Generalized Locally Toeplitz Sequences: A Spectral Analysis Tool for Discretized Differential Equations

Carlo Garoni and Stefano Serra-Capizzano

Contents

1 Introduction

Origin and purpose of the theory of GLT sequences The theory of Generalized Locally Toeplitz (GLT) sequences stems from Tilli's work on Locally Toeplitz (LT) sequences [56] and from the spectral theory of Toeplitz matrices [2, 10, 11, 12, 13, 36, 44, 55, 57, 58, 59, 60]. It was then developed by the authors in [33, 34, 50, 51] and has been recently extended by Barbarino in [3]. It was devised in order to solve a specific application problem, namely the problem of computing/analyzing the spectral distribution of matrices arising from the numerical discretization of Differential Equations (DEs). A final goal of this spectral analysis is the design of efficient numerical methods for computing the related numerical solutions. The theory of GLT sequences finds applications also in other areas of science (see, e.g., [16] and [33, Sections 10.1–10.4]), but the computation of the spectral distribution of DE discretization matrices remains the main application. The next paragraph is therefore devoted to a general description of this application.

Main application of the theory of GLT sequences Suppose a linear DE

$$\mathscr{A}u = g$$

is discretized by a linear numerical method characterized by a mesh fineness parameter n. In this situation, the computation of the numerical solution reduces to solving a linear system of the form

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

where the size d_n of the matrix A_n increases with n. What is often observed in practice is that A_n enjoys an asymptotic spectral distribution as $n\to\infty$, i.e., as the mesh is progressively refined. More precisely, it often turns out that, for a large class of test functions F,

$$\lim_{n\to\infty}\frac{1}{d_n}\sum_{j=1}^{d_n}F(\lambda_j(A_n)) = \frac{1}{\mu_k(D)}\int_D F(\kappa(\mathbf{y}))\mathrm{d}\mathbf{y},$$

where $\lambda_j(A_n),\ j=1,\ldots,d_n$, are the eigenvalues of A_n, μ_k is the Lebesgue measure in $\mathbb{R}^k$, and $\kappa : D\subset\mathbb{R}^k\to\mathbb{C}$. In this situation, the function κ is referred to as the *spectral symbol* of the sequence $\{A_n\}_n$. The spectral information contained in κ can be informally summarized as follows: assuming that n is large enough, the eigenvalues of A_n, except possibly for $o(d_n)$ outliers, are approximately equal to the samples of κ over a uniform grid in D. For example, if $k=1$, $d_n=n$ and $D=[a,b]$, then, assuming we have no outliers, the eigenvalues of A_n are approximately equal to

$$\kappa\Big(a+i\,\frac{b-a}{n}\Big),\qquad i=1,\ldots,n,$$

for n large enough. Similarly, if $k=2$, $d_n=n^2$ and $D=[a_1,b_1]\times[a_2,b_2]$, then, assuming we have no outliers, the eigenvalues of A_n are approximately equal to

$$\kappa\Big(a_1+i_1\,\frac{b_1-a_1}{n},\ a_2+i_2\,\frac{b_2-a_2}{n}\Big),\qquad i_1,i_2=1,\ldots,n,$$

for n large enough. It is then clear that the symbol κ provides a 'compact' and quite accurate description of the spectrum of the matrices A_n (for n large enough).

The theory of GLT sequences is a powerful apparatus for computing the spectral symbol κ. Indeed, the sequence of discretization matrices $\{A_n\}_n$ turns out to be a GLT sequence with symbol (or kernel) κ for many classes of DEs and numerical methods, especially if the numerical method belongs to the family of the so-called 'local methods'. Local methods are, for example, Finite Difference (FD) methods, Finite Element (FE) methods with 'locally supported' basis functions, and collocation methods; in short, all standard numerical methods for the approximation of DEs. We refer the reader to Section 3.2 and [9, 33, 34, 50, 51, 52] for applications of the theory of GLT sequences in the context of FD discretizations of DEs; to Section 3.3 and [5, 9, 25, 26, 33, 34, 51] for the FE and collocation settings; to Section 3.4 and [22, 27, 33, 34, 28, 30, 31, 32, 48] for the case of Isogeometric Analysis (IgA) discretizations, both in the collocation and Galerkin frameworks; and to [24] for a further recent application to fractional DEs.

Practical use of the spectral symbol It is worth emphasizing that the knowledge of the spectral symbol κ, which can be attained through the theory of GLT sequences, is not only interesting in itself, but may also be exploited for practical purposes. Let us mention some of them.

