximation: $\mathbf{u}^{(k+1)} = \mathbf{u}^{(k)} + \alpha^{(k)}\mathbf{p}^{(k)}$, using the optimal step length $\alpha^{(k)} = \dfrac{\mathbf{r}^{(k)\,T}\mathbf{z}^{(k)}}{\mathbf{p}^{(k)\,T}A\mathbf{p}^{(k)}}$;

2. compute the residual: $\mathbf{r}^{(k+1)} = \mathbf{b} - A\mathbf{u}^{(k+1)} = \mathbf{r}^{(k)} - \alpha^{(k)} A\mathbf{p}^{(k)}$;

3. compute the preconditioned residual: $\mathbf{z}^{(k+1)} = M^{-1}\mathbf{r}^{(k+1)}$;

4. compute the A-conjugate search direction: $\mathbf{p}^{(k+1)} = \mathbf{z}^{(k+1)} + \beta^{(k)}\mathbf{p}^{(k)}$, with $\beta^{(k)} = \dfrac{\mathbf{z}^{(k+1)\,T}\mathbf{r}^{(k+1)}}{\mathbf{z}^{(k)\,T}\mathbf{r}^{(k)}}$.

If the vectors $\mathbf{r}^{(k)}$, $\mathbf{z}^{(k)}$, $\mathbf{p}^{(k)}$ are not yet computed by the algorithm in a previous step, then we initialize them as $\mathbf{r}^{(k)} = \mathbf{b} - A\mathbf{u}^{(k)}$, $\mathbf{z}^{(k)} = M^{-1}\mathbf{r}^{(k)}$, $\mathbf{p}^{(k)} = \mathbf{z}^{(k)}$.

Now let us assume that $\boldsymbol{\beta} = \mathbf{0}$ in our model problem (6.1). Under this assumption, we know that $A_{\boldsymbol{n}}^{[\boldsymbol{p}]}$ is SPD (see (6.2) and recall that $K_n^{[p]}$, $M_n^{[p]}$ are SPD), and so the PCG method can be applied.

Remark 6.1. If $\boldsymbol{\beta} \neq \mathbf{0}$, the matrix $A_{\boldsymbol{n}}^{[\boldsymbol{p}]}$ is not symmetric and we cannot apply to it the PCG method. In such a case, we simply suggest to replace the PCG with the PGMRES (using for the PGMRES the same preconditioner that we are going to design for the PCG).

In this subsection, we focus on the construction of a preconditioner such that the PCG applied to our matrix $n^{d-2}A_{\boldsymbol{n}}^{[\boldsymbol{p}]}$ will be $\boldsymbol{p}$-independent. The idea of a $\boldsymbol{p}$-independent PCG method has its theoretical foundation in the spectral results concerning Toeplitz systems with Toeplitz preconditioners [22, 58], and in the study of the specific symbol $f_{\boldsymbol{p}}^{(\boldsymbol{\nu})}$ of our matrix-sequence $\{n^{d-2}A_{\boldsymbol{n}}^{[\boldsymbol{p}]}\}_n$, $\boldsymbol{n} = \boldsymbol{\nu} n$.

Let h be a nonnegative, a.e. nonzero and Lebesgue integrable function over $[-\pi,\pi]^d$. Then $T_{\boldsymbol{m}}(h)$ is a HPD d-level Toeplitz matrix; see Theorem 1.8. Let f be a real-valued and Lebesgue integrable function over $[-\pi,\pi]^d$, so that $T_{\boldsymbol{m}}(f)$ is a Hermitian matrix. By following [22, 58], we know that all the eigenvalues of $T_{\boldsymbol{m}}^{-1}(h)T_{\boldsymbol{m}}(f)$ belong to the set $[r,R]$, with $r = \operatorname{ess\,inf} f/h$, $R = \operatorname{ess\,sup} f/h$, and

$$\left\{T_{\boldsymbol{m}}^{-1}(h)T_{\boldsymbol{m}}(f)\right\} \sim_\lambda f/h.$$

For $d = 1$, the symbol of $\{\frac{1}{n}A_n^{[p]}\}$ is $f_p(\theta) = (2 - 2\cos\theta)h_{p-1}(\theta)$. Since $\frac{1}{n}A_n^{[p]}$ is a 'small' perturbation of $T_{n+p-2}(f_p)$, it can be shown that

$$\left\{T_{n+p-2}^{-1}(h_{p-1})\frac{1}{n}A_n^{[p]}\right\} \sim_\lambda f_p/h_{p-1} = 2 - 2\cos\theta,$$

which is the symbol of the standard FD approximation given in (6.19) for $d = 1$, and is indeed p-independent. Hence, if we apply to $\frac{1}{n}A_n^{[p]}$ the PCG with $T_{n+p-2}(h_{p-1})$ as preconditioner, we expect to have a p-independent method. Unfortunately, it is not optimal, because it is slowly convergent when the fineness parameter n is large (see Table 6.5 for a numerical example); this is due to the fact that $2 - 2\cos\theta$ has a zero at $\theta = 0$. However, in view of the multi-iterative approach, we can build a totally robust method as follows: we consider a basic coarse-grid operator with projector P_{n+p-2} as in (6.20)–(6.21) working in the low frequencies (like in the case of a standard FD approximation), and we include the PCG method, with preconditioner $T_{n+p-2}(h_{p-1})$, in the smoothing strategy. Thus, the coarse-grid operator will be responsible for the optimality of the method (a convergence speed independent of the fineness parameter n) and the PCG-smoother will bring the p-independence, taking care of the numerical zero of f_p/M_{f_p} at π for large p. In conclusion, the global multi-iterative method will be optimal in n and robust in p at the same time, while the standard coarse-grid correction alone is not convergent and the PCG method alone is p-independent, but slowly convergent when n is large (see Section 6.3 for numerical illustrations).

The good news is that the above idea can be generalized to any dimensionality d. Indeed, for $d \geq 2$, thanks to Lemmas 4.4–4.5, the symbol $f_{\boldsymbol{p}}^{(\boldsymbol{\nu})}$ in (6.3) can be factored as follows:

$$f_{\boldsymbol{p}}^{(\boldsymbol{\nu})}(\boldsymbol{\theta}) = \prod_{j=1}^{d} h_{p_j-1}(\theta_j)\left[\sum_{k=1}^{d} c_k(\boldsymbol{\nu})(2 - 2\cos(\theta_k)) \prod_{\substack{j=1 \\ j\neq k}}^{d} w_{p_j}(\theta_j)\right], \tag{6.23}$$

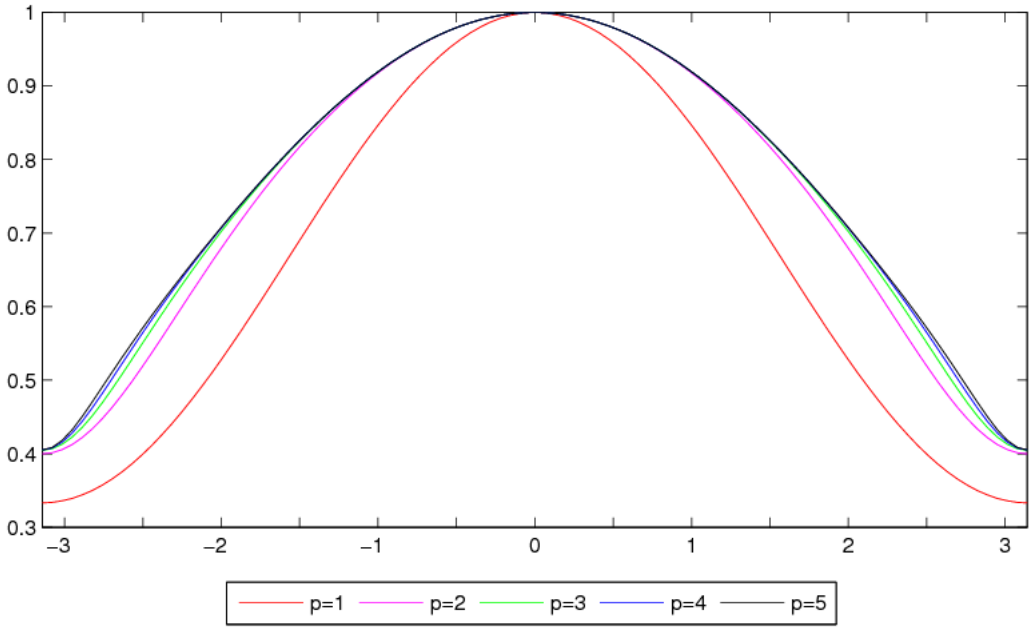


Figure 6.1: graph of $w_p = h_p/h_{p-1}$ for $p = 1, \ldots, 5$.

where $w_p(\theta) := h_p(\theta)/h_{p-1}(\theta)$ is a function well-separated from zero, uniformly with respect to $\theta \in [0, \pi]$ and with respect to $p \geq 1$. In short, w_p is p-independent, in the sense that it is bounded from above and below by two positive constants independent of p. Actually, w_p seems to converge uniformly to some function with range in $[0.4, 1]$; see Figure 6.1. This means that the function between square brackets in (6.23) does not have numerical zeros and only has an actual zero at $\boldsymbol{\theta} = \mathbf{0}$. This zero does not create problems to our two-grid schemes, because the standard coarse-grid correction (6.18) with classical full-weighting projector (6.20)–(6.21) takes care of it. Therefore, the function $\prod_{j=1}^{d} h_{p_j-1}(\theta_j)$ is responsible for the existence of numerical zeros at the π-edge points (6.4) when the p_j's are large. Thus, the same function causes the poor behavior of our two-grid and multigrid schemes, with any classical smoother, when the p_j's are large. We then consider for our matrices $n^{d-2}A_{\boldsymbol{n}}^{[\boldsymbol{p}]}$ the following preconditioner:

$$T_{\boldsymbol{n}+\boldsymbol{p}-\mathbf{2}}\left(\prod_{j=1}^{d} h_{p_j-1}(\theta_j)\right) = T_{\boldsymbol{n}+\boldsymbol{p}-\mathbf{2}}(h_{p_1-1} \otimes \cdots \otimes h_{p_d-1}) = T_{n_1+p_1-2}(h_{p_1-1}) \otimes \cdots \otimes T_{n_d+p_d-2}(h_{p_d-1}). \tag{6.24}$$

Note that the matrix (6.24) is a 'small' perturbation of the (normalized) B-spline mass matrix $M_{n_1+1}^{[p_1-1]} \otimes \cdots \otimes M_{n_d+1}^{[p_d-1]}$ related to the fineness parameters $\boldsymbol{n} + \mathbf{1}$ and the spline degrees $\boldsymbol{p} - \mathbf{1}$; see (4.85)–(4.86) and recall (1.18). The choice of using a PCG method with preconditioner (6.24) as a smoother is made in order to 'erase' all the numerical zeros at the π-edge points (6.4). Due to (6.23) and to the fact that

$$\left(\prod_{j=1}^{d} h_{p_j-1}(\theta_j)\right)^{-1} f_{\boldsymbol{p}}^{(\nu)}(\boldsymbol{\theta}) = \sum_{k=1}^{d} c_k(\boldsymbol{\nu})(2 - 2\cos(\theta_k)) \prod_{\substack{j=1 \\ j\neq k}}^{d} w_{p_j}(\theta_j)$$

is $\boldsymbol{p}$-independent, the PCG with preconditioner (6.24) turns out to have a $\boldsymbol{p}$-robust convergence rate for our matrices in the d-dimensional setting. For a numerical illustration we refer to Tables 6.5, 6.10 and 6.14 for $d = 1, 2, 3$. Note from these tables that, for fixed n, the number of iterations only slightly increases with p. On the other hand, we clearly observe the bad dependence on n, as expected.

At this point, it is important to stress that the proposed preconditioner (6.24) is effectively solvable: due to the tensor-product structure and to the bandedness of the matrices $T_{n_i+p_i-2}(h_{p_i-1})$, the computational cost

for solving a linear system with matrix (6.24) is linear in the matrix size $N(\boldsymbol{n}+\boldsymbol{p}-\boldsymbol{2})$. Let us illustrate this in the case $d=2$, for a general tensor product $X\otimes Y$ of two invertible matrices $X\in\mathbb{C}^{m\times m}$ and $Y\in\mathbb{C}^{\ell\times\ell}$. By the properties of the Kronecker product, it holds that

$$(X\otimes Y)^{-1}=X^{-1}\otimes Y^{-1}.$$

Let $\mathbf{b}:=\mathrm{vec}(B)\in\mathbb{R}^{m\ell}$ be the vector obtained by stacking the columns of $B\in\mathbb{R}^{m\times\ell}$, where vec denotes the stacking operator. Then, the linear system

$$(X\otimes Y)\mathbf{u}=\mathbf{b} \tag{6.25}$$

can be solved by

$$\mathbf{u}=(X^{-1}\otimes Y^{-1})\mathbf{b}=\mathrm{vec}(Y^{-1}BX^{-T}).$$

This requires to solve m linear systems with matrix Y, plus ℓ linear systems with matrix X; see [40, Lemma 4.3.1]. If X and Y are banded, like the Toeplitz matrices $T_{n_i+p_i-2}(h_{p_i-1})$ in (6.24), then the cost for solving a linear system with matrix X or Y is linear in the matrix size, and so the overall cost for solving (6.25) is linear in the matrix size $m\ell$. Of course, this trick applies to the preconditioner (6.24) but not to the matrix $n^{d-2}A_{\boldsymbol{n}}^{[\boldsymbol{p}]}$, which consists of sums of tensor-product matrices; see (6.2).

Summarizing, in the spirit of the multi-iterative approach, our proposal for solving a linear system with coefficient matrix $n^{d-2}A_{\boldsymbol{n}}^{[\boldsymbol{p}]}$, $\boldsymbol{n}=\boldsymbol{\nu}n$, is as follows:

- as a solver, we use a two-grid, or a V-cycle/W-cycle multigrid, using (at each level) the standard coarse-grid correction with classical full-weighting projector (6.21). This basic coarse-grid operator is very effective in low frequencies, and it is all we need if we had to deal with a symbol like (6.19), coming from the standard FD approximation;

- since our normalized symbol $f_{\boldsymbol{p}}^{(\boldsymbol{\nu})}/M_{f_{\boldsymbol{p}}^{(\boldsymbol{\nu})}}$, for large $\boldsymbol{p}$, shows numerical zeros at the π-edge points (6.4), we include the PCG method with preconditioner (6.24) in the smoothing strategy. In particular, we will use it at the finest level.

In this way, the coarse-grid operator will be responsible for the optimality of the method (a convergence speed independent of the fineness parameters $\boldsymbol{n}$), while the chosen PCG-smoother will induce the $\boldsymbol{p}$-independence, taking care of the numerical zeros at the π-edge points (6.4). The global multi-iterative method is expected to be optimal and robust at the same time, meaning that its convergence rate should be independent of both $\boldsymbol{n}$ and $\boldsymbol{p}$. As we shall see from the numerical experiments, the convergence rate will be independent of $\boldsymbol{n}$, substantially independent of $\boldsymbol{p}$, and, surprisingly enough, substantially independent also of the dimensionality d.

Remark 6.2. From the discussion above (in the 1D case), one could guess that the PCG method with preconditioner $T_{n+p-2}(f_p)$ is substantially robust both with respect to n and p, because $\{T_{n+p-2}^{-1}(f_p)\frac{1}{n}A_n^{[p]}\}\sim_\lambda f_p/f_p=1$. This is numerically illustrated in Table 6.6. Unfortunately, this naive choice is not so practical, because it cannot be effectively generalized to the higher dimensional setting. For example, in the 2D case, the PCG method with preconditioner $T_{n_1+p_1-2}(f_{p_1})\otimes T_{n_2+p_2-2}(f_{p_2})$ does not work (see Table 6.11). The explanation is clear: the function $f_{p_1}\otimes f_{p_2}$ and the symbol of our 2D matrices $f_{p_1,p_2}^{(\nu_1,\nu_2)}$ possess two sets of zeros with a completely different structure. On the other hand, the use of $T_{n_1+p_1-2,n_2+p_2-2}(f_{p_1,p_2}^{(\nu_1,\nu_2)})$ as a possible preconditioner is also unsuccessful, because its cost is prohibitive due to the lack of the tensor-product structure in the preconditioner.

6.3 Two-grid algorithms and their performances: 1D

We start with a careful testing of the standard two-grid methods with the classical full-weighting projector (6.20)–(6.21) and with different combinations of the traditional smoothers. We note that the V-cycle and W-cycle convergence cannot be better than the one of the two-grid method. Then, we proceed with the full multi-iterative approach, sketched in Subsection 6.2.5, involving the PCG method as a smoother.

6.3.1 Classical two-grid methods

Let us illustrate the performances of standard two-grid methods with the classical projector $P_n^{[p]} := P_{n+p-2}$ given in (6.20)–(6.21), which induces a coarse-grid correction effective in the low frequencies. We only consider two-grid methods without pre-smoothing steps and with a single post-smoothing step.

Table 6.1 shows the results of the numerical experiments for $TG(\overline{S}_n^{[p]}, P_n^{[p]})$, with $\overline{S}_n^{[p]}$ being the iteration matrix of the relaxed Richardson method with parameter $\omega^{[p]}$; see (6.14). In problem (6.1), we fixed $d = 1$, $\beta = \gamma = 0$, so that $\frac{1}{n}A_n^{[p]} = K_n^{[p]}$ and $\overline{S}_n^{[p]} = I - \omega^{[p]}K_n^{[p]}$. Then, for $p = 1, \ldots, 6$ we determined experimentally the best Richardson parameter $\omega^{[p]}$, in the sense that $\omega^{[p]}$ minimizes $\overline{\rho}_n^{[p]} := \rho(TG(\overline{S}_n^{[p]}, P_n^{[p]}))$ with $n = 2560$ (if p is odd) and $n = 2561$ (if p is even) among all $\omega \in \mathbb{R}$ with at most four nonzero decimal digits after the comma. We note that the choice $\omega^{[1]} = 1/3$ has a theoretical motivation, because it imposes a fast convergence both in high and intermediate frequencies. Finally, we computed the spectral radii $\overline{\rho}_n^{[p]}$ for increasing values of n. In all the considered experiments, the proposed two-grid scheme is optimal. Moreover, as $n \to \infty$, $\overline{\rho}_n^{[p]}$ converges to a limit $\overline{\rho}_\infty^{[p]}$, which is minimal not for $p = 1$ but for $p = 2$. A theoretical explanation of this phenomenon is given in [25]. When p increases from 2 to 6, we observe that $\overline{\rho}_\infty^{[p]}$ increases as well. In view of the theoretical interpretation based on the symbol f_p given in [25], which passes through the identification of $K_n^{[p]}$ with the τ-matrix $\tau_{n+p-2}(f_p)$ (see Subsection 6.1.2), $\overline{\rho}_\infty^{[p]}$ is expected to converge exponentially to 1 as $p \to \infty$, and in fact, even for moderate values of p such as $p = 5, 6$, we see from Table 6.1 that the value $\overline{\rho}_\infty^{[p]}$ is not satisfactory. This 'exponentially poor' behavior can be related to the fact that $f_p(\pi)/M_{f_p}$ exponentially approaches 0 when p increases (see Lemma 4.5, Figure 4.3, Table 4.1). Finally, from some numerical experiments we observe that $\rho(K_n^{[4]}) \approx 1.8372$, $\forall n \geq 15$. Therefore, the best parameter $\omega^{[4]} = 1.2229$ produces a non-convergent smoother $\overline{S}_n^{[4]} = I - 1.2229\, K_n^{[4]}$ having $\rho(\overline{S}_n^{[4]}) \approx 1.2467$. This shows that the two-grid scheme can be convergent even when the smoother $\overline{S}_n^{[p]}$ is not and, moreover, $\overline{\rho}_n^{[p]}$ can attain its minimum at a value of $\omega^{[p]}$ for which $\rho(\overline{S}_n^{[p]}) > 1$, according to the multi-iterative idea (see Section 6.2).

Table 6.2 illustrates the behavior of $TG(\hat{S}_n^{[p]}, P_n^{[p]})$ in the case $\beta = \gamma = 0$, for $p = 1, \ldots, 6$, with $\hat{S}_n^{[p]}$ being the iteration matrix of the relaxed Gauss-Seidel method for $A = K_n^{[p]}$; see (6.15). Like in Table 6.1, the relaxation parameter $\omega^{[p]}$ was chosen so as to minimize $\hat{\rho}_n^{[p]} := \rho(TG(\hat{S}_n^{[p]}, P_n^{[p]}))$ with $n = 2560$ (if p is odd) and $n = 2561$ (if p is even) among all $\omega \in \mathbb{R}$ with at most four nonzero decimal digits after the comma. It follows from Table 6.2 that, except for the particular case $p = 2$, the use of the Gauss-Seidel smoother improves the convergence rate of the two-grid. However, we also observe that $\hat{\rho}_n^{[p]}$ presents the same dependence on p as $\overline{\rho}_n^{[p]}$: the scheme is optimal, but its asymptotic convergence rate attains its minimum for $p = 2$ and then worsens as p increases from 2 to 6. As explained in the discussion after Lemma 6.1 and in Subsection 6.2.4, we know that such a worsening is an intrinsic feature of the problem and is related to the fact that $f_p(\pi)/M_{f_p}$ converges exponentially to 0 for increasing p. In other words, the normalized symbol f_p/M_{f_p} shows a numerical zero at π, inducing an ill-conditioning in the high frequencies, where our coarse-grid operator is not effective and all the considered smoothers (Richardson and Gauss-Seidel) are weakly contractive.

The rapid worsening of the convergence rate with p is well illustrated in Table 6.3, where we fixed $n = 320$ (if p is odd) or $n = 321$ (if p is even), and we computed for increasing p the spectral radii $\overline{\rho}_n^{[p]}$

n	$\overline{\rho}_n^{[1]}$ $[\omega^{[1]} = 1/3]$	$\overline{\rho}_n^{[3]}$ $[\omega^{[3]} = 1.0368]$	$\overline{\rho}_n^{[5]}$ $[\omega^{[5]} = 1.2576]$
80	0.3333333	0.4479733	0.8927544
160	0.3333333	0.4474586	0.8926293
320	0.3333333	0.4472015	0.8925948
640	0.3333333	0.4470729	0.8925948
1280	0.3333333	0.4470366	0.8925948
2560	0.3333333	0.4470391	0.8925948
n	$\overline{\rho}_n^{[2]}$ $[\omega^{[2]} = 0.7311]$	$\overline{\rho}_n^{[4]}$ $[\omega^{[4]} = 1.2229]$	$\overline{\rho}_n^{[6]}$ $[\omega^{[6]} = 1.2235]$
81	0.0257459	0.7373412	0.9596516
161	0.0254342	0.7371979	0.9595077
321	0.0252866	0.7371256	0.9594351
641	0.0252153	0.7371016	0.9593993
1281	0.0252000	0.7371016	0.9593993
2561	0.0252000	0.7371016	0.9593993

Table 6.1: values of $\overline{\rho}_n^{[p]} := \rho(TG(\overline{S}_n^{[p]}, P_n^{[p]}))$ in the case $\beta = \gamma = 0$, for the specified parameter $\omega^{[p]}$.

n	$\hat{\rho}_n^{[1]}$ $[\omega^{[1]} = 0.9065]$	$\hat{\rho}_n^{[3]}$ $[\omega^{[3]} = 0.9483]$	$\hat{\rho}_n^{[5]}$ $[\omega^{[5]} = 1.1999]$
80	0.1762977	0.1486937	0.4279346
160	0.1771878	0.1534242	0.4491173
320	0.1956301	0.1567792	0.4628558
640	0.2228058	0.1589204	0.4710180
1280	0.2358223	0.1602392	0.4758293
2560	0.2416926	0.1609750	0.4786945
n	$\hat{\rho}_n^{[2]}$ $[\omega^{[2]} = 0.9109]$	$\hat{\rho}_n^{[4]}$ $[\omega^{[4]} = 1.0602]$	$\hat{\rho}_n^{[6]}$ $[\omega^{[6]} = 1.3292]$
81	0.0648736	0.2972510	0.5631940
161	0.0648736	0.3110761	0.5852798
321	0.0648736	0.3201033	0.6002364
641	0.0648736	0.3255332	0.6104147
1281	0.0648736	0.3286511	0.6164439
2561	0.0649656	0.3304592	0.6197837

Table 6.2: values of $\hat{\rho}_n^{[p]} := \rho(TG(\hat{S}_n^{[p]}, P_n^{[p]}))$ in the case $\beta = \gamma = 0$, for the specified parameter $\omega^{[p]}$.

p	1	2	3	4	5	6	7	8	9	10	11	12
$\overline{\rho}_n^{[p]}$	0.3333	0.0253	0.4471	0.7371	0.8926	0.9594	0.9853	0.9947	0.9981	0.9994	0.9998	0.9999
$\hat{\rho}_n^{[p]}$	0.1941	0.0639	0.1567	0.3156	0.4608	0.5990	0.7173	0.8114	0.8813	0.9289	0.9627	0.9818

Table 6.3: values of $\overline{\rho}_n^{[p]}$ and $\hat{\rho}_n^{[p]}$ in the case $\beta = \gamma = 0$, corresponding to the optimal parameters $\omega^{[p]}$ with four nonzero decimal digits after the comma. We fixed $n = 320$ (if p is odd) or $n = 321$ (if p is even).

n	$\overline{c}_n^{[1]}[1/3]$	$\hat{c}_n^{[1]}[0.9065]$	$\overline{c}_n^{[3]}[1.0368]$	$\hat{c}_n^{[3]}[0.9483]$	$\overline{c}_n^{[5]}[1.2576]$	$\hat{c}_n^{[5]}[1.1999]$
80	17	14	24	10	164	22
160	17	14	25	10	166	23
320	18	14	25	10	169	23
640	18	14	25	11	172	24
1280	18	14	26	11	175	24
2560	18	14	26	11	178	25
n	$\overline{c}_n^{[2]}[0.7311]$	$\hat{c}_n^{[2]}[0.9109]$	$\overline{c}_n^{[4]}[1.2229]$	$\hat{c}_n^{[4]}[1.0602]$	$\overline{c}_n^{[6]}[1.2235]$	$\hat{c}_n^{[6]}[1.3292]$
81	6	7	62	15	456	32
161	6	8	62	16	460	33
321	6	8	63	16	467	33
641	6	8	64	16	475	34
1281	6	8	66	17	483	35
2561	6	8	67	17	492	36

Table 6.4: number of iterations $\overline{c}_n^{[p]}$ and $\hat{c}_n^{[p]}$ needed by $TG(\overline{S}_n^{[p]}, P_n^{[p]})$ and $TG(\hat{S}_n^{[p]}, P_n^{[p]})$ respectively, for solving $\frac{1}{n}A_n^{[p]}\mathbf{u} = \mathbf{f}$ with $\beta = -5$, $\gamma = 1$, $f = 1$, up to a precision of 10^{-8}. The methods have been started with $\mathbf{u}^{(0)} = \mathbf{0}$. The parameter $\omega^{[p]}$ is specified between brackets $[\cdot]$.

and $\hat{\rho}_n^{[p]}$ obtained with the best parameters $\omega^{[p]}$ among all $\omega \in \mathbb{R}$ with four nonzero decimal digits after the comma.

We now compare $TG(\overline{S}_n^{[p]}, P_n^{[p]})$ and $TG(\hat{S}_n^{[p]}, P_n^{[p]})$ on the linear system $\frac{1}{n}A_n^{[p]}\mathbf{u} = \mathbf{f}$, coming from the B-spline Galerkin approximation of the model problem (6.1) in the case $d = 1$, with $\beta = -5$, $\gamma = 1$ and $f = 1$. In Table 6.4, the considered linear system was solved for $p = 1, \ldots, 6$ and for increasing values of n by means of $TG(\overline{S}_n^{[p]}, P_n^{[p]})$ (with $\omega^{[p]}$ as in Table 6.1) and $TG(\hat{S}_n^{[p]}, P_n^{[p]})$ (with $\omega^{[p]}$ as in Table 6.2). For each pair (p, n), $\overline{c}_n^{[p]}$ and $\hat{c}_n^{[p]}$ are, respectively, the number of iterations needed by $TG(\overline{S}_n^{[p]}, P_n^{[p]})$ and $TG(\hat{S}_n^{[p]}, P_n^{[p]})$, both started with initial guess $\mathbf{u}^{(0)} = \mathbf{0}$, to compute a vector $\mathbf{u}^{(c)}$ whose relative residual in the 2-norm is less than 10^{-8}, i.e.,

$$\left\| \mathbf{f} - \frac{1}{n}A_n^{[p]}\mathbf{u}^{(c)} \right\| \leq 10^{-8}\|\mathbf{f}\|. \tag{6.26}$$

6.3.2 Multi-iterative two-grid method with PCG as smoother

Despite their optimality, the basic two-grid schemes $TG(\overline{S}_n^{[p]}, P_n^{[p]})$ and $TG(\hat{S}_n^{[p]}, P_n^{[p]})$ suffer from a 'pathology', because, as already discussed, their convergence rate rapidly worsens when p increases. To overcome this problem, we follow the multi-iterative idea outlined in Subsection 6.2.5 and we replace, in the two-grid Algorithm 6.1, the smoothers $\overline{S}_n^{[p]}$ and $\hat{S}_n^{[p]}$ with the PCG method, whose preconditioner $T_{n+p-2}(h_{p-1})$ takes care of dampening the high frequencies corresponding to values of θ near π.

We first illustrate the PCG method (see Algorithm 6.2) applied to the linear system $\frac{1}{n}A_n^{[p]}\mathbf{u} = \mathbf{f}$, coming from the B-spline Galerkin approximation of the model problem (6.1) in the case $d = 1$, with $\beta = 0$, $\gamma = 1$ and $f = 1$. Table 6.5 reports the number of iterations needed by the PCG method with preconditioner $T_{n+p-2}(h_{p-1})$ to compute a vector $\mathbf{u}^{(c)}$ satisfying a relative residual less than 10^{-8}; see (6.26). We observe that the PCG method is essentially p-independent, but slowly convergent when the matrix size increases. On the other hand, as shown in Table 6.6, the number of iterations needed by the PCG method with preconditioner $T_{n+p-2}(f_p)$ is essentially independent of both n and p; see Remark 6.2.

As discussed in Subsection 6.2.5, the convergence rate of the two-grid method can be improved for large p by using the PCG method as smoother. In the following experiments, we replace the Richardson and Gauss-Seidel post-smoothers $\overline{S}_n^{[p]}$ and $\hat{S}_n^{[p]}$, used in the previous subsection, with a few PCG post-smoothing iterations (say $s^{[p]}$ iterations) with preconditioner $T_{n+p-2}(h_{p-1})$. Due to the presence of the PCG smoother,

n	$c_n^{[1]}$	$c_n^{[2]}$	$c_n^{[3]}$	$c_n^{[4]}$	$c_n^{[5]}$	$c_n^{[6]}$
80	40	40	41	42	44	44
160	80	80	81	83	86	87
320	160	160	161	166	170	172
640	320	320	321	331	338	343
1280	640	640	641	658	671	684
2560	1280	1280	1281	1310	1339	1362

Table 6.5: number of iterations $c_n^{[p]}$ needed by the PCG method with preconditioner $T_{n+p-2}(h_{p-1})$, for solving the system $\frac{1}{n}A_n^{[p]}\mathbf{u} = \mathbf{f}$ in the case $\beta = 0$, $\gamma = 1$, $f = 1$, up to a precision of 10^{-8}. The method has been started with $\mathbf{u}^{(0)} = \mathbf{0}$.

n	$c_n^{[1]}$	$c_n^{[2]}$	$c_n^{[3]}$	$c_n^{[4]}$	$c_n^{[5]}$	$c_n^{[6]}$
80	4	6	6	7	8	10
160	4	6	7	7	8	10
320	4	6	7	8	8	10
640	4	6	7	8	8	12
1280	4	6	7	8	10	11
2560	4	6	7	8	10	12

Table 6.6: number of iterations $c_n^{[p]}$ needed by the PCG method with preconditioner $T_{n+p-2}(f_p)$, for solving the system $\frac{1}{n}A_n^{[p]}\mathbf{u} = \mathbf{f}$ in the case $\beta = 0$, $\gamma = 1$, $f = 1$, up to a precision of 10^{-8}. The method has been started with $\mathbf{u}^{(0)} = \mathbf{0}$.

the resulting method is no more a stationary iterative method, and hence it is not a two-grid in the classical sense. However, using an expressive notation, we denote it by $TG((\text{PCG})^{s^{[p]}}, P_n^{[p]})$, where the exponent $s^{[p]}$ simply indicates that we apply $s^{[p]}$ steps of the PCG algorithm with preconditioner $T_{n+p-2}(h_{p-1})$.

Then, the same system $\frac{1}{n}A_n^{[p]}\mathbf{u} = \mathbf{f}$ considered in Tables 6.5–6.6 was solved for $p = 1, \ldots, 6$ and for increasing values of n by means of $TG((\text{PCG})^{s^{[p]}}, P_n^{[p]})$ and $TG((\hat{S}_n^{[p]})^{s^{[p]}}, P_n^{[p]})$. The latter method, as indicated by the notation, is the same as $TG(\hat{S}_n^{[p]}, P_n^{[p]})$, except that now we apply $s^{[p]}$ post-smoothing iterations by $\hat{S}_n^{[p]}$ instead of one. This is done for making a fair comparison with $TG((\text{PCG})^{s^{[p]}}, P_n^{[p]})$, in which $s^{[p]}$ steps of PCG are applied. For the smoother $\hat{S}_n^{[p]}$ we used the same (optimal) $\omega^{[p]}$ as in Table 6.2. Both $TG((\text{PCG})^{s^{[p]}}, P_n^{[p]})$ and $TG((\hat{S}_n^{[p]})^{s^{[p]}}, P_n^{[p]})$ were started with $\mathbf{u}^{(0)} = \mathbf{0}$ and stopped at the first term $\mathbf{u}^{(c)}$ satisfying (6.26). The corresponding numbers of iterations are collected in Table 6.7.

We observe from Table 6.7 that $TG((\text{PCG})^{s^{[p]}}, P_n^{[p]})$ performs better than $TG((\hat{S}_n^{[p]})^{s^{[p]}}, P_n^{[p]})$ not only for large p but also for small p, though the difference between the two methods is much more appreciable when p is large. In the 2D case, the difference in performances between their 2D variants is even more significant; see Table 6.12. Another observation from Table 6.7 is the following: provided we increase $s^{[p]}$ a little bit when p increases, the number of iterations $\tilde{c}_n^{[p]}$ needed by $TG((\text{PCG})^{s^{[p]}}, P_n^{[p]})$ to reach the preassigned accuracy 10^{-8} is essentially independent of both n and p. This implies that $TG((\text{PCG})^{s^{[p]}}, P_n^{[p]})$ is robust not only with respect to n but also with respect to p.

Summarizing, $TG((\text{PCG})^{s^{[p]}}, P_n^{[p]})$ is a totally robust method, not only with respect to n but also with respect to p. This property does not hold for the classical two-grid schemes $TG(\overline{S}_n^{[p]}, P_n^{[p]})$ and $TG(\hat{S}_n^{[p]}, P_n^{[p]})$, because we have seen that $\overline{\rho}_n^{[p]}$ and $\hat{\rho}_n^{[p]}$ increase with p.

n	$\tilde{c}_n^{[1]}$ [2]	$\hat{c}_n^{[1]}$ [0.9065]	$\tilde{c}_n^{[3]}$ [2]	$\hat{c}_n^{[3]}$ [0.9483]	$\tilde{c}_n^{[5]}$ [3]	$\hat{c}_n^{[5]}$ [1.1999]
80	4	7	6	6	5	8
160	3	7	6	6	5	8
320	3	7	6	6	5	8
640	3	7	6	6	6	9
1280	3	7	6	6	6	9
2560	3	7	6	6	6	9
n	$\tilde{c}_n^{[2]}$ [2]	$\hat{c}_n^{[2]}$ [0.9109]	$\tilde{c}_n^{[4]}$ [3]	$\hat{c}_n^{[4]}$ [1.0602]	$\tilde{c}_n^{[6]}$ [3]	$\hat{c}_n^{[6]}$ [1.3292]
81	6	7	5	6	6	12
161	6	7	5	6	6	12
321	6	7	5	6	6	12
641	7	7	5	6	6	12
1281	7	7	5	6	6	13
2561	7	8	6	6	6	13

Table 6.7: number of iterations $\tilde{c}_n^{[p]}$ and $\hat{c}_n^{[p]}$ needed by $TG((\mathrm{PCG})^{s^{[p]}}, P_n^{[p]})$ and $TG((\hat{S}_n^{[p]})^{s^{[p]}}, P_n^{[p]})$ respectively, for solving $\frac{1}{n}A_n^{[p]}\mathbf{u} = \mathbf{f}$ in the case $\beta = 0$, $\gamma = 1$, $f = 1$, up to a precision of 10^{-8}. The methods have been started with $\mathbf{u}^{(0)} = \mathbf{0}$. The parameters $s^{[p]}$ and $\omega^{[p]}$ are specified between brackets $[\cdot]$ near the labels $\tilde{c}_n^{[p]}$ and $\hat{c}_n^{[p]}$, respectively.

6.4 Two-grid algorithms and their performances: 2D

In this section, we consider specialized two-grid methods for solving linear systems with coefficient matrix $A_{n,n}^{[p,p]} = A_{\boldsymbol{n}}^{[\boldsymbol{p}]}$, where $\boldsymbol{p} = (p, p)$, $\boldsymbol{n} = (n, n) = n\boldsymbol{\nu}$ and $\boldsymbol{\nu} = (1, 1)$. We first examine the numerical behavior of classical two-grid schemes, which will prove to be unsatisfactory for large p. We will then consider the multi-iterative two-grid scheme analogous to the one tested in Subsection 6.3.2 and we shall see that this solver turns out to have a convergence rate that is at the same time optimal and robust, i.e., n-independent and p-independent.

6.4.1 Classical two-grid methods

We consider two-grid methods with the classical full-weighting projector $P_{n,n}^{[p,p]} := P_{n+p-2,n+p-2}$, as given by (6.20)–(6.21) for $\boldsymbol{m} = (n + p - 2, n + p - 2)$. As already pointed out, such a projector induces a coarse-grid correction effective in the low frequencies. Like in the 1D setting, we only consider two-grid methods without pre-smoothing steps and with a single post-smoothing step, and we provide two choices of the smoother: the relaxed Richardson smoother with iteration matrix $\overline{S}_{n,n}^{[p,p]}$ and the relaxed Gauss-Seidel smoother with iteration matrix $\hat{S}_{n,n}^{[p,p]}$; cf. (6.14)–(6.15). With the smoothers and the projector as above, our two-grid procedure is defined completely for $A = A_{n,n}^{[p,p]}$; see Algorithm 6.1.

Table 6.8 shows the results of some numerical experiments in the case $\boldsymbol{\beta} = \mathbf{0}$, $\gamma = 0$. For $p = 1, \ldots, 6$, we determined experimentally the parameter $\omega^{[p,p]}$ minimizing the quantity $\overline{\rho}_{n,n}^{[p,p]} := \rho(TG(\overline{S}_{n,n}^{[p,p]}, P_{n,n}^{[p,p]}))$, where n is chosen to be 52 (if p is odd) or 53 (if p is even). Then, we computed the spectral radii $\overline{\rho}_{n,n}^{[p,p]}$ for increasing values of n. In all the considered experiments, the proposed two-grid method is optimal. However, for $p = 4, 5, 6$ the spectral radii are very close to 1, and this is not satisfactory for practical purposes. The numerical experiments in Table 6.9, obtained as those in Table 6.8, show a certain improvement in the two-grid convergence rate when using the relaxed Gauss-Seidel smoother instead of Richardson's. However, for large p, the values $\hat{\rho}_{n,n}^{[p,p]}$ are still unsatisfactory.

n	$\overline{\rho}_{n,n}^{[1,1]}$ $[\omega^{[1,1]} = 0.3335]$	$\overline{\rho}_{n,n}^{[3,3]}$ $[\omega^{[3,3]} = 1.3739]$	$\overline{\rho}_{n,n}^{[5,5]}$ $[\omega^{[5,5]} = 1.3293]$
16	0.3287279	0.9248227	0.9984590
28	0.3316020	0.9239241	0.9983433
40	0.3323146	0.9231361	0.9983185
52	0.3325944	0.9229755	0.9983134
n	$\overline{\rho}_{n,n}^{[2,2]}$ $[\omega^{[2,2]} = 1.1009]$	$\overline{\rho}_{n,n}^{[4,4]}$ $[\omega^{[4,4]} = 1.4000]$	$\overline{\rho}_{n,n}^{[6,6]}$ $[\omega^{[6,6]} = 1.2505]$
17	0.6085689	0.9885344	0.9997977
29	0.6085689	0.9881173	0.9997766
41	0.6085689	0.9880112	0.9997724
53	0.6085689	0.9879839	0.9997715

Table 6.8: values of $\overline{\rho}_{n,n}^{[p,p]} := \rho(TG(\overline{S}_{n,n}^{[p,p]}, P_{n,n}^{[p,p]}))$ in the case $\boldsymbol{\beta} = \mathbf{0}$, $\gamma = 0$, for the specified parameter $\omega^{[p,p]}$.

n	$\hat{\rho}_{n,n}^{[1,1]}$ $[\omega^{[1,1]} = 1.0035]$	$\hat{\rho}_{n,n}^{[3,3]}$ $[\omega^{[3,3]} = 1.3143]$	$\hat{\rho}_{n,n}^{[5,5]}$ $[\omega^{[5,5]} = 1.3990]$
16	0.1588106	0.6420608	0.9629505
28	0.1678248	0.6411764	0.9633667
40	0.1753106	0.6418579	0.9626834
52	0.1804148	0.6465563	0.9620579
n	$\hat{\rho}_{n,n}^{[2,2]}$ $[\omega^{[2,2]} = 1.1695]$	$\hat{\rho}_{n,n}^{[4,4]}$ $[\omega^{[4,4]} = 1.3248]$	$\hat{\rho}_{n,n}^{[6,6]}$ $[\omega^{[6,6]} = 1.4914]$
17	0.2661407	0.8798035	0.9913084
29	0.2689991	0.8779954	0.9903263
41	0.2901481	0.8773914	0.9898795
53	0.3045791	0.8778602	0.9897372

Table 6.9: values of $\hat{\rho}_{n,n}^{[p,p]} := \rho(TG(\hat{S}_{n,n}^{[p,p]}, P_{n,n}^{[p,p]}))$ in the case $\boldsymbol{\beta} = \mathbf{0}$, $\gamma = 0$, for the specified parameter $\omega^{[p,p]}$.