(a) Compare the spectrum of A_n, compactly described by κ, with the spectrum of the differential operator $\mathscr{A}$.

(b) Understand whether the numerical method used to discretize the DE $\mathscr{A}u = g$ is appropriate or not to spectrally approximate the operator $\mathscr{A}$.

(c) Analyze the convergence and predict the behavior of iterative methods (especially, multigrid and preconditioned Krylov methods), when they are applied to A_n.

(d) Design fast iterative solvers (especially, multigrid and preconditioned Krylov methods) for linear systems with coefficient matrix A_n.

The goal (b) can be achieved through the spectral comparison mentioned in (a) and allows one to classify the various numerical methods on the basis of their spectral approximation properties. In this way, it is possible to select the best approximation technique among a set of given methods. In this regard, we point out that the symbol-based analysis carried out in [28] proved that IgA is superior to classical FE methods in the spectral approximation of the underlying differential operator $\mathscr{A}$. The reason for which the spectral symbol κ can be exploited for the purposes (c)–(d) is the following: the convergence properties of iterative solvers in general (and of multigrid and preconditioned Krylov methods in particular) strongly depend on the spectral features of the matrix to which they are applied; hence, the spectral information provided by κ can be conveniently used for designing fast solvers of this kind and/or analyzing their convergence properties. In this respect, we recall that noteworthy estimates on 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 [4]. We also refer the reader to [20, 21, 23] for recent developments in the IgA framework, where the spectral symbol was exploited to design ad hoc iterative solvers for IgA discretization matrices.

Description of the present work The present work is an excerpt of the book [33]. Its purpose is to introduce the reader to the theory of GLT sequences and its applications in the context of DE discretizations. Following [33], we will here consider only unidimensional DEs both for simplicity and because the key 'GLT ideas' are better conveyed in the univariate setting. For the multivariate setting, the reader is referred to the literature cited above and, especially, to the book [34].

2 The Theory of GLT Sequences: A Summary

In this chapter we present a *self-contained* summary of the theory of GLT sequences. Despite its conciseness, our presentation contains *everything one needs to know* in order to understand the applications presented in the next chapter.

Matrix-sequences Throughout this work, by a *matrix-sequence* we mean a sequence of the form $\{A_n\}_n$, where A_n is an $n\times n$ matrix. We say that the matrix-sequence $\{A_n\}_n$ is *Hermitian* if each A_n is Hermitian.

Singular value and eigenvalue distribution of a matrix-sequence Let μ_k be the Lebesgue measure in $\mathbb{R}^k$. Throughout this work, all the terminology coming from measure theory (such as 'measurable set', 'measurable function', 'almost everywhere (a.e.)', etc.) is always referred to the Lebesgue measure. Let $C_c(\mathbb{R})$ (resp., $C_c(\mathbb{C})$) be the space of continuous complex-valued functions with bounded support defined on $\mathbb{R}$ (resp., $\mathbb{C}$). If A is a square matrix of size n, the singular values and the eigenvalues of A are denoted by $\sigma_1(A),\ldots,\sigma_n(A)$ and $\lambda_1(A),\ldots,\lambda_n(A)$, respectively. The set of the eigenvalues (i.e., the spectrum) of A is denoted by $\Lambda(A)$.

Definition 2.1. Let $\{A_n\}_n$ be a matrix-sequence and let $f: D\subset\mathbb{R}^k\to\mathbb{C}$ be a measurable function defined on a set D with $0<\mu_k(D)<\infty$.

- We say that $\{A_n\}_n$ has a singular value distribution described by f, and we write $\{A_n\}_n\sim_\sigma f$, if

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}F(\sigma_i(A_n))=\frac{1}{\mu_k(D)}\int_D F(|f(\mathbf{x})|)\mathrm{d}\mathbf{x},\qquad \forall F\in C_c(\mathbb{R}).$$

  In this case, f is called the *singular value symbol* of $\{A_n\}_n$.

- We say that $\{A_n\}_n$ has a spectral (or eigenvalue) distribution described by f, and we write $\{A_n\}_n\sim_\lambda f$, if

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}F(\lambda_i(A_n))=\frac{1}{\mu_k(D)}\int_D F(f(\mathbf{x}))\mathrm{d}\mathbf{x},\qquad \forall F\in C_c(\mathbb{C}).$$

  In this case, f is called the *spectral (or eigenvalue) symbol* of $\{A_n\}_n$.