6.4.2 Multi-iterative two-grid method with PCG as smoother

The convergence rate of both the two-grid schemes $TG(\overline{S}_{n,n}^{[p,p]}, P_{n,n}^{[p,p]})$ and $TG(\hat{S}_{n,n}^{[p,p]}, P_{n,n}^{[p,p]})$ rapidly worsens when p increases. The main reason, as explained in Subsection 6.2.4, is the presence of a large set of numerical zeros of the symbol $f_{p,p}^{(1,1)}$; see (6.4). Following the suggestion from Subsection 6.2.5, we now adopt a multi-iterative method, identical to the one tested in Subsection 6.3.2, which involves the PCG method as smoother.

Let us first illustrate the PCG method (see Algorithm 6.2) applied to the linear system $A_{n,n}^{[p,p]}\mathbf{u} = \mathbf{f}$, coming from the B-spline Galerkin approximation of the model problem (6.1) in the case $d = 2$ with $\boldsymbol{\beta} = \mathbf{0}$, $\gamma = 1$ and $f = 1$. Tables 6.10 and 6.11 report the number of iterations needed by the PCG with preconditioners $T_{n+p-2}(h_{p-1}) \otimes T_{n+p-2}(h_{p-1})$ and $T_{n+p-2}(f_p) \otimes T_{n+p-2}(f_p)$, respectively, for computing a vector $\mathbf{u}^{(c)}$ satisfying a relative residual less than 10^{-8}. As illustrated in Table 6.10, the former PCG is not efficient for large n, but its convergence rate is quite robust with respect to p; see Subsection 6.2.5 for an explanation of this phenomenon. On the other hand, Table 6.11 shows that the latter PCG is not at all effective, because the dependency on n and p is unsatisfactory; see Remark 6.2.

Then, the same system $A_{n,n}^{[p,p]}\mathbf{u} = \mathbf{f}$ considered in Tables 6.10–6.11 was solved for $p = 1, \ldots, 6$ and for increasing n, by means of $TG((\text{PCG})^{s^{[p,p]}}, P_{n,n}^{[p,p]})$ and $TG((\hat{S}_{n,n}^{[p,p]})^{s^{[p,p]}}, P_{n,n}^{[p,p]})$. The corresponding numbers of iterations are given in Table 6.12. For $\hat{S}_{n,n}^{[p,p]}$ we used the same (optimal) parameter $\omega^{[p,p]}$ as in Table 6.9. Both $TG((\text{PCG})^{s^{[p,p]}}, P_{n,n}^{[p,p]})$ and $TG((\hat{S}_{n,n}^{[p,p]})^{s^{[p,p]}}, P_{n,n}^{[p,p]})$ were started with $\mathbf{u}^{(0)} = \mathbf{0}$ and stopped at the first term $\mathbf{u}^{(c)}$ satisfying a criterion of relative residual less than 10^{-8}. Analogously to the 1D case (see Subsection 6.3.2), we can conclude that $TG((\text{PCG})^{s^{[p,p]}}, P_{n,n}^{[p,p]})$ is robust not only with respect to n but also with respect to p. The only unpleasant point is that, similarly to the 1D case, $s^{[p,p]}$ increases a little bit when p increases.

We end this subsection with a numerical experiment involving a nonzero convection term $\boldsymbol{\beta}$. To be precise, we consider the linear system $A_{n,n}^{[p,p]}\mathbf{u} = \mathbf{f}$ coming from the B-spline Galerkin approximation of the

n	$c_{n,n}^{[1,1]}$	$c_{n,n}^{[2,2]}$	$c_{n,n}^{[3,3]}$	$c_{n,n}^{[4,4]}$	$c_{n,n}^{[5,5]}$	$c_{n,n}^{[6,6]}$
20	25	24	26	29	33	40
40	52	49	49	57	65	77
60	78	75	75	83	96	118
80	104	100	100	111	130	157
100	131	125	126	140	165	198
120	157	151	151	168	200	241

Table 6.10: number of iterations $c_{n,n}^{[p,p]}$ needed by the PCG with preconditioner $T_{n+p-2}(h_{p-1}) \otimes T_{n+p-2}(h_{p-1})$, for solving the system $A_{n,n}^{[p,p]}\mathbf{u} = \mathbf{f}$ in the case $\boldsymbol{\beta} = \mathbf{0}$, $\gamma = 1$, $f = 1$, up to a precision of 10^{-8}. The method has been started with $\mathbf{u}^{(0)} = \mathbf{0}$.

n	$c_{n,n}^{[1,1]}$	$c_{n,n}^{[2,2]}$	$c_{n,n}^{[3,3]}$	$c_{n,n}^{[4,4]}$	$c_{n,n}^{[5,5]}$	$c_{n,n}^{[6,6]}$
20	64	79	100	120	153	184
40	133	166	195	232	293	364
60	203	249	286	342	419	518
80	266	328	374	444	538	662
100	328	403	462	546	660	808
120	391	480	549	649	773	952

Table 6.11: number of iterations $c_{n,n}^{[p,p]}$ needed by the PCG with preconditioner $T_{n+p-2}(f_p) \otimes T_{n+p-2}(f_p)$, for solving the system $A_{n,n}^{[p,p]}\mathbf{u} = \mathbf{f}$ in the case $\boldsymbol{\beta} = \mathbf{0}$, $\gamma = 1$, $f = 1$, up to a precision of 10^{-8}. The method has been started with $\mathbf{u}^{(0)} = \mathbf{0}$.

n	$\tilde{c}_{n,n}^{[1,1]}$ [2]	$\hat{c}_{n,n}^{[1,1]}$ [1.0035]	$\tilde{c}_{n,n}^{[3,3]}$ [2]	$\hat{c}_{n,n}^{[3,3]}$ [1.3143]	$\tilde{c}_{n,n}^{[5,5]}$ [4]	$\hat{c}_{n,n}^{[5,5]}$ [1.3990]
20	6	7	6	16	7	65
40	6	7	6	14	6	54
60	6	7	6	14	6	49
80	5	7	6	13	6	46
100	5	7	6	13	6	44
120	5	7	6	13	6	42
n	$\tilde{c}_{n,n}^{[2,2]}$ [2]	$\hat{c}_{n,n}^{[2,2]}$ [1.1695]	$\tilde{c}_{n,n}^{[4,4]}$ [3]	$\hat{c}_{n,n}^{[4,4]}$ [1.3248]	$\tilde{c}_{n,n}^{[6,6]}$ [6]	$\hat{c}_{n,n}^{[6,6]}$ [1.4914]
21	6	8	6	32	6	140
41	6	8	6	29	6	115
61	6	8	6	27	5	104
81	6	9	6	26	5	97
101	6	9	6	26	5	91
121	6	9	6	25	5	87

Table 6.12: number of iterations $\tilde{c}_{n,n}^{[p,p]}$ and $\hat{c}_{n,n}^{[p,p]}$ needed by $TG((\text{PCG})^{s^{[p,p]}}, P_{n,n}^{[p,p]})$ and $TG((\hat{S}_{n,n}^{[p,p]})^{s^{[p,p]}}, P_{n,n}^{[p,p]})$ respectively, for solving $A_{n,n}^{[p,p]}\mathbf{u} = \mathbf{f}$ in the case $\boldsymbol{\beta} = \mathbf{0}$, $\gamma = 1$, $f = 1$, up to a precision of 10^{-8}. The methods have been started with $\mathbf{u}^{(0)} = \mathbf{0}$. The parameters $s^{[p,p]}$ and $\omega^{[p,p]}$ are specified between brackets $[\cdot]$ near the labels $\tilde{c}_{n,n}^{[p,p]}$ and $\hat{c}_{n,n}^{[p,p]}$, respectively.

n	$\tilde{c}_{n,n}^{[1,1]}$ [2]	$\tilde{c}_{n,n}^{[3,3]}$ [2]	$\tilde{c}_{n,n}^{[5,5]}$ [4]
20	7	6	7
40	6	6	6
60	6	6	6
80	6	6	6
100	6	6	6
120	7	6	6

n	$\tilde{c}_{n,n}^{[2,2]}$ [2]	$\tilde{c}_{n,n}^{[4,4]}$ [3]	$\tilde{c}_{n,n}^{[6,6]}$ [6]
21	6	6	6
41	6	6	6
61	6	6	6
81	6	6	6
101	6	6	5
121	6	6	6

Table 6.13: number of iterations $\tilde{c}_{n,n}^{[p,p]}$ needed by $TG((\text{PGMRES})^{s^{[p,p]}}, P_{n,n}^{[p,p]})$ for solving $A_{n,n}^{[p,p]}\mathbf{u} = \mathbf{f}$ in the case $\boldsymbol{\beta} = (5,-5)$, $\gamma = 1$, $f = 1$, up to a precision of 10^{-8}. The method has been started with $\mathbf{u}^{(0)} = \mathbf{0}$. The parameter $s^{[p,p]}$ is specified between brackets $[\cdot]$.

model problem (6.1) in the case $d = 2$ with $\boldsymbol{\beta} = (5,-5)$, $\gamma = 1$ and $f = 1$. Due to the presence of the convection term, the matrix $A_{n,n}^{[p,p]}$ is no more symmetric. According to Remark 6.1, we replace the PCG smoother in the two-grid method $TG((\text{PCG})^{s^{[p,p]}}, P_{n,n}^{[p,p]})$ with the PGMRES smoother and, of course, we keep on using the preconditioner $T_{n+p-2}(h_{p-1}) \otimes T_{n+p-2}(h_{p-1})$ also in the PGMRES case. The results of the numerical experiment are shown in Table 6.13.

6.5 Two-grid algorithms and their performances: 3D

We are now convinced, on the basis of the results in the previous sections, that standard smoothers such as Richardson or Gauss-Seidel do not produce robust two-grid methods with respect to p. Hence, a fortiori, they cannot produce p-robust V-cycles or W-cycles. On the contrary, if we take as smoother the PCG or the PGMRES method with preconditioner given by (6.24), the resulting two-grid method is robust with respect to both n and p.

In this section we provide a 3D evidence of this (n,p)-robustness. In analogy with the previous sections, we consider the linear system $n^{d-2}A_{\boldsymbol{n}}^{[\boldsymbol{p}]}\mathbf{u} = \mathbf{f}$, coming from the B-spline Galerkin approximation of the model problem (6.1), in the case $d = 3$, with $\boldsymbol{\beta} = \mathbf{0}$, $\gamma = 1$, $f = 1$ and $\boldsymbol{n} = (n,n,n)$, $\boldsymbol{p} = (p,p,p)$. Then we solve this system up to a precision of 10^{-8}, using either the PCG method alone or the two-grid method $TG((\text{PCG})^{s^{[\boldsymbol{p}]}}, P_{\boldsymbol{n}}^{[\boldsymbol{p}]})$, where $P_{\boldsymbol{n}}^{[\boldsymbol{p}]} = P_{\boldsymbol{n}+\boldsymbol{p}-\mathbf{2}}$ is the projector defined in (6.20)–(6.21) for $\boldsymbol{m} = \boldsymbol{n} + \boldsymbol{p} - \mathbf{2}$, while $s^{[\boldsymbol{p}]}$ is the number of PCG post-smoothing iterations.

We see from Table 6.14 that the considered PCG method alone is p-robust. Table 6.15 shows the (n,p)-robustness of $TG((\text{PCG})^{s^{[\boldsymbol{p}]}}, P_{\boldsymbol{n}}^{[\boldsymbol{p}]})$. By comparing Tables 6.7, 6.12 and 6.15, we see that $TG((\text{PCG})^{s^{[\boldsymbol{p}]}}, P_{\boldsymbol{n}}^{[\boldsymbol{p}]})$ is also robust with respect to the dimensionality d. The only unpleasant point is that $s^{[\boldsymbol{p}]}$ slightly increases when p and d increase. Note, however, that the p-growth of $s^{[\boldsymbol{p}]}$ could be expected, because Tables 6.5, 6.10 and 6.14 show that the PCG method alone is p-robust, but not completely p-independent: for fixed n, the number of iterations slightly increases with p. Nevertheless, we should also say that if we decrease $s^{[\boldsymbol{p}]}$ a little bit, the number of iterations does not increase so much. For instance, if in Table 6.15 we had chosen $s^{[6,6,6]} = 6$ (instead of $s^{[6,6,6]} = 9$), then the resulting number of iterations $\tilde{c}_{n,n,n}^{[6,6,6]}$ for $n = 45$ would be 10.

n	$c_{n,n,n}^{[1,1,1]}$	$c_{n,n,n}^{[2,2,2]}$	$c_{n,n,n}^{[3,3,3]}$	$c_{n,n,n}^{[4,4,4]}$	$c_{n,n,n}^{[5,5,5]}$	$c_{n,n,n}^{[6,6,6]}$
15	21	19	22	28	35	54
25	35	32	33	40	50	67
35	49	46	46	53	68	84
45	64	60	60	68	84	105

Table 6.14: number of iterations $c_{n,n,n}^{[p,p,p]}$ needed by the PCG with preconditioner $T_{n+p-2}(h_{p-1}) \otimes T_{n+p-2}(h_{p-1}) \otimes T_{n+p-2}(h_{p-1})$, for solving the system $nA_{n,n,n}^{[p,p,p]}\mathbf{u} = \mathbf{f}$ in the case $\boldsymbol{\beta} = \mathbf{0}$, $\gamma = 1$, $f = 1$, up to a precision of 10^{-8}. The method has been started with $\mathbf{u}^{(0)} = \mathbf{0}$.

n	$\tilde{c}_{n,n,n}^{[1,1,1]}$ [2]	$\tilde{c}_{n,n,n}^{[3,3,3]}$ [3]	$\tilde{c}_{n,n,n}^{[5,5,5]}$ [5]
14	6	6	8
24	6	6	7
34	6	6	7
44	6	6	6
n	$\tilde{c}_{n,n,n}^{[2,2,2]}$ [2]	$\tilde{c}_{n,n,n}^{[4,4,4]}$ [4]	$\tilde{c}_{n,n,n}^{[6,6,6]}$ [9]
15	8	6	7
25	7	6	7
35	7	6	6
45	6	6	6

Table 6.15: number of iterations $\tilde{c}_{n,n,n}^{[p,p,p]}$ needed by $TG((\text{PCG})^{s^{[p,p,p]}}, P_{n,n,n}^{[p,p,p]})$ for solving $nA_{n,n,n}^{[p,p,p]}\mathbf{u} = \mathbf{f}$ in the case $\boldsymbol{\beta} = \mathbf{0}$, $\gamma = 1$, $f = 1$, up to a precision of 10^{-8}. The method has been started with $\mathbf{u}^{(0)} = \mathbf{0}$. The parameter $s^{[p,p,p]}$ is specified between brackets $[\cdot]$.

6.6 Multigrid: V-cycle and W-cycle

This section illustrates the numerical behavior of the V-cycle and W-cycle multigrid algorithms. Like for the two-grid algorithms, we observe an optimal convergence rate; see Tables 6.16–6.18. In all the numerical experiments of this section, we considered the linear systems $n^{d-2}A_{\boldsymbol{n}}^{[\boldsymbol{p}]}\mathbf{u} = \mathbf{f}$, coming from the B-spline Galerkin approximation of (6.1) in the cases $d = 1$, $d = 2$ and $d = 3$, respectively, with $\boldsymbol{\beta} = \mathbf{0}$, $\gamma = 0$, $f = 1$ and $\boldsymbol{n} = (n, \ldots, n)$, $\boldsymbol{p} = (p, \ldots, p)$. The V-cycle and W-cycle algorithms were started with initial guess $\mathbf{u}^{(0)} = \mathbf{0}$ and stopped with the criterion of the relative residual less than 10^{-8}, i.e., $\|n^{d-2}A_{\boldsymbol{n}}^{[\boldsymbol{p}]}\mathbf{u}^{(c)} - \mathbf{f}\| \leq 10^{-8}\|\mathbf{f}\|$.

6.6.1 1D case

Table 6.16 reports the numbers of iterations needed to solve the system $\frac{1}{n}A_n^{[p]}\mathbf{u} = \mathbf{f}$ with the V-cycle and the W-cycle multigrid. We now explain in detail how our multigrid algorithms were constructed.

The finest level is indicated by index 0 and the coarsest level by index $\ell_n^{[p]} := \log_2(n+p-1)-1$, assuming that $n+p-1$ is a power of 2. Let $A_{n,i}^{[p]}$ be the matrix at level i and let $m_{n,i}^{[p]}$ denote its dimension, $0 \leq i \leq \ell_n^{[p]}$. In this notation, we have $A_{n,0}^{[p]} = \frac{1}{n}A_n^{[p]}$,

$$A_{n,i+1}^{[p]} = P_{n,i}^{[p]} A_{n,i}^{[p]} (P_{n,i}^{[p]})^T, \quad i = 0, \ldots, \ell_n^{[p]} - 1,$$

and $A_{n,\ell_n^{[p]}}^{[p]}$ has dimension 1. In the above expression,

$$P_{n,i}^{[p]} := P_{m_{n,i}^{[p]}}, \quad i = 0, \ldots, \ell_n^{[p]} - 1,$$

is the projector at level i, defined by (6.20)–(6.21) for $d = 1$ and $\boldsymbol{m} = m_{n,i}^{[p]}$. Given the shape of $P_{m_{n,i}^{[p]}}$, one can show by induction on i that $m_{n,i+1}^{[p]} = (m_{n,i}^{[p]} - 1)/2$, $i = 0, \ldots, \ell_n^{[p]} - 1$, and $m_{n,i}^{[p]} = \frac{n+p-1}{2^i} - 1$, $i = 0, \ldots, \ell_n^{[p]}$.

n	$\tilde{c}_n^{[1]}$ [2]		$\hat{c}_n^{[1]}$ [0.9065]		n	$\tilde{c}_n^{[3]}$ [2]		$\hat{c}_n^{[3]}$ [0.9483]		n	$\tilde{c}_n^{[5]}$ [3]		$\hat{c}_n^{[5]}$ [1.1999]	
16	10	7	9	7	14	8	6	7	5	12	7	5	7	7
32	11	7	10	7	30	9	6	8	5	28	9	5	8	8
64	12	7	11	7	62	10	6	9	6	60	10	5	9	8
128	13	7	12	8	126	11	6	9	6	124	11	5	10	8
256	13	7	12	8	254	11	6	10	6	252	12	6	11	8
512	14	7	13	8	510	12	6	11	6	508	13	6	12	9
1024	14	7	14	8	1022	12	6	12	6	1020	13	6	13	9

n	$\tilde{c}_n^{[2]}$ [2]		$\hat{c}_n^{[2]}$ [0.9109]		n	$\tilde{c}_n^{[4]}$ [3]		$\hat{c}_n^{[4]}$ [1.0602]		n	$\tilde{c}_n^{[6]}$ [3]		$\hat{c}_n^{[6]}$ [1.3292]	
15	8	6	7	6	13	8	6	6	5	11	7	5	10	10
31	10	6	9	7	29	9	6	8	6	27	9	6	12	12
63	11	6	10	7	61	10	6	9	6	59	9	6	12	12
127	11	6	11	7	125	11	6	10	6	123	11	6	12	12
255	12	7	11	7	253	12	6	11	6	251	12	6	12	12
511	13	7	12	7	509	12	6	12	6	507	13	6	13	12
1023	13	7	12	7	1021	13	6	13	6	1019	14	6	13	13

Table 6.16: number of iterations $\tilde{c}_n^{[p]}$ (resp. $\hat{c}_n^{[p]}$) needed for solving $\frac{1}{n}A_n^{[p]}\mathbf{u} = \mathbf{f}$ in the case $\beta = \gamma = 0$ and $f = 1$, up to a precision of 10^{-8}, when using the multigrid cycle with $s^{[p]}$ post-smoothing steps by the PCG algorithm (resp. by the relaxed Gauss-Seidel smoother $\hat{S}_{n,0}^{[p]}$) at the finest level, and one post-smoothing step by the simple Gauss-Seidel smoother $\hat{S}_{n,i}^{[p]}$ at all other levels. The parameters $s^{[p]}$ and $\omega^{[p]}$ are specified between brackets $[\cdot]$ near the labels $\tilde{c}_n^{[p]}$ and $\hat{c}_n^{[p]}$, respectively. For each pair (p, n), the first entry in the cell corresponding to $\tilde{c}_n^{[p]}$ refers to the V-cycle, the second to the W-cycle. The same holds for $\hat{c}_n^{[p]}$.