When we write a relation such as $\{A_n\}_n\sim_\sigma f$ or $\{A_n\}_n\sim_\lambda f$, it is understood that $\{A_n\}_n$ is a matrix-sequence and f is a measurable function defined on a subset D of some $\mathbb{R}^k$ with $0<\mu_k(D)<\infty$. If $\{A_n\}_n$ has both a singular value and a spectral distribution described by f, we write $\{A_n\}_n\sim_{\sigma,\lambda} f$.

We report in **S 1** and **S 2** the statements of two useful results concerning the spectral distribution of matrix-sequences. Throughout this work, if A is an $n\times n$ matrix and $1\le p\le\infty$, we denote by $\|A\|_p$ the Schatten p-norm of A, i.e., the p-norm of the vector $(\sigma_1(A),\ldots,\sigma_n(A))$ formed by the singular values of A; see [7]. The Schatten ∞-norm $\|A\|_\infty$ is the largest singular value of A and coincides with the classical 2-norm $\|A\|$. The Schatten 1-norm $\|A\|_1$ is the sum of all the singular values of A and is often referred to as the trace-norm of A. The (topological) closure of a set S is denoted by $\overline{S}$.

S 1. If $\{A_n\}_n\sim_\lambda f$ and $\Lambda(A_n)\subseteq S$ for all n then $f\in\overline{S}$ a.e.

S 2. If $A_n=X_n+Y_n$ where

- each X_n is Hermitian and $\{X_n\}_n\sim_\lambda f$,
- $\|X_n\|,\|Y_n\|\le C$ for all n, where C is a constant independent of n,
- $n^{-1}\|Y_n\|_1\to 0$,

then $\{A_n\}_n\sim_\lambda f$.

Informal meaning Assuming f is Riemann-integrable, the spectral distribution $\{A_n\}_n\sim_\lambda f$ has the following informal meaning: all the eigenvalues of A_n, except possibly for $o(n)$ outliers, are approximately equal to the samples of f over a uniform grid in D (for n large enough). For instance, if $k=1$ and $D=[a,b]$, then, assuming we have no outliers, the eigenvalues of A_n are approximately equal to

$$f\Big(a+i\,\frac{b-a}{n}\Big),\qquad i=1,\ldots,n,$$

for n large enough. Similarly, if $k=2$, $n=m^2$ and $D=[a_1,b_1]\times[a_2,b_2]$, then, assuming we have no outliers, the eigenvalues of A_n are approximately equal to

$$f\Big(a_1+i\,\frac{b_1-a_1}{m},\ a_2+j\,\frac{b_2-a_2}{m}\Big),\qquad i,j=1,\ldots,m,$$

for n large enough. A completely analogous meaning can also be given for the singular value distribution $\{A_n\}_n\sim_\sigma f$.

Zero-distributed sequences A matrix-sequence $\{Z_n\}_n$ such that $\{Z_n\}_n\sim_\sigma 0$ is referred to as a zero-distributed sequence. In other words, $\{Z_n\}_n$ is zero-distributed if and only if

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}F(\sigma_i(Z_n))=F(0),\qquad \forall F\in C_c(\mathbb{R}).$$

Z 1 – Z 2 will provide us with an important characterization of zero-distributed sequences together with a useful sufficient condition for detecting such sequences. For convenience, throughout this work we use the natural convention $1/\infty=0$.

Z 1. $\{Z_n\}_n\sim_\sigma 0$ if and only if $Z_n=R_n+N_n$ with $\lim_{n\to\infty}n^{-1}\mathrm{rank}(R_n)=\lim_{n\to\infty}\|N_n\|=0$.

Z 2. $\{Z_n\}_n\sim_\sigma 0$ if there is a $p\in[1,\infty]$ such that $\lim_{n\to\infty}n^{-1/p}\|Z_n\|_p=0$.

Sequences of diagonal sampling matrices If $n\in\mathbb{N}$ and $a:[0,1]\to\mathbb{C}$, the nth diagonal sampling matrix generated by a is the $n\times n$ diagonal matrix given by

$$D_n(a)=\operatorname*{diag}_{i=1,\ldots,n}a\Big(\frac{i}{n}\Big).$$

$\{D_n(a)\}_n$ is called the sequence of diagonal sampling matrices generated by a.