We note that the choice of the projector $P_{n,i}^{[p]}$ at each level i has the same motivation as the projector $P_n^{[p]}$ for $\frac{1}{n}A_n^{[p]}$. Indeed, we know that $A_{n,0}^{[p]} = \frac{1}{n}A_n^{[p]}$ has the symbol $f_{p,0} := f_p$. Then, referring to [62, Proposition 2.2] or [2, Proposition 2.5], it follows that $A_{n,i}^{[p]}$ has a symbol $f_{p,i}$ at level i sharing the same properties of the symbol $f_{p,0}$ at level 0: $f_{p,i}(0) = 0$, with $\theta = 0$ a zero of order two, and $f_{p,i}(\theta) > 0$ for all $\theta \in [-\pi, \pi]\backslash\{0\}$ (see also Subsection 3.7.1 in [64]). These properties make it necessary to use for $A_{n,i}^{[p]}$ a projector like $P_{n,i}^{[p]}$, which is effective in low frequencies.

Regarding the smoother, at each coarse level $i \geq 1$ we chose the standard Gauss-Seidel smoother without relaxation $\hat{S}_{n,i}^{[p]}$, as given in (6.15) for $A = \frac{1}{n}A_n^{[p]}$ and $\omega = 1$. However, at the finest level $i = 0$ we considered two alternatives: $s^{[p]}$ smoothing iterations by the PCG method with preconditioner $T_{n+p-2}(h_{p-1})$, as in Subsection 6.3.2, or $s^{[p]}$ smoothing iterations by the relaxed Gauss-Seidel method $\hat{S}_{n,0}^{[p]}$ with the relaxation parameter $\omega^{[p]}$ as in Table 6.2. Note that, due to the presence of the (optimal) parameter $\omega^{[p]}$, $\hat{S}_{n,0}^{[p]}$ is different from $\hat{S}_{n,i}^{[p]}$, $i \geq 1$.

At each level i, we first performed a coarse-grid correction, with one recursive call in the V-cycle and two recursive calls in the W-cycle, and then we applied one post-smoothing iteration by $\hat{S}_{n,i}^{[p]}$ (if $i \geq 1$), or $s^{[p]}$ post-smoothing iterations by the PCG algorithm or $\hat{S}_{n,0}^{[p]}$ (if $i = 0$). From Table 6.16 we can conclude that all the proposed multigrid methods have an optimal convergence rate, independent of n. Moreover, the versions with a few PCG smoothing steps are also robust in p.

Finally, we want to motivate why the $s^{[p]}$ PCG smoothing steps were used only at the finest level. Let $M_{f_{p,i}} := \max_{\theta\in[-\pi,\pi]} f_{p,i}(\theta)$. Referring to [62, Proposition 2.2 (item 2)], and taking into account some additional numerical experiments that we performed, it seems that the numerical zero $\theta = \pi$ of $f_{p,0}/M_{f_{p,0}}$ disappears for $i \geq 1$, and each $f_{p,i}/M_{f_{p,i}}$, $i \geq 1$, only possesses the actual zero $\theta = 0$, like the symbol $2 - 2\cos\theta$ associated with the FD discretization matrices in one dimension; see (6.19). Hence, a single smoothing iteration by the standard Gauss-Seidel method is all we need at the coarse levels $i \geq 1$.

n	$\tilde{c}_{n,n}^{[1,1]}$ [2]		$\hat{c}_{n,n}^{[1,1]}$ [1.0035]		n	$\tilde{c}_{n,n}^{[3,3]}$ [2]		$\hat{c}_{n,n}^{[3,3]}$ [1.3143]		n	$\tilde{c}_{n,n}^{[5,5]}$ [4]		$\hat{c}_{n,n}^{[5,5]}$ [1.3990]	
16	10	7	9	7	14	7	6	16	16	12	7	7	85	85
32	11	7	10	7	30	9	6	15	15	28	8	6	59	59
64	12	7	11	7	62	9	6	14	14	60	10	6	49	49
128	13	7	12	7	126	10	6	13	13	124	11	6	42	42
256	13	7	13	7	254	11	6	13	13	252	12	6	38	38
n	$\tilde{c}_{n,n}^{[2,2]}$ [2]		$\hat{c}_{n,n}^{[2,2]}$ [1.1695]		n	$\tilde{c}_{n,n}^{[4,4]}$ [3]		$\hat{c}_{n,n}^{[4,4]}$ [1.3248]		n	$\tilde{c}_{n,n}^{[6,6]}$ [6]		$\hat{c}_{n,n}^{[6,6]}$ [1.4914]	
15	8	6	8	8	13	7	6	37	37	11	7	7	204	204
31	9	6	8	8	29	8	6	30	30	27	8	6	129	129
63	10	6	9	9	61	10	6	27	28	59	10	6	105	105
127	11	6	10	9	125	11	6	25	25	123	11	6	86	87
255	12	7	11	9	253	12	6	23	23	251	12	6	71	72

Table 6.17: number of iterations $\tilde{c}_{n,n}^{[p,p]}$ (resp. $\hat{c}_{n,n}^{[p,p]}$) needed for solving $A_{n,n}^{[p,p]}\mathbf{u} = \mathbf{f}$ in the case $\boldsymbol{\beta} = \mathbf{0}$, $\gamma = 0$, $f = 1$, up to a precision of 10^{-8}, when using the multigrid cycle with $s^{[p,p]}$ post-smoothing steps by the PCG algorithm (resp. by the relaxed Gauss-Seidel smoother $\hat{S}_{n,n,0}^{[p,p]}$) at the finest level and one post-smoothing step by the simple Gauss-Seidel smoother $\hat{S}_{n,n,i}^{[p,p]}$ at all other levels. The parameters $s^{[p,p]}$ and $\omega^{[p,p]}$ are specified between brackets $[\cdot]$ near the labels $\tilde{c}_{n,n}^{[p,p]}$ and $\hat{c}_{n,n}^{[p,p]}$, respectively. For each pair (p,n), the first entry in the cell corresponding to $\tilde{c}_{n,n}^{[p,p]}$ refers to the V-cycle, the second to the W-cycle. The same holds for $\hat{c}_{n,n}^{[p,p]}$.

6.6.2 2D case

Table 6.17 reports the numbers of iterations needed to solve the system $A_{n,n}^{[p,p]}\mathbf{u} = \mathbf{f}$ with the V-cycle and the W-cycle multigrid. The multigrid algorithms were constructed in a similar way as in the 1D case.

The finest level is again indicated by index 0 and the coarsest level by index $\ell_n^{[p]} := \log_2(n+p-1)-1$. Let $A_{n,n,i}^{[p,p]}$ be the matrix at level i, whose dimension is $(m_{n,i}^{[p]})^2$, $0 \le i \le \ell_n^{[p]}$, with $m_{n,i}^{[p]} := \frac{n+p-1}{2^i} - 1$ as in Subsection 6.6.1. We have

$$A_{n,n,i+1}^{[p,p]} = P_{n,n,i}^{[p,p]} A_{n,n,i}^{[p,p]} (P_{n,n,i}^{[p,p]})^T, \quad i = 0, \ldots, \ell_n^{[p]} - 1,$$

where

$$P_{n,n,i}^{[p,p]} := P_{m_{n,i}^{[p]}, m_{n,i}^{[p]}}, \quad i = 0, \ldots, \ell_n^{[p]} - 1,$$

is the projector at level i, defined by (6.20)–(6.21) for $d = 2$ and $\boldsymbol{m} = (m_{n,i}^{[p]}, m_{n,i}^{[p]})$.

Regarding the smoother, we took the same choices as in the 1D case. At each coarse level $i \ge 1$ we used the standard Gauss-Seidel smoother without relaxation. However, at the finest level $i = 0$ we used either $s^{[p,p]}$ smoothing iterations by the PCG algorithm with preconditioner (6.24) or $s^{[p,p]}$ smoothing iterations by the relaxed Gauss-Seidel method $\hat{S}_{n,n,0}^{[p,p]}$ with the relaxation parameter $\omega^{[p,p]}$ as in Table 6.9.

At each level i, we first performed a coarse-grid correction, with one recursive call in the V-cycle and two recursive calls in the W-cycle, and then we applied one post-smoothing iteration by $\hat{S}_{n,n,i}^{[p,p]}$ (if $i \ge 1$), or $s^{[p,p]}$ post-smoothing iterations by the PCG algorithm or $\hat{S}_{n,n,0}^{[p,p]}$ (if $i = 0$).

6.6.3 3D case

Table 6.18 reports the numbers of iterations needed to solve $nA_{n,n,n}^{[p,p,p]}\mathbf{u} = \mathbf{f}$ with the V-cycle and W-cycle multigrid. The multigrid algorithms were constructed as in the 1D and 2D case.

The finest level is again indicated by index 0 and the coarsest level by index $\ell_n^{[p]} := \log_2(n+p-1)-1$. Let $A_{n,n,n,i}^{[p,p,p]}$ be the matrix at level i, whose dimension is $(m_{n,i}^{[p]})^3$, $0 \le i \le \ell_n^{[p]}$, with $m_{n,i}^{[p]} := \frac{n+p-1}{2^i} - 1$ as in Subsections 6.6.1–6.6.2. We have $A_{n,n,n,0}^{[p,p,p]} = nA_{n,n,n}^{[p,p,p]}$ and

$$A_{n,n,n,i+1}^{[p,p,p]} = P_{n,n,n,i}^{[p,p,p]} A_{n,n,n,i}^{[p,p,p]} (P_{n,n,n,i}^{[p,p,p]})^T, \quad i = 0, \ldots, \ell_n^{[p]} - 1,$$

n	$\tilde{c}^{[1,1,1]}_{n,n,n}$ [2]		n	$\tilde{c}^{[3,3,3]}_{n,n,n}$ [3]		n	$\tilde{c}^{[5,5,5]}_{n,n,n}$ [5]	
16	10	7	14	7	6	12	8	8
32	11	7	30	8	6	28	8	7
64	12	7	62	9	6	60	9	6
n	$\tilde{c}^{[2,2,2]}_{n,n,n}$ [2]		n	$\tilde{c}^{[4,4,4]}_{n,n,n}$ [4]		n	$\tilde{c}^{[6,6,6]}_{n,n,n}$ [9]	
15	9	8	13	7	6	11	9	9
31	8	7	29	8	6	27	8	6
63	9	7	61	9	6	59	10	6

Table 6.18: number of iterations $\tilde{c}^{[p,p,p]}_{n,n,n}$ needed for solving $nA^{[p,p,p]}_{n,n,n}\mathbf{u} = \mathbf{f}$ in the case $\boldsymbol{\beta} = \mathbf{0}$, $\gamma = 0$, $f = 1$, up to a precision of 10^{-8}, when using the multigrid cycle with $s^{[p,p,p]}$ post-smoothing steps by the PCG algorithm at the finest level and one post-smoothing step by the simple Gauss-Seidel smoother $\hat{S}^{[p,p,p]}_{n,n,n,i}$ at all other levels. The parameter $s^{[p,p,p]}$ is specified between brackets $[\cdot]$. For each pair (p,n), the first entry in the cell corresponding to $\tilde{c}^{[p,p,p]}_{n,n,n}$ refers to the V-cycle, the second to the W-cycle.

where

$$P^{[p,p,p]}_{n,n,n,i} := P_{m^{[p]}_{n,i},m^{[p]}_{n,i},m^{[p]}_{n,i}}, \quad i = 0,\dots,\ell^{[p]}_n - 1,$$

is the projector at level i, as given by (6.20)–(6.21) for $d = 3$ and $\boldsymbol{m} = (m^{[p]}_{n,i}, m^{[p]}_{n,i}, m^{[p]}_{n,i})$.

Regarding the smoother, we took the same choices as in the 1D and 2D case. At each coarse level $i \geq 1$ we used the standard Gauss-Seidel smoother without relaxation; at the finest level $i = 0$ we used $s^{[p,p,p]}$ smoothing iterations by the PCG algorithm with preconditioner (6.24).

At each level i, we first performed a coarse-grid correction, with one recursive call in the V-cycle and two recursive calls in the W-cycle, and then we applied one post-smoothing iteration by $\hat{S}^{[p,p,p]}_{n,n,n,i}$ (if $i \geq 1$), or $s^{[p,p,p]}$ post-smoothing iterations by the PCG algorithm (if $i = 0$).

When using a few PCG smoothing steps at the finest level, we can conclude from Tables 6.16–6.18 that the resulting V-cycle and W-cycle multigrid algorithms have a convergence rate which is substantially independent not only of n but also of p. This means that they are robust with respect to both n and p. We also note that the W-cycle convergence rate is essentially the same as the corresponding two-grid convergence rate: compare Tables 6.16–6.18 with Tables 6.7, 6.12 and 6.15.

6.7 Further insights: fast multi-iterative solver for Galerkin B-spline IgA stiffness matrices associated with full elliptic problems

In Section 6.6 we have designed optimal and robust multi-iterative methods of multigrid type for solving linear systems with coefficient matrix $A^{[\boldsymbol{p}]}_{\boldsymbol{n}}$ as in (6.2) with $\boldsymbol{\beta} = \mathbf{0}$ and $\gamma = 0$; this is the B-spline discretization matrix related to the Laplacian on the hypercube and will be referred to as the Parametric Laplacian matrix (or PL-matrix). In this section we show that the solution of linear systems related to the Galerkin B-spline IgA approximation of more general elliptic problems with variable coefficients and with a domain deformation can be reduced to the solution of linear systems involving the PL-matrix. Indeed, the PL-matrix itself is an optimal and robust GMRES preconditioner for the full IgA stiffness matrices.

We begin with a brief description of the isogeometric Galerkin method for the solution of full elliptic problems with variable coefficients on general domains. Here, we do not confine ourselves to the isogeometric approach in the strict sense, since we allow the geometry map to be any function, not necessarily described by B-splines. Then, we provide the expression of the resulting stiffness matrices. Finally, we give a numerical evidence of the optimality of the PL-matrix as a preconditioner for such matrices.

Let us consider the following full elliptic differential problem:

$$\begin{cases} -\nabla \cdot K\nabla u + \boldsymbol{\beta} \cdot \nabla u + \gamma u = f & \text{in } \Omega, \\ u = 0 & \text{on } \partial\Omega, \end{cases} \tag{6.27}$$

where Ω is a bounded open domain in $\mathbb{R}^d$, $K : \Omega \to \mathbb{R}^{d\times d}$ is an SPD matrix of functions in $L^\infty(\Omega)$, $\boldsymbol{\beta} : \Omega \to \mathbb{R}^d$ is a vector of functions in $L^\infty(\Omega)$, $\gamma \in L^\infty(\Omega)$, $\gamma \geq 0$ and $f \in L^2(\Omega)$. The weak form of (6.27) consists in finding $u \in H^1_0(\Omega)$ such that

$$\int_\Omega (K\nabla u \cdot \nabla v + \boldsymbol{\beta} \cdot \nabla u\, v + \gamma u v) = \int_\Omega f v, \quad \forall v \in H^1_0(\Omega). \tag{6.28}$$

Suppose that the physical domain Ω can be described by a global geometry map $\mathbf{G} : \hat{\Omega} \to \overline{\Omega}$, which is invertible in the parametric domain $\hat{\Omega} := [0,1]^d$ and satisfies $\mathbf{G}(\partial\hat{\Omega}) = \partial\overline{\Omega}$. Let $\{\hat{\varphi}_1, \dots, \hat{\varphi}_m\}$ be a set of basis functions defined on $\hat{\Omega}$ and vanishing on the boundary $\partial\hat{\Omega}$. We approximate the solution of (6.28) by the Galerkin method using the approximation space $\mathscr{W} := \langle \varphi_i : \ i = 1, \dots, m\rangle \subset H^1_0(\Omega)$, where

$$\varphi_i(\mathbf{x}) := \hat{\varphi}_i(\mathbf{G}^{-1}(\mathbf{x})) = \hat{\varphi}_i(\hat{\mathbf{x}}), \qquad \mathbf{x} = \mathbf{G}(\hat{\mathbf{x}}).$$

More precisely, we look for $u_{\mathscr{W}} \in \mathscr{W}$ such that

$$\int_\Omega (K\nabla u_{\mathscr{W}} \cdot \nabla v + \boldsymbol{\beta} \cdot \nabla u_{\mathscr{W}}\, v + \gamma u_{\mathscr{W}} v) = \int_\Omega f v, \quad \forall v \in \mathscr{W}, \tag{6.29}$$

which is equivalent to solving the linear system $A_{\mathbf{G}}\mathbf{u} = \mathbf{f}_{\mathbf{G}}$, where

$$A_{\mathbf{G}} := \left[\int_\Omega (K\nabla\varphi_j \cdot \nabla\varphi_i + \boldsymbol{\beta} \cdot \nabla\varphi_j\ \varphi_i + \gamma\varphi_j\varphi_i)\right]_{i,j=1}^m, \qquad \mathbf{f}_{\mathbf{G}} := \left[\int_\Omega f\varphi_i\right]_{i=1}^m,$$

and $\mathbf{u}$ is the coefficient vector of $u_{\mathscr{W}}$ with respect to $\{\varphi_1, \dots, \varphi_m\}$: $u_{\mathscr{W}} = \sum_{j=1}^m u_j\varphi_j$. Assuming that $\mathbf{G}$ and $\hat{\varphi}_i$, $i = 1, \dots, m$, are sufficiently regular, we can apply standard differential calculus and we get the following expressions for $A_{\mathbf{G}}$ and $\mathbf{f}_{\mathbf{G}}$ in terms of $\mathbf{G}$ and $\hat{\varphi}_i$, $i = 1, \dots, m$:

$$A_{\mathbf{G}} = \left[\int_{\hat{\Omega}} \left((\nabla\hat{\varphi}_j)^T (J_{\mathbf{G}})^{-1} K(\mathbf{G}) (J_{\mathbf{G}})^{-T} \nabla\hat{\varphi}_i + (\nabla\hat{\varphi}_j)^T (J_{\mathbf{G}})^{-1} \boldsymbol{\beta}(\mathbf{G})\, \hat{\varphi}_i + \gamma(\mathbf{G})\hat{\varphi}_j\hat{\varphi}_i\right) |\det(J_{\mathbf{G}})|\right]_{i,j=1}^m, \tag{6.30}$$

$$\mathbf{f}_{\mathbf{G}} = \left[\int_{\hat{\Omega}} f(\mathbf{G})\, \hat{\varphi}_i\, |\det(J_{\mathbf{G}})|\right]_{i=1}^m, \tag{6.31}$$

where

$$J_{\mathbf{G}} := \left[\frac{\partial G_i}{\partial \hat{x}_j}\right]_{i,j=1}^m = \left[\frac{\partial x_i}{\partial \hat{x}_j}\right]_{i,j=1}^m$$

is the Jacobian matrix of $\mathbf{G}$. In the framework of IgA based on B-splines, the basis functions $\hat{\varphi}_i$, $i = 1, \dots, m$, are tensor-product B-splines as in (4.5) and (5.14). The resulting stiffness matrix $A_{\mathbf{G}}$ in (6.30) is denoted by $A^{[\boldsymbol{p}]}_{\mathbf{G},\boldsymbol{n}}$ to emphasize its dependence on the B-spline degrees $\boldsymbol{p}$ and the fineness parameters $\boldsymbol{n}$.

We now focus on a specific example in the case $d = 2$, in which we illustrate that the PL-matrix is an optimal and robust GMRES preconditioner for the matrix (6.30). We consider problem (6.27) on a quarter of an annulus, namely

$$\Omega = \{(x, y) \in \mathbb{R}^2 : \ r^2 < x^2 + y^2 < R^2, \ x > 0, \ y > 0\}, \quad r = 1, \quad R = 4,$$

n	$p=1$		$p=2$		$p=3$		$p=4$		$p=5$		$p=6$	
	I	P	I	P	I	P	I	P	I	P	I	P
10	29	16	25	18	42	19	72	21	119	22	164	23
20	61	20	42	21	50	22	84	23	140	24	223	25
30	94	22	63	23	60	23	90	24	154	25	240	26
40	128	23	84	24	77	24	95	25	161	26	249	26
50	161	24	106	24	96	25	106	26	168	26	256	27

Table 6.19: number of GMRES iterations without (I) and with (P) preconditioning for solving $A^{[p,p]}_{\mathbf{G},n,n}\mathbf{u} = \mathbf{f}_{\mathbf{G}}$ up to a precision of 10^{-8}, varying the fineness parameter $n_1 = n_2 = n$ and the spline degree $p_1 = p_2 = p$.

and with

$$
\begin{aligned}
K(x,y) &= \begin{bmatrix} (2+\cos x)(1+y) & \cos(x+y)\sin(x+y) \\ \cos(x+y)\sin(x+y) & (2+\sin y)(1+x) \end{bmatrix}, \\
\boldsymbol{\beta}(x,y) &= \sqrt{x^2+y^2}\begin{bmatrix} \cos\frac{x}{\sqrt{x^2+y^2}} \\ \sin\frac{y}{\sqrt{x^2+y^2}} \end{bmatrix}, \\
\gamma(x,y) &= xy, \\
f(x,y) &= x\cos y + y\sin x.
\end{aligned}
$$

The geometry map is given by

$$
\mathbf{G}(\hat{x},\hat{y}) = (x,y), \quad \text{where} \quad \begin{cases} x = [r+\hat{x}(R-r)]\cos(\frac{\pi}{2}\hat{y}), \\ y = [r+\hat{x}(R-r)]\sin(\frac{\pi}{2}\hat{y}). \end{cases}
$$

We solved the corresponding IgA Galerkin system $A^{[p,p]}_{\mathbf{G},n,n}\mathbf{u} = \mathbf{f}_{\mathbf{G}}$ using GMRES without restarting and with a tolerance of 10^{-8}. The results are collected in Table 6.19. The GMRES method was applied first without preconditioning and then with the preconditioner $A^{[p,p]}_{n,n}$, given by (6.2) for $d = 2$, $\boldsymbol{n} = (n,n)$, $\boldsymbol{p} = (p,p)$ and $\boldsymbol{\beta} = \mathbf{0}$, $\gamma = 0$. The table clearly illustrates that the PL-matrix $A^{[p,p]}_{n,n}$ is an optimal and robust GMRES preconditioner for $A^{[p,p]}_{\mathbf{G},n,n}$. Indeed, the number of iterations to reach the fixed accuracy 10^{-8} is substantially independent of both n and p.

Summarizing, our proposal for solving linear systems associated with the B-spline IgA Galerkin approximation of full elliptic problems such as (6.27) is the following.

- As external solver, we use a PGMRES method, with preconditioner given by the PL-matrix $A^{[p,\dots,p]}_{n,\dots,n}$. The theoretical foundation of such a proposal falls beyond the scope of the paper. However, we can anticipate that the observed (optimal and robust) convergence rate is related to a conditioning measure of K, i.e.,

  $$\frac{\sup_{(x,y)\in\Omega}\lambda_{\max}(K(x,y))}{\inf_{(x,y)\in\Omega}\lambda_{\min}(K(x,y))},$$

  and to the same measure for $(J_{\mathbf{G}})^T J_{\mathbf{G}}$. Again, the analysis is based on the study of the symbol, in a similar way as carried out in Chapter 5 for the IgA collocation setting.

- The PL-matrix (or, more precisely, its scaled version $n^{d-2}A^{[p,\dots,p]}_{n,\dots,n}$) is treated by the specific multi-iterative solver of multigrid type designed in Section 6.6. This consists of a V-cycle or W-cycle multigrid method, which applies the standard full-weighting projector (6.21) at each level, a few post-smoothing iterations by the PCG method with preconditioner $\bigotimes_{j=1}^{d} T_{n+p-2}(h_{p-1})$ at the finest level, and one single post-smoothing iteration by the standard Gauss-Seidel method at all other levels.

Chapter 7

Fast iterative solvers for B-spline IgA collocation linear systems

This chapter is in many respects analogous to the previous one. In Chapter 5, we studied the spectral properties of the collocation matrices $A_{\mathbf{G},\boldsymbol{n}}^{[\boldsymbol{p}]}$ coming from the B-spline IgA collocation approximation of the second-order full elliptic problem

$$\begin{cases} -\nabla \cdot K\nabla u + \boldsymbol{\alpha} \cdot \nabla u + \gamma u = f & \text{in } \Omega, \\ u = 0 & \text{on } \partial\Omega, \end{cases} \iff \begin{cases} -\mathbf{1}(K \circ Pu)\mathbf{1}^T + \boldsymbol{\beta} \cdot \nabla u + \gamma u = f & \text{in } \Omega, \\ u = 0 & \text{on } \partial\Omega, \end{cases} \tag{7.1}$$

where Ω is a bounded open domain in $\mathbb{R}^d$, $K : \overline{\Omega} \to \mathbb{R}^{d\times d}$ is a SPD matrix of functions in $C^1(\Omega) \cap C(\overline{\Omega})$, $\boldsymbol{\alpha} : \overline{\Omega} \to \mathbb{R}^d$ is a vector of functions in $C(\overline{\Omega})$, $\gamma, f \in C(\overline{\Omega})$, $\gamma \geq 0$, and $Pu, \boldsymbol{\beta}$ are given in (5.3)–(5.4). In particular, we have computed the spectral symbol $f_{\mathbf{G},\boldsymbol{p}}^{(\nu)}$ of the normalized matrix-sequence $\{\frac{1}{n^2}A_{\mathbf{G},\boldsymbol{n}}^{[\boldsymbol{p}]}\}_n$, $\boldsymbol{n} = \boldsymbol{\nu}n$, and observed that, in the case where $\mathbf{G}$ is the identity map over the parameteric domain $\hat{\Omega} := [0,1]^d$ and K is the identity matrix, the symbol $f_{\mathbf{G},\boldsymbol{p}}^{(\nu)}$ reduces to the function $f_{\boldsymbol{p}}^{(\nu)}$ in (5.93); see Remark 5.3. We will now exploit the properties of the symbol in order to design fast iterative algorithms for solving linear systems with coefficient matrix $A_{\mathbf{G},\boldsymbol{n}}^{[\boldsymbol{p}]}$. As in Chapter 6, our goal is to obtain an iterative method that is optimal and robust at the same time, meaning that its convergence rate is simultaneously $\boldsymbol{n}$-independent and $\boldsymbol{p}$-independent. Using the properties of $f_{\mathbf{G},\boldsymbol{p}}^{(\nu)}$ and $f_{\boldsymbol{p}}^{(\nu)}$, we will succeed in designing a multi-iterative solver with these features, which will be essentially identical to the one presented in Chapter 6 (see in particular Section 6.7). The solver consists of the following two-step strategy.

1. An external PGMRES for $A_{\mathbf{G},\boldsymbol{n}}^{[\boldsymbol{p}]}$, with preconditioner equal to the so-called Parametric Laplacian (PL) matrix $A_{\boldsymbol{n}}^{[\boldsymbol{p}]}$, that is the matrix coming from the IgA collocation approximation of (7.1) in the case where K is the identity matrix, $\boldsymbol{\alpha} = \mathbf{0}$, $\gamma = 0$ and $\mathbf{G}$ is the identity map on the parametric domain $\hat{\Omega} = [0,1]^d$.

2. The PL-matrix $A_{\boldsymbol{n}}^{[\boldsymbol{p}]}$, or, more precisely, its scaled version $\frac{1}{n^2}A_{\boldsymbol{n}}^{[\boldsymbol{p}]}$ ($\boldsymbol{n} = \boldsymbol{\nu}n$), is treated by a specific multi-iterative multigrid solver consisting of a V-cycle (or W-cycle) formed by:

   (a) a standard full-weighting restriction operator at each level, chosen as in (6.20)–(6.21), which reduces the error in the low frequencies (a subspace of ill-conditioning due to the zero of the symbol $f_{\boldsymbol{p}}^{(\nu)}$ at $\boldsymbol{\theta} = \mathbf{0}$);

   (b) one standard post-smoothing iteration by the classical Gauss-Seidel method at all the coarse levels and a few post-smoothing iterations by a certain PGMRES at the finest level, where the latter is designed for reducing the error in the high frequencies (a subspace of ill-conditioning due to the numerical zeros of the normalized symbol $f_{\boldsymbol{p}}^{(\nu)}/M_{f_{\boldsymbol{p}}^{(\nu)}}$ at the π-edge points (7.3); see Lemma 7.1 below). In particular, the PGMRES preconditioner is chosen as the Toeplitz matrix

$$T_{\boldsymbol{n}+\boldsymbol{p}-\mathbf{2}}(h_{p_1-2} \otimes \cdots \otimes h_{p_d-2}) = \bigotimes_{j=1}^{d} T_{n_j+p_j-2}(h_{p_j-2}),$$

which is generated by the specific function $h_{p_1-2} \otimes \cdots \otimes h_{p_d-2}$ coming from a factorization of the symbol $f_{\boldsymbol{p}}^{(\boldsymbol{\nu})}$ completely analogous to the one considered in Chapter 6.

The chapter is organized as follows. In the remainder of this introductory discussion, we highlight some properties of the symbols $f_{\mathbf{G},\boldsymbol{p}}^{(\boldsymbol{\nu})}$ and $f_{\boldsymbol{p}}^{(\boldsymbol{\nu})}$. Section 7.1 deals with the external PGMRES and shows that a fast (optimal and robust) solver for the general IgA collocation matrix $A_{\mathbf{G},\boldsymbol{n}}^{[\boldsymbol{p}]}$ is obtained if we have a fast solver for the PL-matrix $A_{\boldsymbol{n}}^{[\boldsymbol{p}]}$. Section 7.2 is devoted to the description of the multi-iterative solver of multigrid type for the PL-matrix $A_{\boldsymbol{n}}^{[\boldsymbol{p}]}$ and contains several numerical experiments demonstrating its optimality and robustness.

We recall from Chapter 5 that the spectral symbol $f_{\mathbf{G},\boldsymbol{p}}^{(\boldsymbol{\nu})} : [0,1]^d \times [-\pi,\pi]^d \to \mathbb{R}$ of the normalized matrix-sequence $\{\frac{1}{n^2} A_{\mathbf{G},\boldsymbol{n}}^{(\boldsymbol{\nu})}\}_n$, with $\boldsymbol{n} = \boldsymbol{\nu} n$ and $\boldsymbol{\nu} \in \mathbb{Q}_+^d$, is

$$f_{\mathbf{G},\boldsymbol{p}}^{(\boldsymbol{\nu})} := [\nu_1 \ \cdots \ \nu_d] \left(K_{\mathbf{G}} \circ P_{p_1,\ldots,p_d} \right) [\nu_1 \ \cdots \ \nu_d]^T, \tag{7.2}$$

with

$$\left(P_{p_1,\ldots,p_d}\right)_{rs} := \begin{cases} h_{p_1} \otimes \cdots \otimes h_{p_{r-1}} \otimes f_{p_r} \otimes h_{p_{r+1}} \otimes \cdots \otimes h_{p_d}, & \text{if } r = s, \\ h_{p_1} \otimes \cdots \otimes h_{p_{r-1}} \otimes g_{p_r} \otimes h_{p_{r+1}} \otimes \cdots \otimes h_{p_{s-1}} \otimes g_{p_s} \otimes h_{p_{s+1}} \otimes \cdots \otimes h_{p_d}, & \text{if } r < s, \\ h_{p_1} \otimes \cdots \otimes h_{p_{s-1}} \otimes g_{p_s} \otimes h_{p_{s+1}} \otimes \cdots \otimes h_{p_{r-1}} \otimes g_{p_r} \otimes h_{p_{r+1}} \otimes \cdots \otimes h_{p_d}, & \text{if } r > s; \end{cases}$$

see (5.25) for the expression of $K_{\mathbf{G}}$ and (5.32)–(5.34) for the definitions of h_p, g_p, f_p. In particular,

$$f_{\boldsymbol{p}}^{(\boldsymbol{\nu})} := \sum_{k=1}^{d} \nu_k^2 (h_{p_1} \otimes \cdots \otimes h_{p_{k-1}} \otimes f_{p_k} \otimes h_{p_{k+1}} \otimes \cdots \otimes h_{p_d}) : [-\pi,\pi]^d \to \mathbb{R}$$

is the symbol of the sequence $\{\frac{1}{n^2} A_{\mathbf{G},\boldsymbol{n}}^{[\boldsymbol{p}]}\}_n = \{\frac{1}{n^2} A_{\boldsymbol{n}}^{[\boldsymbol{p}]}\}_n$ obtained when $\overline{\Omega} = \hat{\Omega} = [0,1]^d$, $\mathbf{G} : \hat{\Omega} \to \overline{\Omega}$ is the identity map and K is the identity matrix; see Remark 5.3. $f_{\boldsymbol{p}}^{(\boldsymbol{\nu})}$ only depends on the 'Fourier variables' $\boldsymbol{\theta} = (\theta_1,\ldots,\theta_d) \in [-\pi,\pi]^d$ and, moreover, it is symmetric in each of these variables: $f_{\boldsymbol{p}}^{(\boldsymbol{\nu})}(\pm\theta_1,\ldots,\pm\theta_d) = f_{\boldsymbol{p}}^{(\boldsymbol{\nu})}(\theta_1,\ldots,\theta_d)$. This implies that $f_{\boldsymbol{p}}^{(\boldsymbol{\nu})} : [0,\pi]^d \to \mathbb{R}$, considered on the domain $[0,\pi]^d$, is also a symbol for $\{\frac{1}{n^2} A_{\boldsymbol{n}}^{[\boldsymbol{p}]}\}_n$. The following lemma follows from the properties derived in Chapter 5 (see in particular Lemmas 5.3–5.5).

Lemma 7.1. *Let $\boldsymbol{p} \geq \mathbf{2}$ and $\boldsymbol{\nu} \in \mathbb{Q}_+^d$, then*

$$\left(\frac{2}{\pi}\right)^{\sum_{j=1}^d p_j + d - 2} \min(\nu_1,\ldots,\nu_d)^2 \sum_{k=1}^{d} (2 - 2\cos\theta_k) \leq f_{\boldsymbol{p}}^{(\boldsymbol{\nu})}(\boldsymbol{\theta}) \leq \max(\nu_1,\ldots,\nu_d)^2 \sum_{k=1}^{d} (2 - 2\cos\theta_k).$$

Moreover, setting $M_{f_{\boldsymbol{p}}^{(\boldsymbol{\nu})}} := \max_{\boldsymbol{\theta} \in [0,\pi]^d} f_{\boldsymbol{p}}^{(\boldsymbol{\nu})}(\boldsymbol{\theta})$, for all $j = 1,\ldots,d$ we have

$$f_{\boldsymbol{p}}^{(\boldsymbol{\nu})}(\theta_1,\ldots,\theta_{j-1},\pi,\theta_{j+1},\ldots,\theta_d) \leq 2^{(5-p_j)/2} f_{\boldsymbol{p}}^{(\boldsymbol{\nu})}(\theta_1,\ldots,\theta_{j-1},\tfrac{\pi}{2},\theta_{j+1},\ldots,\theta_d) \leq 2^{(5-p_j)/2} M_{f_{\boldsymbol{p}}^{(\boldsymbol{\nu})}}.$$

In particular, $f_{\boldsymbol{p}}^{(\boldsymbol{\nu})}$ has a unique zero of order two at $\boldsymbol{\theta} = \mathbf{0}$, like the function $\sum_{k=1}^d (2 - 2\cos\theta_k)$, but, for every $j = 1,\ldots,d$, the value $f_{\boldsymbol{p}}^{(\boldsymbol{\nu})}(\theta_1,\ldots,\theta_{j-1},\pi,\theta_{j+1},\ldots,\theta_d)/M_{f_{\boldsymbol{p}}^{(\boldsymbol{\nu})}}$ converges to 0 exponentially when $p_j \to \infty$.

According to Lemma 7.1, the normalized symbol $f_{\boldsymbol{p}}^{(\boldsymbol{\nu})}/M_{f_{\boldsymbol{p}}^{(\boldsymbol{\nu})}}$ has only one actual zero at $\boldsymbol{\theta} = \mathbf{0}$, but, when the spline degrees $\boldsymbol{p}$ are large, it also has infinitely many numerical zeros located at the π-edge points

$$\{\boldsymbol{\theta} \in [0,\pi]^d : \ \exists j \in \{1,\ldots,d\} \text{ with } \theta_j = \pi\}. \tag{7.3}$$

The zero of the symbol at $\boldsymbol{\theta} = \mathbf{0}$ is interpreted by saying that the related IgA collocation matrices $\frac{1}{n^2} A_{\boldsymbol{n}}^{[\boldsymbol{p}]}$ are ill-conditioned in the low frequencies. On the other hand, the fact that the normalized symbol $f_{\boldsymbol{p}}^{(\boldsymbol{\nu})}/M_{f_{\boldsymbol{p}}^{(\boldsymbol{\nu})}}$ shows

infinitely many numerical zeros at the π-edge points (7.3) means that the matrices $\frac{1}{n^2}A_{\boldsymbol{n}}^{[\boldsymbol{p}]}$ are ill-conditioned (for large $\boldsymbol{p}$) also in the high frequencies. The ill-conditioning in the low frequencies is expected, because it is a canonical feature of the symbol associated with the discretization matrices of second-order differential problems like (7.1). However, the ill-conditioning in the high frequencies is not expected and is responsible for the deterioration in the convergence rate of the standard multigrid methods when the approximation parameters $\boldsymbol{p}$ increase. A way to overcome this problem consists in adopting a multi-iterative strategy, as we shall see in Section 7.2.

7.1 Optimal and robust PGMRES for the general IgA collocation matrix $A_{\mathbf{G},\boldsymbol{n}}^{[\boldsymbol{p}]}$

The B-spline discretization matrix related to the Laplacian on the hypercube will be referred to as the Parametric Laplacian matrix (or PL-matrix). This is the matrix coming from the IgA collocation approximation of (7.1) in the case where $\mathbf{G}$ is the identity map (so $\Omega = (0,1)^d$), $K = I$ is the identity matrix, and $\boldsymbol{\beta} = \mathbf{0}$, $\gamma = 0$. In this section we show through numerical experiments that, in many situations, the PL-matrix $A_{\boldsymbol{n}}^{[\boldsymbol{p}]}$ is an optimal and robust GMRES preconditioner for the general IgA collocation matrix $A_{\mathbf{G},\boldsymbol{n}}^{[\boldsymbol{p}]}$ approximating the full elliptic problem (7.1) with arbitrary K, $\boldsymbol{\beta}$, γ, $\mathbf{G}$. This can be explained by means of the theory of GLT sequences (see Subsection 1.4.3), which is a generalization of the standard Fourier Analysis to nonconstant coefficient differential operators, as discussed in [64].