Toeplitz sequences If $n\in\mathbb{N}$ and $f:[-\pi,\pi]\to\mathbb{C}$ is a function in $L^1([-\pi,\pi])$, the nth Toeplitz matrix generated by f is the $n\times n$ matrix

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

where the numbers f_k are the Fourier coefficients of f,

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

$\{T_n(f)\}_n$ is called the Toeplitz sequence generated by f.

T 1. For every $n\in\mathbb{N}$ the map $T_n(\cdot):L^1([-\pi,\pi])\to\mathbb{C}^{n\times n}$

- is linear: $T_n(\alpha f+\beta g)=\alpha T_n(f)+\beta T_n(g)$ for $\alpha,\beta\in\mathbb{C}$ and $f,g\in L^1([-\pi,\pi])$;
- satisfies $(T_n(f))^*=T_n(\overline{f})$ for all $f\in L^1([-\pi,\pi])$, so if f is real then $T_n(f)$ is Hermitian for all n.

T 2. If $f\in L^1([-\pi,\pi])$ then $\{T_n(f)\}_n\sim_\sigma f$. If $f\in L^1([-\pi,\pi])$ and f is real then $\{T_n(f)\}_n\sim_\lambda f$.

T 3. If $n\in\mathbb{N}$, $1\le p\le\infty$ and $f\in L^p([-\pi,\pi])$, then $\|T_n(f)\|_p\le\frac{n^{1/p}}{(2\pi)^{1/p}}\|f\|_{L^p}$.

Approximating classes of sequences The notion of approximating classes of sequences (a.c.s.) is the fundamental concept on which the theory of GLT sequences is based.

Definition 2.2. Let $\{A_n\}_n$ be a matrix-sequence and let $\{\{B_{n,m}\}_n\}_m$ be a sequence of matrix-sequences. We say that $\{\{B_{n,m}\}_n\}_m$ is an approximating class of sequences (a.c.s.) for $\{A_n\}_n$ if the following condition is met: for every m there exists n_m such that, for $n\ge n_m$,

$$A_n=B_{n,m}+R_{n,m}+N_{n,m},\qquad \mathrm{rank}(R_{n,m})\le c(m)n,\qquad \|N_{n,m}\|\le\omega(m),$$

where n_m, $c(m)$, $\omega(m)$ depend only on m, and

$$\lim_{m\to\infty}c(m)=\lim_{m\to\infty}\omega(m)=0.$$

Throughout this work, we use the abbreviation 'a.c.s.' for both the singular 'approximating class of sequences' and the plural 'approximating classes of sequences'; it will be clear from the context whether 'a.c.s.' is singular or plural. Roughly speaking, $\{\{B_{n,m}\}_n\}_m$ is an a.c.s. for $\{A_n\}_n$ if, for all sufficiently large m, the sequence $\{B_{n,m}\}_n$ approximates $\{A_n\}_n$ in the sense that A_n is eventually equal to $B_{n,m}$ plus a small-rank matrix (with respect to the matrix size n) plus a small-norm matrix. It turns out that the notion of a.c.s. is a notion of convergence in the space of matrix-sequences $\mathscr{E} = \{\{A_n\}_n : \{A_n\}_n \text{ is a matrix-sequence}\}$, i.e., there exists a topology $\tau_{\text{a.c.s.}}$ on $\mathscr{E}$ such that $\{\{B_{n,m}\}_n\}_m$ is an a.c.s. for $\{A_n\}_n$ if and only if $\{\{B_{n,m}\}_n\}_m$ converges to $\{A_n\}_n$ in $(\mathscr{E}, \tau_{\text{a.c.s.}})$. The theory of a.c.s. may then be interpreted as an approximation theory for matrix-sequences, and for this reason we will use the convergence notation $\{B_{n,m}\}_n \xrightarrow{\text{a.c.s.}} \{A_n\}_n$ to indicate that $\{\{B_{n,m}\}_n\}_m$ is an a.c.s. for $\{A_n\}_n$.

ACS 1. $\{A_n\}_n \sim_\sigma f$ if and only if there exist matrix-sequences $\{B_{n,m}\}_n \sim_\sigma f_m$ such that $\{B_{n,m}\}_n \xrightarrow{\text{a.c.s.}} \{A_n\}_n$ and $f_m \to f$ in measure.

ACS 2. Suppose the matrices A_n are Hermitian. Then, $\{A_n\}_n \sim_\lambda f$ if and only if there exist Hermitian matrix-sequences $\{B_{n,m}\}_n \sim_\lambda f_m$ such that $\{B_{n,m}\}_n \xrightarrow{\text{a.c.s.}} \{A_n\}_n$ and $f_m \to f$ in measure.

ACS 3. Let $p \in [1,\infty]$ and suppose for every m there exists n_m such that, for $n \ge n_m$, $\|A_n - B_{n,m}\|_p \le \varepsilon(m,n)n^{1/p}$, where $\lim_{m\to\infty}\limsup_{n\to\infty}\varepsilon(m,n) = 0$. Then $\{B_{n,m}\}_n \xrightarrow{\text{a.c.s.}} \{A_n\}_n$.

Generalized locally Toeplitz sequences A Generalized Locally Toeplitz (GLT) sequence $\{A_n\}_n$ is a special matrix-sequence equipped with a measurable function $\kappa : [0,1]\times[-\pi,\pi] \to \mathbb{C}$, the so-called *symbol* (or *kernel*). We use the notation $\{A_n\}_n \sim_{\text{GLT}} \kappa$ to indicate that $\{A_n\}_n$ is a GLT sequence with symbol κ. The symbol of a GLT sequence is unique in the sense that if $\{A_n\}_n \sim_{\text{GLT}} \kappa$ and $\{A_n\}_n \sim_{\text{GLT}} \xi$ then $\kappa = \xi$ a.e. in $[0,1]\times[-\pi,\pi]$. The main properties of GLT sequences are summarized in the following list. If A is a matrix, we denote by $A^\dagger$ the Moore–Penrose pseudoinverse of A; we recall that $A^\dagger = A^{-1}$ whenever A is invertible and we refer the reader to [8, 35] for more details on the pseudoinverse of a matrix. If A is a Hermitian matrix and f is a function defined at each point of $\Lambda(A)$, we denote by $f(A)$ the unique matrix such that $f(A)\mathbf{v} = f(\lambda)\mathbf{v}$ whenever $A\mathbf{v} = \lambda\mathbf{v}$; for more on matrix functions, we refer the reader to Higham's book [37].

GLT 1. If $\{A_n\}_n \sim_{\text{GLT}} \kappa$ then $\{A_n\}_n \sim_\sigma \kappa$. If $\{A_n\}_n \sim_{\text{GLT}} \kappa$ and the matrices A_n are Hermitian then $\{A_n\}_n \sim_\lambda \kappa$.

GLT 2. If $\{A_n\}_n \sim_{\text{GLT}} \kappa$ and $A_n = X_n + Y_n$, where

- every X_n is Hermitian,
- $\|X_n\|, \|Y_n\| \le C$ for some constant C independent of n,
- $n^{-1}\|Y_n\|_1 \to 0$,

then $\{A_n\}_n \sim_\lambda \kappa$.

GLT 3. We have

- $\{T_n(f)\}_n \sim_{\text{GLT}} \kappa(x,\theta) = f(\theta)$ if $f \in L^1([-\pi,\pi])$,
- $\{D_n(a)\}_n \sim_{\text{GLT}} \kappa(x,\theta) = a(x)$ if $a : [0,1] \to \mathbb{C}$ is Riemann-integrable,
- $\{Z_n\}_n \sim_{\text{GLT}} \kappa(x,\theta) = 0$ if and only if $\{Z_n\}_n \sim_\sigma 0$.

GLT 4. If $\{A_n\}_n \sim_{\text{GLT}} \kappa$ and $\{B_n\}_n \sim_{\text{GLT}} \xi$ then

- $\{A_n^*\}_n \sim_{\text{GLT}} \overline{\kappa}$,
- $\{\alpha A_n + \beta B_n\}_n \sim_{\text{GLT}} \alpha\kappa + \beta\xi$ for all $\alpha, \beta \in \mathbb{C}$,
- $\{A_nB_n\}_n \sim_{\text{GLT}} \kappa\xi$.

GLT 5. If $\{A_n\}_n \sim_{\text{GLT}} \kappa$ and $\kappa \ne 0$ a.e. then $\{A_n^\dagger\}_n \sim_{\text{GLT}} \kappa^{-1}$.

GLT 6. If $\{A_n\}_n \sim_{\text{GLT}} \kappa$ and each A_n is Hermitian, then $\{f(A_n)\}_n \sim_{\text{GLT}} f(\kappa)$ for every continuous function $f : \mathbb{C} \to \mathbb{C}$.