Let us illustrate in the bivariate case $d = 2$, without entering into the details, why the PL-matrix $A_{\boldsymbol{n}}^{[\boldsymbol{p}]}$ should work fairly well as a preconditioner for $A_{\mathbf{G},\boldsymbol{n}}^{[\boldsymbol{p}]}$. From the analysis in Chapter 5, it follows that both $\{\frac{1}{n^2}A_{\mathbf{G},\boldsymbol{n}}^{[\boldsymbol{p}]}\}_n$ and $\{\frac{1}{n^2}A_{\boldsymbol{n}}^{[\boldsymbol{p}]}\}_n$ ($\boldsymbol{n} = n\boldsymbol{\nu}$) are GLT sequences, with corresponding symbols

$$f_{\mathbf{G},\boldsymbol{p}}^{(\boldsymbol{\nu})} = [\nu_1\ \nu_2]\left(K_{\mathbf{G}} \circ P_{p_1,p_2}\right)[\nu_1\ \nu_2]^T,$$

and

$$f_{\boldsymbol{p}}^{(\boldsymbol{\nu})} = [\nu_1\ \nu_2]\left(I \circ P_{p_1,p_2}\right)[\nu_1\ \nu_2]^T,$$

respectively. Since the GLT class is an algebra and since $f_{\boldsymbol{p}}^{(\boldsymbol{\nu})}$ vanishes only at $\boldsymbol{\theta} = \mathbf{0}$ (so that $\{\frac{1}{n^2}A_{\boldsymbol{n}}^{[\boldsymbol{p}]}\}_n$ is sparsely vanishing according to the terminology in [64]), it follows that $\{(\frac{1}{n^2}A_{\boldsymbol{n}}^{[\boldsymbol{p}]})^{-1}\frac{1}{n^2}A_{\mathbf{G},\boldsymbol{n}}^{[\boldsymbol{p}]}\}_n = \{(A_{\boldsymbol{n}}^{[\boldsymbol{p}]})^{-1}A_{\mathbf{G},\boldsymbol{n}}^{[\boldsymbol{p}]}\}_n$ is still a GLT sequence with symbol $(f_{\boldsymbol{p}}^{(\boldsymbol{\nu})})^{-1}f_{\mathbf{G},\boldsymbol{p}}^{(\boldsymbol{\nu})}$:

$$\{(A_{\boldsymbol{n}}^{[\boldsymbol{p}]})^{-1}A_{\mathbf{G},\boldsymbol{n}}^{[\boldsymbol{p}]}\}_n \sim_\lambda (f_{\boldsymbol{p}}^{(\boldsymbol{\nu})})^{-1}f_{\mathbf{G},\boldsymbol{p}}^{(\boldsymbol{\nu})} = \frac{[\nu_1\ \nu_2]\left(K_{\mathbf{G}} \circ P_{p_1,p_2}\right)[\nu_1\ \nu_2]^T}{[\nu_1\ \nu_2]\left(I \circ P_{p_1,p_2}\right)[\nu_1\ \nu_2]^T}. \tag{7.4}$$

Now, suppose that there exist two positive constants c, C such that

$$c\,I \le K_{\mathbf{G}}(\hat{\mathbf{x}}) \le C\,I, \qquad \forall \hat{\mathbf{x}} \in \hat{\Omega}, \tag{7.5}$$

where we recall that the notation $X \ge Y$ means that $X - Y$ is HPSD. Condition (7.5) is equivalent to the following:

$$\min_{\hat{\mathbf{x}}\in\hat{\Omega}} \lambda_{\min}(K_{\mathbf{G}}(\hat{\mathbf{x}})) \ge c > 0, \qquad \max_{\hat{\mathbf{x}}\in\hat{\Omega}} \lambda_{\max}(K_{\mathbf{G}}(\hat{\mathbf{x}})) \le C < \infty. \tag{7.6}$$

Note that (7.5) is usually satisfied in practice. For instance, it is satisfied if

1. $c_K I \le K(\mathbf{x}) \le C_K I$ for some positive constants c_K, C_K and for all $\mathbf{x} \in \overline{\Omega}$,
2. $c_{\mathbf{G}}\,I \le (J_{\mathbf{G}}(\hat{\mathbf{x}}))^T J_{\mathbf{G}}(\hat{\mathbf{x}}) \le C_{\mathbf{G}}\,I$ for some positive constants $c_{\mathbf{G}}$, $C_{\mathbf{G}}$ and for all $\hat{\mathbf{x}} \in \hat{\Omega}$;

in this case we can take $c = c_K/C_{\mathbf{G}}$ and $C = C_K/c_{\mathbf{G}}$. Under the assumption (7.5), Lemma 1.6 yields

$$c\, I \circ P_{p_1,p_2} \le K_{\mathbf{G}} \circ P_{p_1,p_2} \le C\, I \circ P_{p_1,p_2}.$$

This implies that the 'preconditioned symbol' in (7.4) satisfies

$$c \le (f_{\boldsymbol{p}}^{(\nu)})^{-1} f_{\mathbf{G},\boldsymbol{p}}^{(\nu)} \le C,$$

i.e., it is uniformly bounded from above and below by two positive constants C and c, which of course depend on K, $\mathbf{G}$, but not on $\boldsymbol{n}$, $\boldsymbol{p}$. This explains why the PL-matrix $A_{\boldsymbol{n}}^{[\boldsymbol{p}]}$ is expected to be an optimal and robust GMRES preconditioner for $A_{\mathbf{G},\boldsymbol{n}}^{[\boldsymbol{p}]}$. In particular, the PGMRES convergence rate should be independent of $\boldsymbol{n}$ and $\boldsymbol{p}$. This reduces the fast solution of linear systems associated with the general IgA collocation matrix $A_{\mathbf{G},\boldsymbol{n}}^{[\boldsymbol{p}]}$ to the fast solution of linear systems related to the PL-matrix $A_{\boldsymbol{n}}^{[\boldsymbol{p}]}$.

As we will see in Section 7.2, a fast solver is available for systems related to the PL-matrix. The solver is of multi-iterative type, combining a standard multigrid strategy and a certain PGMRES employed as a smoother at the finest level. The first method is effective for approximating the solution especially in the space of low frequencies, where a source of ill-conditioning exists, due to the fact that the symbol $f_{\boldsymbol{p}}^{(\nu)}$ vanishes at $\boldsymbol{\theta} = \mathbf{0}$. The second method is equipped with a specific preconditioner for dampening the high frequency error components, or, equivalently, for approximating the solution in the high frequencies, where another (unexpected) source of ill-conditioning shows up when the spline degrees $\boldsymbol{p}$ are large, due to the presence of the numerical zeros of the normalized symbol $f_{\boldsymbol{p}}^{(\nu)}/M_{f_{\boldsymbol{p}}^{(\nu)}}$ at the π-edge points (7.3). The combination of these two methods, in the spirit of a multi-iterative strategy, leads to a solver whose convergence speed is optimal and robust, i.e., independent of the matrix-size and substantially independent of the other relevant parameters, like the approximation parameters $\boldsymbol{p}$ and the dimensionality d.

In the following examples, we show through numerical experiments the optimality of the PL-matrix $A_{\boldsymbol{n}}^{[\boldsymbol{p}]}$ as a GMRES preconditioner for the general IgA collocation matrix $A_{\mathbf{G},\boldsymbol{n}}^{[\boldsymbol{p}]}$. In all the examples, we use the MATLAB `gmres` function without restarting and with a tolerance of 10^{-6}. The method is started with $\mathbf{u}^{(0)} = \mathbf{0}$ and stopped at the first vector $\mathbf{u}^{(c)}$ whose relative residual in 2-norm is less than 10^{-6}:

$$\|A_{\mathbf{G},\boldsymbol{n}}^{[\boldsymbol{p}]}\mathbf{u}^{(c)} - \mathbf{f}\| \le 10^{-6}\|\mathbf{f}\|. \tag{7.7}$$

Example 1. Consider problem (7.1) in the case $d = 2$, defined on the unit square

$$\Omega = (0,1)^2, \qquad \mathbf{G}(\hat{x},\hat{y}) = (\hat{x},\hat{y}),$$

with

$$K(x,y) = \begin{bmatrix} (2+\cos x)(1+y) & \cos(x+y)\sin(x+y) \\ \cos(x+y)\sin(x+y) & (2+\sin y)(1+x) \end{bmatrix},$$
$$\boldsymbol{\beta}(x,y) = \begin{bmatrix} 11+\sin x + y\sin x - 2\cos^2(x+y) \\ -9-\cos y - x\cos y - 2\cos^2(x+y) \end{bmatrix},$$
$$\gamma(x,y) = f(x,y) = 1.$$

To solve the linear system $A_{\mathbf{G},n_1,n_2}^{[p_1,p_2]}\mathbf{u} = \mathbf{f}$ resulting from the IgA collocation approximation of this problem, the GMRES method was applied first without preconditioning and then with the PL-matrix as preconditioner. The results are collected in Table 7.1. We note that the PGMRES has an optimal and robust convergence rate, completely independent of n and p. This is in contrast with the behavior of the simple GMRES, whose convergence rate worsens with respect to both n and p and, in particular, grows linearly with n (the system size is $(n+p-2)^2 \sim n^2$). From Table 7.1 we can conclude that the PL-matrix is an optimal and robust GMRES preconditioner for the general IgA collocation matrix.

$n_1 \times n_2$	$p=2$		$p=3$		$p=4$		$p=5$		$p=6$		$p=7$		$p=8$		$p=9$	
	I	P	I	P	I	P	I	P	I	P	I	P	I	P	I	P
20×20	51	8	59	8	71	8	83	8	94	8	104	8	130	8	166	8
30×30	76	8	88	8	105	8	122	8	137	8	148	8	159	8	196	8
40×40	102	8	117	8	138	8	161	8	180	8	193	8	204	8	219	8
50×50	128	8	146	8	172	8	200	8	223	8	239	8	252	8	266	8
60×60	154	8	176	8	206	8	239	8	266	8	284	8	300	8	316	8

Table 7.1: Example 1: number of GMRES iterations without (I) and with (P) preconditioning for solving $A^{[p_1,p_2]}_{\mathbf{G},n_1,n_2}\mathbf{u} = \mathbf{f}$ up to a precision of 10^{-6}, varying the fineness parameter $n_1 = n_2 = n$ and the spline degree $p_1 = p_2 = p$.

$n_1 \times n_2$	$p=2$		$p=3$		$p=4$		$p=5$		$p=6$		$p=7$		$p=8$		$p=9$	
	I	P	I	P	I	P	I	P	I	P	I	P	I	P	I	P
20×20	62	17	66	17	71	18	81	19	91	19	100	19	108	20	114	20
30×30	97	17	103	18	106	18	120	19	133	19	145	19	155	19	163	20
40×40	134	18	139	18	141	18	159	19	177	19	192	19	204	19	214	19
50×50	170	18	176	18	176	18	199	19	221	19	239	19	253	19	265	19
60×60	208	18	214	18	211	18	239	19	265	19	286	20	302	19	316	19

Table 7.2: Example 2: number of GMRES iterations without (I) and with (P) preconditioning for solving $A^{[p_1,p_2]}_{\mathbf{G},n_1,n_2}\mathbf{u} = \mathbf{f}$ up to a precision of 10^{-6}, varying the fineness parameter $n_1 = n_2 = n$ and the spline degree $p_1 = p_2 = p$.

Example 2. Consider problem (7.1) in the case $d = 2$, defined on a quarter of annulus

$$\Omega = \{(x,y) \in \mathbb{R}^2 : \ r^2 < x^2 + y^2 < R^2, \ x > 0, \ y > 0\}, \qquad r = 1, \qquad R = 4,$$

with

$$\mathbf{G}(\hat{x},\hat{y}) = (x,y), \qquad \begin{cases} x = [r + \hat{x}(R-r)]\cos(\frac{\pi}{2}\hat{y}), \\ y = [r + \hat{x}(R-r)]\sin(\frac{\pi}{2}\hat{y}). \end{cases}$$

Note that the map $\mathbf{G}$ provides an exact representation of the domain Ω, but is not expressed in terms of tensor-product B-splines. In this sense, our analysis is general, since we are not restricted to use a B-spline approximation of the domain (following the isoparametric approach), but we may use any exact representation of the domain. Moreover, we take

$$K(x,y) = \begin{bmatrix} (2+\cos x)(1+y) & \cos(x+y)\sin(x+y) \\ \cos(x+y)\sin(x+y) & (2+\sin y)(1+x) \end{bmatrix}, \qquad \boldsymbol{\beta}(x,y) = \begin{bmatrix} -5y \\ 5x \end{bmatrix}, \qquad \gamma(x,y) = xy,$$

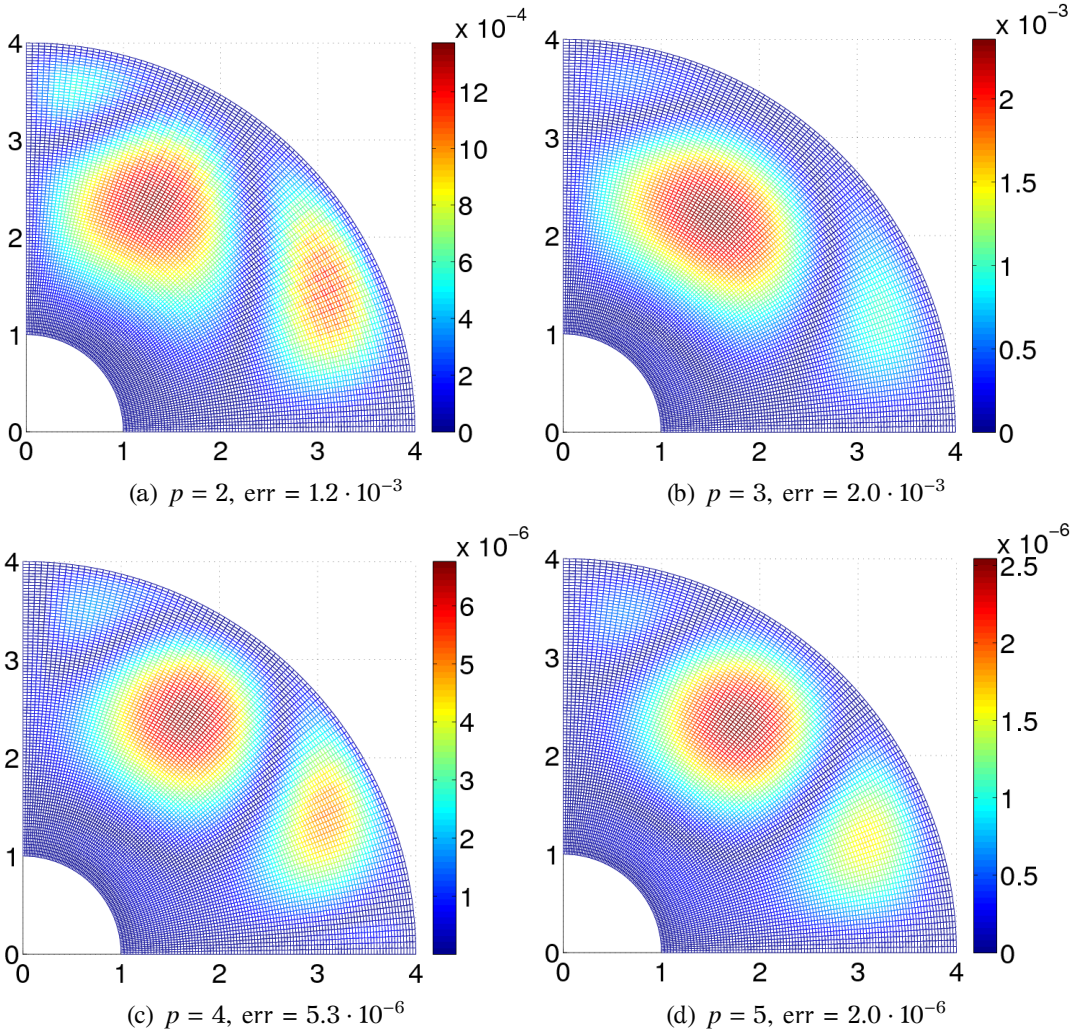
and $f(x,y)$ computed from the exact solution

$$u(x,y) = (x^2 + y^2 - 1)(x^2 + y^2 - 16)\sin x \, \sin y.$$

To solve the corresponding B-spline IgA collocation linear system $A^{[p_1,p_2]}_{\mathbf{G},n_1,n_2}\mathbf{u} = \mathbf{f}$, the GMRES method was applied first without preconditioning and then with the PL-matrix as preconditioner. The results are collected in Table 7.2, and they clearly indicate that the PL-matrix is an optimal and robust GMRES preconditioner for the general IgA collocation matrix $A^{[p_1,p_2]}_{\mathbf{G},n_1,n_2}$.

When increasing p, but keeping n fixed, the number of PGMRES iterations to reach the preassigned accuracy 10^{-6} is slowly increasing for moderate n, whereas it seems practically constant for large n: the observed convergence rate is about $10^{-6/19} \approx 0.483$. In this example, we also computed the best constants for which the relations (7.5)–(7.6) are satisfied, i.e.

$$c := \min_{(\hat{x},\hat{y})\in\hat{\Omega}} \lambda_{\min}(K_{\mathbf{G}}(\hat{x},\hat{y})) \approx 0.111, \qquad C := \max_{(\hat{x},\hat{y})\in\hat{\Omega}} \lambda_{\max}(K_{\mathbf{G}}(\hat{x},\hat{y})) \approx 2.436. \tag{7.8}$$



(a) $p = 2$, err $= 1.2 \cdot 10^{-3}$ (b) $p = 3$, err $= 2.0 \cdot 10^{-3}$

(c) $p = 4$, err $= 5.3 \cdot 10^{-6}$ (d) $p = 5$, err $= 2.0 \cdot 10^{-6}$

Figure 7.1: Example 2: error of the computed solution for $n = 30$ varying p, where err $= \|u - \tilde{u}\|/\|u\|$.

Let us assume that $(A_{n,n}^{[p,p]})^{-1} A_{\mathbf{G},n,n}^{[p,p]}$ is 'almost' symmetric positive definite with its spectrum contained in $[c, C]$. Note that this makes sense because the corresponding symbol $(f_{p,p}^{(1,1)})^{-1} f_{\mathbf{G},p,p}^{(1,1)}$ is nonnegative with range in $[c, C]$. Then, the classical GMRES convergence analysis based on the values (7.8) provides an upper bound of 0.648 for the asymptotic convergence rate; see [52, Proposition 6.32] and recall the classical estimate for the quantity $\epsilon^{(m)}$ appearing in the proposition, which in our case becomes

$$\epsilon^{(m)} \leq 2 \left(\frac{\sqrt{C/c} - 1}{\sqrt{C/c} + 1} \right)^m .$$

Luckily, the observed convergence rate 0.483 is even better. Thus, the presence of eigenvalues with small imaginary part, the existence of outliers, and the fact that the matrix of the eigenvectors of $(A_{n,n}^{[p,p]})^{-1} A_{\mathbf{G},n,n}^{[p,p]}$ is not exactly unitary do not seem to negatively influence the observed convergence rate.

We conclude the numerical example by showing in Figure 7.1 the error $|u(x, y) - \tilde{u}(x, y)|/\|u\|_\infty$, where $\tilde{u}(x, y)$ is the computed solution, for $n = 30$ and for different values of p. The 2-norm of the relative error is also given in the figure.

Example 3. Consider problem (7.1) in the case $d = 3$, defined on the unit cube

$$\Omega = (0, 1)^3, \qquad \mathbf{G}(\hat{x}, \hat{y}, \hat{z}) = (\hat{x}, \hat{y}, \hat{z}),$$

$n_1 \times n_2 \times n_3$	$p = 2$	$p = 3$	$p = 4$	$p = 5$	$p = 6$	$p = 7$	$p = 8$	$p = 9$
$15 \times 15 \times 15$	20	21	22	23	24	25	26	27
$20 \times 20 \times 20$	20	22	23	23	24	24	25	26
$25 \times 25 \times 25$	21	22	23	23	24	24	25	25
$30 \times 30 \times 30$	21	22	23	23	24	24	24	25

Table 7.3: Example 3: number of PGMRES iterations for solving $A^{[p_1,p_2,p_3]}_{\mathbf{G},n_1,n_2,n_3}\mathbf{u} = \mathbf{f}$ up to a precision of 10^{-6}, varying the mesh size $n_1 \times n_2 \times n_3$ and the spline degree $p_1 = p_2 = p_3 = p$.

with

$$K(x,y,z) = \begin{bmatrix} e^{xyz} & \frac{xy}{2} & \frac{xz}{2} \\ \frac{xy}{2} & e^{x+y+z} & \frac{yz}{2} \\ \frac{xz}{2} & \frac{yz}{2} & xyz+3 \end{bmatrix}, \qquad \boldsymbol{\beta}(x,y,z) = \begin{bmatrix} 5xy+z \\ -10yz+x \\ 5xz+y \end{bmatrix}, \qquad \gamma(x,y,z) = \frac{x^2y-y^3}{1+z}, \qquad f(x,y,z) = 1.$$

We solved the corresponding B-spline IgA collocation linear system $A^{[p_1,p_2,p_3]}_{\mathbf{G},n_1,n_2,n_3}\mathbf{u} = \mathbf{f}$ by means of the PGMRES method, with the PL-matrix as preconditioner. The results are collected in Table 7.3, and once again, they show that the PL-matrix is an optimal and robust GMRES preconditioner for the general IgA collocation matrix.

7.2 Optimal and robust multi-iterative multigrid solver for the PL-matrix $A^{[\boldsymbol{p}]}_{\boldsymbol{n}}$

Let us consider the linear system

$$A^{[\boldsymbol{p}]}_{\boldsymbol{n}}\mathbf{u} = \mathbf{f} \tag{7.9}$$

coming from the IgA collocation approximation of the d-dimensional problem (7.1) with $\boldsymbol{n} := (n_1, \dots, n_d)$, $\boldsymbol{p} := (p_1, \dots, p_d)$, in the case where $K = I$, $\boldsymbol{\beta} = \mathbf{0}$, $\gamma = 0$, $\mathbf{G}$ is the identity map on the parametric domain $\hat{\Omega}$, and $f = 1$. The matrix in (7.9) is just the PL-matrix. In this section we present optimal and robust two-grid and multigrid methods to solve the linear system (7.9). The used machinery is very similar to the work in Chapter 6 in the IgA Galerkin context, so we refer the reader to Chapter 6 for a description of the tools.