GLT 7. $\{A_n\}_n \sim_{\text{GLT}} \kappa$ if and only if there exist GLT sequences $\{B_{n,m}\}_n \sim_{\text{GLT}} \kappa_m$ such that $\{B_{n,m}\}_n \xrightarrow{\text{a.c.s.}} \{A_n\}_n$ and $\kappa_m \to \kappa$ in measure.

3 Applications

In this chapter we present several applications of the theory of GLT sequences to the spectral analysis of DE discretization matrices. Our aim is to show how to compute the singular value and eigenvalue distribution of matrix-sequences arising from a DE discretization through the ‘GLT tools’ presented in the previous chapter. We begin by considering FD discretizations, then we will move to FE discretizations, and finally we will focus on IgA discretizations. Before starting, we collect in the next section some auxiliary results.

3.1 Preliminaries

3.1.1 Matrix-Norm Inequalities

If $1 \le p \le \infty$, the symbol $|\cdot|_p$ denotes both the p-norm of vectors and the associated operator norm for matrices:

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

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

The 2-norm $|\cdot|_2$ is also known as the spectral (or Euclidean) norm and it is preferably denoted by $\|\cdot\|$. Important inequalities involving the p-norms with $p = 1, 2, \infty$ are the following:

$$\|X\| \le \sqrt{|X|_1 |X|_\infty}, \qquad X \in \mathbb{C}^{m\times m}, \tag{3.1}$$

$$\|X\| \ge |x_{ij}|, \qquad i, j = 1, \ldots, m, \qquad X \in \mathbb{C}^{m\times m}; \tag{3.2}$$

see [8, 35]. Recalling that $|X|_1 = \max_{j=1,\ldots,m} \sum_{i=1}^m |x_{ij}|$ and $|X|_\infty = \max_{i=1,\ldots,m} \sum_{j=1}^m |x_{ij}|$, the inequalities (3.1)–(3.2) are particularly useful to estimate the spectral norm of a matrix when we have bounds for its components.

As mentioned in Chapter 2, the Schatten p-norm of an $n \times n$ matrix A is defined as the p-norm of the vector $(\sigma_1(A), \ldots, \sigma_n(A))$ formed by the singular values of A. The Schatten ∞-norm $\|A\|_\infty$ is the largest singular value $\sigma_{\max}(A)$ and coincides with the spectral norm $\|A\|$. The Schatten 1-norm $\|A\|_1$ is the sum of all the singular values of A and is often referred to as the trace-norm of A. The Schatten p-norms are deeply studied in Bhatia’s book [7]. Here, we just recall a couple of basic trace-norm inequalities that we shall need in what follows:

$$\|X\|_1 \le \mathrm{rank}(X)\|X\| \le m\|X\|, \qquad X \in \mathbb{C}^{m\times m}, \tag{3.3}$$

$$\|X\|_1 \le \sum_{i,j=1}^m |x_{ij}|, \qquad X \in \mathbb{C}^{m\times m}. \tag{3.4}$$

The inequality (3.3) follows from the equation $\sigma_{\max}(X) = \|X\|$ and the definition $\|X\|_1 = \sum_{i=1}^m \sigma_i(X) = \sum_{i=1}^{\mathrm{rank}(X)} \sigma_i(X)$. For the proof of (3.4), see, e.g., [33, Section 2.4.3].

3.1.2 GLT Preconditioning

The next theorem is an important result in the context of GLT preconditioning, but it will be used only in Section 3.4.3. The reader may then decide to skip it on first reading and come back here afterwards, just before going into Section 3.4.3.

Theorem 3.1. *Let $\{A_n\}_n$ be a sequence of Hermitian matrices such that $\{A_n\}_n \sim_{\mathrm{GLT}} \kappa$, and let $\{P_n\}_n$ be a sequence of Hermitian Positive Definite (HPD) matrices such that $\{P_n\}_n \sim_{\mathrm{GLT}} \xi$ with $\xi \ne 0$ a.e. Then, the sequence of preconditioned matrices $P_n^{-1}A_n$ satisfies*

$$\{P_n^{-1}A_n\}_n \sim_{\mathrm{GLT}} \xi^{-1}\kappa,$$

and

$$\{P_n^{-1}A_n\}_n \sim_{\sigma,\lambda} \xi^{-1}\kappa.$$

Proof. The GLT relation $\{P_n^{-1}A_n\}_n \sim_{\mathrm{GLT}} \xi^{-1}\kappa$ is a direct consequence of **GLT 4 – GLT 5**. The singular value distribution $\{P_n^{-1}A_n\}_n \sim_\sigma \xi^{-1}\kappa$ follows immediately from **GLT 1**. The only difficult part is the spectral distribution $\{P_n^{-1}A_n\}_n \sim_\lambda \xi^{-1}\kappa$, which does not follow from **GLT 1** because $P_n^{-1}A_n$ is not Hermitian in general.