7.2.1 Two-grid

We consider the two-grid method $TG((\text{PGMRES})^{s^{[\boldsymbol{p}]}}, P^{[\boldsymbol{p}]}_{\boldsymbol{n}})$ which is formed by:

1. a canonical coarse-grid correction, with standard full-weighting projector

$$P^{[\boldsymbol{p}]}_{\boldsymbol{n}} := P_{\boldsymbol{n}+\boldsymbol{p}-\mathbf{2}} = P_{n_1+p_1-2} \otimes \cdots \otimes P_{n_d+p_d-2}, \tag{7.10}$$

as given by (6.21) for $\boldsymbol{m} = \boldsymbol{n} + \boldsymbol{p} - \mathbf{2}$. We recall that P_m is defined for any odd $m \geq 3$ by

$$P_m := \frac{1}{2}\begin{bmatrix} 1 & 2 & 1 & & & & \\ & & 1 & 2 & 1 & & \\ & & & & & \ddots & \\ & & & & & 1 & 2 & 1 \end{bmatrix} \in \mathbb{R}^{\frac{m-1}{2} \times m}. \tag{7.11}$$

The prolongation operator is just the transpose of the projector (7.10), so that the coarse-grid correction matrix is

$$CGC := I - (P^{[\boldsymbol{p}]}_{\boldsymbol{n}})^T \left(P^{[\boldsymbol{p}]}_{\boldsymbol{n}} A^{[\boldsymbol{p}]}_{\boldsymbol{n}} (P^{[\boldsymbol{p}]}_{\boldsymbol{n}})^T\right)^{-1} P^{[\boldsymbol{p}]}_{\boldsymbol{n}} A^{[\boldsymbol{p}]}_{\boldsymbol{n}};$$

n	$c_n^{[2]}$ [2]	$c_n^{[4]}$ [3]	$c_n^{[6]}$ [4]	$c_n^{[8]}$ [5]
81	6	6	5	4
161	7	6	5	4
321	7	6	5	4
641	7	6	5	4
1281	7	6	5	5
2561	7	6	5	5

n	$c_n^{[3]}$ [2]	$c_n^{[5]}$ [3]	$c_n^{[7]}$ [4]	$c_n^{[9]}$ [5]
80	8	6	5	4
160	8	6	5	4
320	9	7	5	4
640	9	7	5	4
1280	9	7	5	4
2560	9	7	6	5

Table 7.4: number of iterations $c_n^{[p]}$ needed by the two-grid method $TG((\text{PGMRES})^{s^{[p]}}, P_n^{[p]})$ for solving $(1/n^2)A_n^{[p]}\mathbf{u} = \mathbf{f}/n^2$ up to a precision of 10^{-6}. The parameter $s^{[p]}$ is specified between brackets $[\cdot]$.

n	$c_{n,n}^{[2,2]}$ [2]	$c_{n,n}^{[4,4]}$ [4]	$c_{n,n}^{[6,6]}$ [5]	$c_{n,n}^{[8,8]}$ [7]
21	6	6	7	8
41	6	6	7	8
61	6	6	7	8
81	6	6	7	8
101	6	6	7	8

n	$c_{n,n}^{[3,3]}$ [2]	$c_{n,n}^{[5,5]}$ [4]	$c_{n,n}^{[7,7]}$ [6]	$c_{n,n}^{[9,9]}$ [9]
20	8	6	7	7
40	8	7	7	8
60	8	7	7	7
80	8	7	7	7
100	8	7	7	7

Table 7.5: number of iterations $c_{n,n}^{[p,p]}$ needed by the two-grid method $TG((\text{PGMRES})^{s^{[p,p]}}, P_{n,n}^{[p,p]})$ for solving $(1/n^2)A_{n,n}^{[p,p]}\mathbf{u} = \mathbf{f}/n^2$ up to a precision of 10^{-6}. The parameter $s^{[p,p]}$ is specified between brackets $[\cdot]$.

2. $s^{[\boldsymbol{p}]}$ post-smoothing iterations by the PGMRES with preconditioner

$$T_{\boldsymbol{n}+\boldsymbol{p}-\boldsymbol{2}}(h_{p_1-2}\otimes\cdots\otimes h_{p_d-2}) = T_{n_1+p_1-2}(h_{p_1-2})\otimes\cdots\otimes T_{n_d+p_d-2}(h_{p_d-2}). \tag{7.12}$$

In Tables 7.4 and 7.5, we solved the (normalized) system (7.9) for $d = 1, 2$, using the two-grid method $TG((\text{PGMRES})^{s^{[p]}}, P_{\boldsymbol{n}}^{[\boldsymbol{p}]})$. The two-grid procedure has been started with $\mathbf{u}^{(0)} = \mathbf{0}$ and stopped at the first vector $\mathbf{u}^{(c)}$ whose relative residual in 2-norm is less than 10^{-6}; cf. (7.7).

Let us give a motivation for the choice of our two-grid method through the symbol. We first consider the case $d = 1$ and then we generalize the argument to the case $d \geq 2$. We will not provide all the necessary details, since they were already described in Section 6.2 in the (analogous) context of Galerkin IgA. For a better understanding of the following discussion, the reader is recommended to read Chapter 6 first.

For $d = 1$, the symbol of $\{\frac{1}{n^2}A_n^{[p]}\}_n$ is

$$f_p(\theta) = (2 - 2\cos\theta)h_{p-2}(\theta).$$

As already pointed out (see the discussion after Lemma 7.1), the symbol f_p has a unique zero at $\theta = 0$, implying that the low frequency subspace is ill-conditioned for $\frac{1}{n^2}A_n^{[p]}$. However, this ill-conditioning in low frequencies is canonical when dealing with matrices coming from the approximation of elliptic problems like (7.1), and, in fact, it causes no problems for any standard two-grid or multigrid procedure which employs the usual full-weighting projector $P_n^{[p]}$. Indeed, $P_n^{[p]}$ is designed to be highly contractive in low frequencies and hence any classical two-grid or multigrid method using such a projector combined with any standard smoother (e.g. Gauss-Seidel) will have a convergence rate independent of the matrix size. However, when p is large, a numerical zero of the (normalized) symbol f_p/M_{f_p} occurs at $\theta = \pi$; see Lemma 7.1, Figure 5.1 and Table 5.1. Therefore, for large p, also the high frequency subspace is ill-conditioned for $\frac{1}{n^2}A_n^{[p]}$, and this non-canonical ill-conditioning in high frequencies is completely ignored by the full-weighting projector $P_n^{[p]}$. This is the reason why classical two-grid and multigrid procedures with full-weighting projector and standard Gauss-Seidel smoother have a convergence rate that, despite being independent of the matrix size, worsens with p.

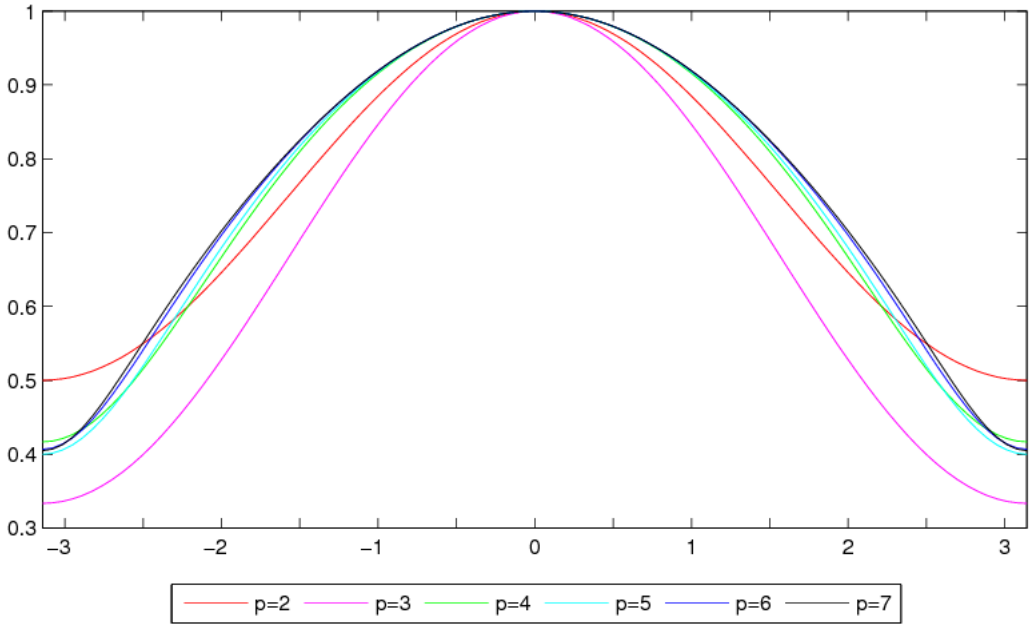


Figure 7.2: graph of $w_p(\theta) := h_p(\theta)/h_{p-2}(\theta)$ for $p = 2, \ldots, 7$.

The choice of using as a smoother, instead of the Gauss-Seidel method, the PGMRES with preconditioner $T_{n+p-2}(h_{p-2})$ (as given by (7.12) for $d = 1$, $\boldsymbol{p} = p$, $\boldsymbol{n} = n$), is made in order to gain a p-independent convergence rate. Actually, we see from Table 7.4 that the resulting two-grid method is quite successful, its convergence rate being independent of both n and p. This success was not unexpected. Indeed, the idea of using the preconditioner $T_{n+p-2}(h_{p-2})$ follows from the observation that $[h_{p-2}(\theta)]^{-1} f_p(\theta) = 2 - 2\cos\theta$ is p-independent, which means that the symbol h_{p-2} of the Toeplitz preconditioner $T_{n+p-2}(h_{p-2})$ 'erases' the numerical zero of the symbol f_p of $\frac{1}{n^2} A_n^{[p]}$ at $\theta = \pi$; see Subsection 6.2.5 for more detailed explanations. Therefore, we expect that

- the PGMRES alone for $\frac{1}{n^2} A_n^{[p]}$ has a convergence rate substantially independent of p but worsening with n;
- the standard two-grid and multigrid procedures with full-weighting projector and classical smoothers (e.g. Gauss-Seidel) have a convergence rate independent of n but worsening with p;
- the combination of this two methods in a unique two-grid or multigrid procedure has a convergence rate independent of both n and p, according to the multi-iterative idea.

In the case $d = 2$, the symbol of $\{\frac{1}{n^2} A_{n_1,n_2}^{[p_1,p_2]}\}_n$,with $n_1 = \nu_1 n$ and $n_2 = \nu_2 n$, is

$$\begin{aligned} f_{p_1,p_2}^{(\nu_1,\nu_2)}(\theta_1,\theta_2) &= \nu_1^2 f_{p_1}(\theta_1) h_{p_2}(\theta_2) + \nu_2^2 h_{p_1}(\theta_1) f_{p_2}(\theta_2) \\ &= h_{p_1-2}(\theta_1) h_{p_2-2}(\theta_2) \left[\nu_1^2 (2 - 2\cos\theta_1) \frac{h_{p_2}(\theta_2)}{h_{p_2-2}(\theta_2)} + \nu_2^2 \frac{h_{p_1}(\theta_1)}{h_{p_1-2}(\theta_1)} (2 - 2\cos\theta_2) \right] \\ &= (h_{p_1-2} \otimes h_{p_2-2})(\theta_1,\theta_2) \left[\nu_1^2 (2 - 2\cos\theta_1) w_{p_2}(\theta_2) + \nu_2^2 w_{p_1}(\theta_1) (2 - 2\cos\theta_2) \right], \end{aligned}$$

where $w_p(\theta) := h_p(\theta)/h_{p-2}(\theta)$, and, for general $d \geq 2$, the symbol of $\{\frac{1}{n^2}A_{\boldsymbol{n}}^{[\boldsymbol{p}]}\}_n$, with $\boldsymbol{n} = \boldsymbol{\nu}n = (\nu_1 n, \ldots, \nu_d n)$, is

$$\begin{aligned} f_{\boldsymbol{p}}^{(\boldsymbol{\nu})}(\boldsymbol{\theta}) &= \sum_{k=1}^{d} \nu_k^2 (h_{p_1} \otimes \cdots \otimes h_{p_{k-1}} \otimes f_{p_k} \otimes h_{p_{k+1}} \otimes \cdots \otimes h_{p_d})(\boldsymbol{\theta}) \\ &= (h_{p_1-2} \otimes \cdots \otimes h_{p_d-2})(\boldsymbol{\theta}) \sum_{k=1}^{d} \nu_k^2 w_{p_1}(\theta_1) \cdots w_{p_{k-1}}(\theta_{k-1})(2-2\cos\theta_k) w_{p_{k+1}}(\theta_{k+1}) \cdots w_{p_d}(\theta_d). \end{aligned}$$

We see from Figure 7.2 that the function w_p is p-independent, in the sense that it is uniformly bounded from above and below by two positive constants independent of p. Actually, it seems that w_p converges uniformly to some function with range in $[0.4, 1]$. Therefore, the idea of using the PGMRES with preconditioner (7.12) as smoother has the same motivation as for the case $d = 1$: the preconditioned symbol

$$[(h_{p_1-2} \otimes \cdots \otimes h_{p_d-2})(\boldsymbol{\theta})]^{-1} f_{\boldsymbol{p}}^{(\boldsymbol{\nu})}(\boldsymbol{\theta}) = \sum_{k=1}^{d} \nu_k^2 w_{p_1}(\theta_1) \cdots w_{p_{k-1}}(\theta_{k-1})(2-2\cos\theta_k) w_{p_{k+1}}(\theta_{k+1}) \cdots w_{p_d}(\theta_d)$$

is $\boldsymbol{p}$-independent. This implies that the numerical zeros of $f_{\boldsymbol{p}}^{(\boldsymbol{\nu})}$ at the π-edge points (7.3) are completely 'erased' by the symbol $h_{p_1-2} \otimes \cdots \otimes h_{p_d-2}$ of the Toeplitz preconditioner (7.12), and this motivates why such a preconditioner is effective; we refer again to Subsection 6.2.5 for more detailed explanations.

Finally, we point out that the preconditioner (7.12) is effectively solvable. Indeed, due to the tensor-product structure and to the bandedness of the matrices $T_{n_j+p_j-2}(h_{p_j-2})$, the computational cost for solving a linear system with matrix (7.12) is linear in the matrix size $N(\boldsymbol{n}+\boldsymbol{p}-\boldsymbol{2}) = \prod_{j=1}^{d}(n_j + p_j - 2)$.

7.2.2 Multigrid: V-cycle and W-cycle

We now focus on the V-cycle and W-cycle multigrid methods for the PL-matrix $\frac{1}{n^2}A_{\boldsymbol{n}}^{[\boldsymbol{p}]}$ formed by:

1. standard coarse-grid corrections at each level, which use, as restriction operator, the full-weighting restriction (6.21) (with properly adjusted size), and, as prolongation operator, the transpose of the projector;
2. $s^{[\boldsymbol{p}]}$ post-smoothing iterations by the PGMRES with preconditioner (7.12) at the finest level, and a single standard Gauss-Seidel post-smoothing iteration at all the other levels.

Let us assume that $\boldsymbol{n} = (n, \ldots, n)$ and $\boldsymbol{p} = (p, \ldots, p)$. We denote by index 0 the finest level and by index $\ell_n^{[p]} := \log_2(n+p-1) - 1$ the coarsest level. Let $A_{\boldsymbol{n},i}^{[\boldsymbol{p}]}$ be the matrix at level i, whose dimension is $(m_{n,i}^{[p]})^d$, $0 \leq i \leq \ell_n^{[p]}$, with $m_{n,i}^{[p]} := \frac{n+p-1}{2^i} - 1$. In this notation, we have $A_{\boldsymbol{n},0}^{[\boldsymbol{p}]} = (1/n^2)A_{\boldsymbol{n}}^{[\boldsymbol{p}]}$ and

$$A_{\boldsymbol{n},i+1}^{[\boldsymbol{p}]} = P_{\boldsymbol{n},i}^{[\boldsymbol{p}]} A_{\boldsymbol{n},i}^{[\boldsymbol{p}]} (P_{\boldsymbol{n},i}^{[\boldsymbol{p}]})^T, \quad i = 0, \ldots, \ell_n^{[p]} - 1,$$

where

$$P_{\boldsymbol{n},i}^{[\boldsymbol{p}]} := P_{m_{n,i}^{[p]}} \otimes \cdots \otimes P_{m_{n,i}^{[p]}}, \quad i = 0, \ldots, \ell_n^{[p]} - 1,$$

is the full-weighting projector at level i, and P_m is defined in (7.11). Regarding the smoother, at each coarse level $i \geq 1$ we used one single post-smoothing iteration with the standard Gauss-Seidel method; at the finest level $i = 0$ we used $s^{[\boldsymbol{p}]}$ post-smoothing iterations by the PGMRES with preconditioner (7.12). At each level i, we first performed a coarse-grid correction, with one recursive call in the V-cycle and two recursive calls in the W-cycle, and then we applied one post-smoothing iteration by the Gauss-Seidel method (if $i \geq 1$), or $s^{[\boldsymbol{p}]}$ post-smoothing iterations by the PGMRES with preconditioner (7.12) (if $i = 0$).

n	$c_n^{[2]}$	[2]	n	$c_n^{[4]}$	[3]	n	$c_n^{[6]}$	[4]	n	$c_n^{[8]}$	[5]
15	7	6	13	10	6	11	9	5	9	8	5
31	8	6	29	11	6	27	13	6	25	13	5
63	9	6	61	12	6	59	14	6	57	17	5
127	10	7	125	13	7	123	15	6	121	17	5
255	10	7	253	13	7	251	15	6	249	18	5
511	11	7	509	14	7	507	16	6	505	18	5
1023	11	7	1021	14	7	1019	17	6	1017	19	6
n	$c_n^{[3]}$	[2]	n	$c_n^{[5]}$	[3]	n	$c_n^{[7]}$	[4]	n	$c_n^{[9]}$	[5]
14	9	8	12	10	6	10	8	5	8	7	5
30	10	8	28	13	6	26	13	5	24	12	5
62	11	8	60	13	6	58	16	6	56	17	5
126	12	8	124	14	7	122	16	6	120	19	5
254	12	9	252	15	7	250	17	6	248	19	5
510	12	9	508	15	7	506	17	6	504	19	5
1022	13	9	1020	16	7	1018	18	6	1016	20	5

Table 7.6: number of iterations $c_n^{[p]}$ needed for solving $(1/n^2)A_n^{[p]}\mathbf{u} = \mathbf{f}/n^2$ up to a precision of 10^{-6}, when using the multigrid cycle with $s^{[p]}$ post-smoothing steps by the PGMRES at the finest level and 1 post-smoothing step by standard Gauss-Seidel at the coarse levels. The parameter $s^{[p]}$ is specified between brackets $[\cdot]$. The methods have been started with $\mathbf{u}^{(0)} = \mathbf{0}$ and stopped at the first term $\mathbf{u}^{(c)}$ satisfying the relative criterion (7.7). For each pair (p, n), the first entry corresponding to $c_n^{[p]}$ refers to the V-cycle, the second entry to the W-cycle.

n	$c_{n,n}^{[2,2]}$	[2]	n	$c_{n,n}^{[4,4]}$	[4]	n	$c_{n,n}^{[6,6]}$	[5]	n	$c_{n,n}^{[8,8]}$	[7]
15	7	6	13	9	6	11	8	6	9	8	7
31	8	6	29	11	6	27	12	7	25	12	8
63	9	6	61	12	7	59	14	7	57	16	8
127	9	6	125	13	7	123	15	7	121	17	9
255	10	7	253	13	7	251	16	8	249	18	9
n	$c_{n,n}^{[3,3]}$	[2]	n	$c_{n,n}^{[5,5]}$	[4]	n	$c_{n,n}^{[7,7]}$	[6]	n	$c_{n,n}^{[9,9]}$	[9]
14	8	8	12	9	7	10	8	6	8	9	9
30	10	8	28	12	7	26	12	7	24	11	7
62	10	8	60	13	7	58	15	7	56	16	8
126	11	8	124	14	8	122	16	8	120	18	8
254	12	8	252	15	7	250	17	7	248	19	8

Table 7.7: number of iterations $c_{n,n}^{[p,p]}$ needed for solving $(1/n^2)A_{n,n}^{[p,p]}\mathbf{u} = \mathbf{f}/n^2$ up to a precision of 10^{-6}, when using the multigrid cycle with $s^{[p,p]}$ post-smoothing steps by the PGMRES at the finest level and 1 post-smoothing step by standard Gauss-Seidel at the coarse levels. The parameter $s^{[p,p]}$ is specified between brackets $[\cdot]$. The methods have been started with $\mathbf{u}^{(0)} = \mathbf{0}$ and stopped at the first term $\mathbf{u}^{(c)}$ satisfying the relative criterion (7.7). For each pair (p, n), the first entry corresponding to $c_{n,n}^{[p,p]}$ refers to the V-cycle, the second entry to the W-cycle.

n	$c_{n,n,n}^{[2,2,2]}$	[2]	n	$c_{n,n,n}^{[4,4,4]}$	[4]	n	$c_{n,n,n}^{[6,6,6]}$	[6]	n	$c_{n,n,n}^{[8,8,8]}$	[9]
15	6	6	13	8	6	11	7	6	9	9	9
31	8	6	29	10	6	27	10	6	25	10	8
63	9	6	61	11	6	59	13	7	57	14	8

n	$c_{n,n,n}^{[3,3,3]}$	[2]	n	$c_{n,n,n}^{[5,5,5]}$	[4]	n	$c_{n,n,n}^{[7,7,7]}$	[7]	n	$c_{n,n,n}^{[9,9,9]}$	[12]
14	8	7	12	8	7	10	8	7	8	9	9
30	9	8	28	11	7	26	10	8	24	9	7
62	10	8	60	12	7	58	14	8	56	14	8

Table 7.8: number of iterations $c_{n,n,n}^{[p,p,p]}$ needed for solving $(1/n^2)A_{n,n,n}^{[p,p,p]}\mathbf{u} = \mathbf{f}/n^2$ up to a precision of 10^{-6}, when using the multigrid cycle with $s^{[p,p,p]}$ post-smoothing steps by the PGMRES at the finest level and 1 post-smoothing step by standard Gauss-Seidel at the coarse levels. The parameter $s^{[p,p,p]}$ is specified between brackets $[\cdot]$. The methods have been started with $\mathbf{u}^{(0)} = \mathbf{0}$ and stopped at the first term $\mathbf{u}^{(c)}$ satisfying the relative criterion (7.7). For each pair (p, n), the first entry corresponding to $c_{n,n,n}^{[p,p,p]}$ refers to the V-cycle, the second entry to the W-cycle.

We observe that these V-cycle and W-cycle essentially coincide with those considered in Chapter 6 (Section 6.6) for the IgA Galerkin PL-matrix, with the only difference that now, at the finest level, we use a PGMRES smoother instead of a PCG smoother, because of the non-symmetry of the IgA collocation PL-matrix.

In Tables 7.6, 7.7, 7.8 we solved the (normalized) system (7.9) for $d = 1, 2, 3$, using the V-cycle and W-cycle multigrid methods described above. We see that the number of V-cycle and W-cycle iterations for reaching the preassigned accuracy 10^{-6} is substantially independent of all the relevant parameters: p, n, d. In particular, the convergence rate of the W-cycle is practically constant (the number of iterations is around 8). The only unpleasant fact is that the number of PGMRES post-smoothing steps $s^{[p,\ldots,p]}$ needed for keeping a fixed number of W-cycle iterations around 8 slightly increases when p and d increase. However, we should also say that, if we decrease $s^{[p,\ldots,p]}$ a little bit, the number of iterations does not increase so much. For instance, if in Table 7.8 we chose $s^{[9,9,9]} = 9$ (instead of $s^{[9,9,9]} = 12$), then the resulting number of W-cycle iterations $c_{n,n,n}^{[9,9,9]}$ for $n = 56$ would be 12 (instead of 8).

Conclusion

In the first part of this thesis (Chapter 2), we provided new tools for computing the asymptotic spectral distribution of matrix-sequences $\{A_n\}$. Then, in Chapters 3–5, we considered the sequences of matrices $\{A_n\}$ associated with the numerical approximation of elliptic PDE by means of various numerical methods: from the classical $\mathbb{Q}_p$ Lagrangian FEM to more recent techniques based on the IgA paradigm, such as the Galerkin B-spline IgA and the B-spline IgA Collocation Method. For each of these matrix-sequences $\{A_n\}$, we computed the corresponding spectral symbol in the sense of Definition 1.1, and we studied its properties in considerable detail. Afterwards, in Chapters 6–7, we used the properties of the symbol to design fast (optimal and robust) multi-iterative solvers of multigrid type for the matrices A_n associated with the IgA-based methods.

It is clear that the nature of this thesis is at the same time classificatory and applicative. In fact, a precise target of all this work is to show that, whenever a linear PDE is given and a linear numerical method for its approximation is chosen, one may ask if a spectral distribution for the corresponding sequence of discretization matrices A_n exists. Usually, the answer is 'yes' and the computation of the symbol describing the spectral distribution can be carried out by using the huge 'GLT machinery', of which here we have seen particular examples of applications. In this sense, the present thesis is classificatory: we chose specific PDE and numerical methods, and we determined the symbol for the resulting discretization matrices A_n. However, we did not limit ourselves to find the symbol: we also studied its properties and used them for designing fast solvers for the matrices A_n. Here is the applicative nature of our work.

From this discussion, it is clear that a lot of open problems remain, because a lot of PDE and numerical methods have not been investigated yet: the 'classification' is still incomplete, since a lot of PDE and numerical methods are still waiting for their symbol! We list some open problems in the following.

1. Compute (and study) the symbol of the matrices arising from the Galerkin B-spline IgA approximation of the full elliptic PDE (5.1). Note that such a symbol has not been computed in this thesis, because in Chapter 4, where we considered the Galerkin B-spline IgA, we only focused on the constant-coefficient PDE (4.1).

2. Compute (and study) the symbol of the matrices arising from the approximation of (5.1) by means of Galerkin-type methods based on B-splines with reduced smoothness. This has been partially done in [32], but without any rigorous theoretical justification and, in any case, the problem (3) addressed in [32] is much simpler than (5.1). Moreover, [32] does not contain a careful study of the symbol, which would shed light on the asymptotic spectral properties of the considered matrices.

3. Compute (and study) the symbol of the matrices arising from the approximation of (4.1) and (5.1) by means of Finite Element Methods that use other bases than the Lagrangian one. It is known in the FEM community that choosing the Lagrangian basis with uniform knots (as in Chapter 3) is a simple but unfortunate choice, due to the instability of the Lagrangian interpolation. A much more interesting basis is, for instance, the so-called integrated Legendre basis [55]. Another possibility is to use the Lagrangian basis, but with Gauss-Lobatto nodes. Both these choices can be the subject of a future research.

4. Compute (and study) the symbol of the matrices arising from the Galkerkin IgA approximation and the IgA collocation approximation of (5.1) in the case where the B-spline basis functions are replaced by NURBS. The current research of our team is moving in this direction: after the identification and the study of the symbol, we will be interested in designing fast iterative solvers for the resulting discretization matrices, in analogy with the program followed in Chapters 6–7.

5. Use the properties of the symbol associated with the matrices coming from other numerical techniques than IgA in order to design fast iterative solvers also for these matrices. In Chapters 6–7 we only considered the IgA case, but one may be interested in fast solvers for other discretization matrices as well (e.g., FEM matrices or matrices associated with the Galerkin-type methods mentioned in item 2).

Besides the specific issues listed above, other more general problems that can be addressed in the future are, on the one hand, the computation/study of the symbol associated with the matrices A_n coming from the discretization of other differential problems of interest in Physics and Engineering (Navier-Stokes equations, elasticity equations, ...), and, on the other hand, the organization of the material concerning the 'GLT machinery' in a book.

Bibliography

[1] Aricò A., Donatelli M. *A V-cycle multigrid for multilevel matrix algebras: proof of optimality.* Numer. Math. **105** (2007) 511–547.

[2] Aricò A., Donatelli M., Serra-Capizzano S. *V-cycle optimal convergence for certain (multilevel) structured linear systems.* SIAM J. Matrix Anal. Appl. **26** (2004) 186–214.

[3] Auricchio F., Beirão da Veiga L., Hughes T. J. R., Reali A., Sangalli G. *Isogeometric collocation methods.* Math. Models Methods Appl. Sci. **20** (2010) 2075–2107.

[4] Axelsson O. *Iterative solution methods.* Cambridge University Press (1996).

[5] Beckermann B., Kuijlaars A. B. J. *Superlinear convergence of Conjugate Gradients.* SIAM J. Numer. Anal. **39** (2001) 300–329.

[6] Beckermann B., Serra-Capizzano S. *On the asymptotic spectrum of Finite Element matrix sequences.* SIAM J. Numer. Anal. **45** (2007) 746–769.

[7] Bhatia R. *Matrix Analysis.* Springer-Verlag, New York (1997).

[8] Bini D., Capovani M., Menchi O. *Metodi numerici per l'algebra lineare.* Zanichelli, Bologna (1988).

[9] Böttcher A., Grudsky S. M. *Spectral properties of banded Toeplitz matrices.* SIAM (2005).

[10] Böttcher A., Grudsky S., Ramirez de Arellano E. *On the asymptotic behavior of the eigenvectors of large banded Toeplitz matrices.* Math. Nachr. **279** (2006) 121–129.

[11] Böttcher A., Silbermann B. *Introduction to large truncated Toeplitz matrices.* Springer-Verlag, New York (1999).

[12] Böttcher A., Widom H. *From Toeplitz eigenvalues through Green's kernels to higher-order Wirtinger-Sobolev inequalities.* Oper Theory Adv. Appl. **171** (2007) 73-87.

[13] Brezis H. *Functional analysis, Sobolev spaces and partial differential equations.* Springer-Verlag, New York (2011).

[14] Brezzi F., Fortin M. *Mixed and hybrid Finite Element Methods.* Springer-Verlag, New York (1991).

[15] Canuto C., Hussaini M. Y., Quarteroni A., Zang T. A. *Spectral methods: evolution to complex geometries and applications to fluid dynamics.* Springer-Verlag, Berlin Heidelberg (2007).

[16] Chan R. H., Ng M. K. *Conjugate gradient method for Toeplitz systems.* SIAM Review **38** (1996) 427–482.

[17] Chui C. K. *An introduction to wavelets.* Academic Press (1992).

[18] Ciarlet P. *The Finite Element Method for elliptic problems.* SIAM (2002).

[19] Cottrell J. A., Hughes T. J. R., Bazilevs Y. *Isogeometric Analysis: toward integration of CAD and FEA.* John Wiley & Sons (2009).

[20] Davis P. J. *Circulant matrices.* 2nd Edition, AMS Chelsea Publishing (1994).

[21] De Boor C. *A practical guide to splines.* Springer-Verlag, New York (2001).

[22] Di Benedetto F., Fiorentino G., Serra S. *C.G. preconditioning for Toeplitz matrices.* Comput. Math. Appl. **25** (1993) 33–45.

[23] Donatelli M. *An algebraic generalization of local Fourier analysis for grid transfer operators in multigrid based on Toeplitz matrices.* Numer. Linear Algebra Appl. **17** (2010) 179–197.

[24] Donatelli M., Garoni C., Manni C., Serra-Capizzano S., Speleers H. *Robust and optimal multi-iterative techniques for IgA Galerkin linear systems.* Comput. Meth. Appl. Mech. Engrg. **284** (2015) 230–264.

[25] Donatelli M., Garoni C., Manni C., Serra-Capizzano S., Speleers H. *Symbol-based multigrid methods for Galerkin B-spline Isogeometric Analysis.* SIAM J. Numer. Anal. (2014) submitted. An extended version is available as Tech. Rep. TW650 (2014), Dept. Computer Science, KU Leuven.

[26] Donatelli M., Garoni C., Manni C., Serra-Capizzano S., Speleers H. *Spectral analysis and spectral symbol of matrices in isogeometric collocation methods.* Math. Comput. (2014) submitted.

[27] Donatelli M., Garoni C., Manni C., Serra-Capizzano S., Speleers H. *Robust and optimal multi-iterative techniques for IgA collocation linear systems.* Comput. Meth. Appl. Mech. Engrg. (2014) submitted.

[28] Donatelli M., Garoni C., Mazza M., Serra-Capizzano S., Sesana D. *Spectral behavior of preconditioned non-Hermitian multilevel block Toeplitz matrices with matrix-valued symbol.* Appl. Math. Comput. **245** (2014) 158–173.

[29] Donatelli M., Garoni C., Mazza M., Serra-Capizzano S., Sesana D. *Preconditioned HSS method for large multilevel block Toeplitz linear systems via the notion of matrix-valued symbol.* Numer. Linear Algebra Appl. (2014) submitted.

[30] Garoni C. *Estimates for the minimum eigenvalue and the condition number of Hermitian (block) Toeplitz matrices.* Linear Algebra Appl. **439** (2013) 707–728.

[31] Garoni C. *Properties of the eigenvalues of Galerkin stiffness and mass matrices associated with classical second-order elliptic problems.* Manuscript (2013).

[32] Garoni C., Hughes T. J. R., Reali A., Serra-Capizzano S., Speleers H. *Smoothness versus polynomial degree: why IgA outperforms FEA in the spectral approximation.* In preparation.

[33] Garoni C., Manni C., Pelosi F., Serra-Capizzano S., Speleers H. *On the spectrum of stiffness matrices arising from Isogeometric Analysis.* Numer. Math. **127** (2014) 751–799. An extended version is available as Tech. Rep. TW632 (2013), Dept. Computer Science, KU Leuven.

[34] Garoni C., Serra-Capizzano S., Vassalos P. *Tools for determining the asymptotic spectral distribution of Hermitian matrix-sequences and applications.* Oper. Matrices (2014) submitted.

[35] Garoni C., Serra-Capizzano S., Sesana D. *Tools for determining the asymptotic spectral distribution of non-Hermitian perturbations of Hermitian matrix-sequences and applications.* Integr. Equat. Oper. Theory (2014) http://dx.doi.org/10.1007/s00020-014-2157-6

[36] Garoni C., Serra-Capizzano S., Sesana D. *Spectral analysis and spectral symbol of d-variate $\mathbb{Q}_p$ Lagrangian FEM stiffness matrices.* SIAM J. Matrix Anal. Appl. (2014) submitted.

[37] Golinskii L., Serra-Capizzano S. *The asymptotic properties of the spectrum of nonsymmetrically perturbed Jacobi matrix sequences.* J. Approx. Theory **144** (2007) 84–102.

[38] Graham A. *Kronecker products and matrix calculus: with applications.* Ellis Horwood Limited, Chichester (1981).

[39] Grenander U., Szegö G. *Toeplitz forms and their applications.* 2nd Edition, Chelsea, New York (1984).

[40] Horn R. A., Johnson C. R. *Topics in Matrix Analysis.* Cambridge University Press (1994).

[41] Hughes T. J. R., Cottrell J. A., Bazilevs Y. *Isogeometric Analysis: CAD, Finite Elements, NURBS, exact geometry and mesh refinement.* Comput. Meth. Appl. Mech. Engrg. **194** (2005) 4135–4195.

[42] Hughes T. J. R., Evans J. A., Reali A. *Finite Element and NURBS approximations of eigenvalue, boundary-value, and initial-value problems.* Comput. Meth. Appl. Mech. Engrg. **272** (2014) 290–320.

[43] Jin X. Q. *Developments and applications of block Toeplitz iterative solvers.* Kluwer Academic Publishers and Science Press (2002).

[44] Olsen E., Douglas J. *Bounds on spectral condition numbers of matrices arising in the p-version of the Finite Element Method.* Numer. Math. **69** (1995) 333–352.

[45] Parter S. V. *On the extreme eigenvalues of truncated Toeplitz matrices.* Bull. American Math. Soc. **67** (1961) 191–197.

[46] Parter S. V. *On the extreme eigenvalues of Toeplitz matrices.* Trans. American Math. Soc. **100** (1961) 263-276.

[47] Quarteroni A. *Modellistica numerica per problemi differenziali.* 4a Edizione, Spinger-Verlag Italia, Milano (2008).

[48] Quarteroni A. *Numerical models for differential problems.* Springer-Verlag Italia, Milan (2009).

[49] Quarteroni A., Valli A. *Numerical approximation of partial differential equations.* Springer-Verlag, Berlin Heidelberg (2008).

[50] Rudin W. *Real and complex analysis.* 3rd Edition, McGraw-Hill (1987).

[51] Ruge J. W., Stüben K. *Algebraic multigrid.* Chapter 4 of the book 'Frontiers in applied mathematics: multigrid methods', by S. F. McCormick, SIAM (1987).

[52] Saad Y. *Iterative methods for sparse linear systems.* SIAM (2003).

[53] Schillinger D., Evans J. A., Reali A., Scott M. A., Hughes T. J. R. *Isogeometric collocation: cost comparison with Galerkin methods and extension to adaptive hierarchical NURBS discretizations.* Comput. Meth. Appl. Mech. Engrg. **267** (2013) 170–232.

[54] Schumaker L. L. *Spline functions: basic theory.* 3rd Edition, Cambridge Mathematical Library (2007).

[55] Schwab C. *p- and hp- Finite Element Methods.* Clarendon Press, Oxford (1998).

[56] Serra S. *Multi-iterative methods.* Comput. Math. Appl. **26** (1993) 65–87.

[57] Serra S. *On the extreme spectral properties of symmetric Toeplitz matrices generated by L^1 functions with several global minima/maxima.* BIT **36** (1996) 135–142.

[58] Serra-Capizzano S. *An ergodic theorem for classes of preconditioned matrices.* Linear Algebra Appl. **282** (1998) 161–183.

[59] Serra-Capizzano S. *Distribution results on the algebra generated by Toeplitz sequences: a finite dimensional approach.* Linear Algebra Appl. **328** (2001) 121–130.

[60] Serra-Capizzano S. *Spectral behavior of matrix sequences and discretized boundary value problems.* Linear Algebra Appl. **337** (2001) 37–78.

[61] Serra-Capizzano S. *More inequalities and asymptotics for matrix valued linear positive operators: the noncommutative case.* Oper. Theory Adv. Appl. **135** (2002) 293–315.

[62] Serra-Capizzano S. *Convergence analysis of two-grid methods for elliptic Toeplitz and PDEs matrix-sequences.* Numer. Math. **92** (2002) 433–465.

[63] Serra-Capizzano S. *Generalized Locally Toeplitz sequences: spectral analysis and applications to discretized Partial Differential Equations.* Linear Algebra Appl. **366** (2003) 371–402.

[64] Serra-Capizzano S. *GLT sequences as a Generalized Fourier Analysis and applications.* Linear Algebra Appl. **419** (2006) 180–233.

[65] Serra-Capizzano S., Tablino-Possio C. *Multigrid methods for multilevel circulant matrices.* SIAM J. Sci. Comput. **26** (2004) 55–85.

[66] Smith G. D. *Numerical solution of partial differential equations: Finite Difference methods.* 3rd Edition, Oxford University Press (1985).

[67] Tilli P. *A note on the spectral distribution of Toeplitz matrices.* Linear Multilinear Algebra **45** (1998) 147–159.

[68] Tilli P. *Locally Toeplitz sequences: spectral properties and applications.* Linear Algebra Appl. **278** (1998) 91–120.

[69] Trottenberg U., Oosterlee C. W., Schüller A. *Multigrid.* Academic Press, London (2001).

[70] Tyrtyshnikov E. E. *A unifying approach to some old and new theorems on distribution and clustering.* Linear Algebra Appl. **232** (1996) 1–43.

[71] Varga R. S. *Matrix iterative analysis.* Prentice Hall, Englewood Cliffs (1962).

[72] Zamarashkin N. L., Tyrtyshnikov E. E. *On the distribution of eigenvectors of Toeplitz matrices with weakened requirements on the generating function.* Russian Math. Survey **522** (1997) 1333–1334.

Acknowledgments

I wish to thank my wonderful supervisor Stefano Serra-Capizzano for his invaluable technical guidance and his incessant moral support (especially at the beginning of my Ph.D., when I was always weeping and claiming that I had no ideas to publish...): he was like a father for me.

Moreover, I wish to thank Carla Manni, who was like a mother for me (indeed, she 'looked after' me in more than one situation...).

Many thanks to Hendrik Speleers, Mariarosa Mazza, Debora Sesana, Marco Donatelli, Francesca Pelosi, Alessandro Reali, Thomas J. R. Hughes, Bruno Iannazzo, Cristina Tablino-Possio, Micol Pennacchio, Annalisa Buffa and Giancarlo Sangalli for their friendship and collaboration.

Finally, I wish to thank also Albrecht Böttcher and Harold Widom for helping me in publishing my very first paper [30] (and it is known the importance of the first paper for a Ph.D. student...).