Since P_n is HPD, the eigenvalues of P_n are positive and the matrices $P_n^{1/2}$ and $P_n^{-1/2}$ are well-defined. Moreover,

$$P_n^{-1}A_n \sim P_n^{-1/2}A_nP_n^{-1/2}, \tag{3.5}$$

where $X \sim Y$ means that X is similar to Y. The good news is that $P_n^{-1/2}A_nP_n^{-1/2}$ is Hermitian and, moreover, by **GLT 4** – **GLT 6** (with **GLT 6** applied to $f(z) = |z|^{1/2}$), we have

$$\{P_n^{-1/2}A_nP_n^{-1/2}\}_n \sim_{\rm GLT} |\xi|^{-1/2}\kappa|\xi|^{-1/2} = |\xi|^{-1}\kappa = \xi^{-1}\kappa;$$

note that the latter equation follows from the fact that $\xi \geq 0$ a.e. by **S 1**, since P_n is HPD and $\{P_n\}_n \sim_\lambda \xi$ by **GLT 1**. Since $P_n^{-1/2}A_nP_n^{-1/2}$ is Hermitian, **GLT 1** yields

$$\{P_n^{-1/2}A_nP_n^{-1/2}\}_n \sim_\lambda \xi^{-1}\kappa.$$

Thus, by the similarity (3.5),

$$\{P_n^{-1}A_n\}_n \sim_\lambda \xi^{-1}\kappa. \qquad \square$$

3.1.3 Arrow-Shaped Sampling Matrices

If $n \in \mathbb{N}$ and $a : [0,1] \to \mathbb{C}$, the nth arrow-shaped sampling matrix generated by a is denoted by $S_n(a)$ and is defined as the following symmetric matrix of size n:

$$(S_n(a))_{i,j} = (D_n(a))_{\min(i,j),\min(i,j)} = a\Big(\frac{\min(i,j)}{n}\Big), \qquad i,j = 1,\ldots,n, \tag{3.6}$$

that is,

$$S_n(a) = \begin{bmatrix} a(\frac{1}{n}) & a(\frac{1}{n}) & a(\frac{1}{n}) & \cdots & \cdots & a(\frac{1}{n}) \\ a(\frac{1}{n}) & a(\frac{2}{n}) & a(\frac{2}{n}) & \cdots & \cdots & a(\frac{2}{n}) \\ a(\frac{1}{n}) & a(\frac{2}{n}) & a(\frac{3}{n}) & \cdots & \cdots & a(\frac{3}{n}) \\ \vdots & \vdots & \vdots & \ddots & & \vdots \\ \vdots & \vdots & \vdots & & \ddots & \vdots \\ a(\frac{1}{n}) & a(\frac{2}{n}) & a(\frac{3}{n}) & \cdots & \cdots & a(1) \end{bmatrix}.$$

The name is due to the fact that, if we imagine to color the matrix $S_n(a)$ by assigning the color i to the entries $a(\frac{i}{n})$, the resulting picture looks like a sort of arrow pointing toward the upper left corner. Throughout this work, if $X, Y \in \mathbb{C}^{m\times m}$, we denote by $X \circ Y$ the componentwise (Hadamard) product of X and Y:

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

Moreover, if $g : D \to \mathbb{C}$ is continuous over D, with $D \subseteq \mathbb{C}^k$ for some k, we denote by $\omega_g(\cdot)$ the modulus of continuity of g,

$$\omega_g(\delta) = \sup_{\substack{\mathbf{x},\mathbf{y}\in D \\ \|\mathbf{x}-\mathbf{y}\|\leq\delta}} |g(\mathbf{x}) - g(\mathbf{y})|, \qquad \delta > 0.$$

If we need/want to specify D, we will say that $\omega_g(\cdot)$ is the modulus of continuity of g over D.

Theorem 3.2. *Let $a : [0,1] \to \mathbb{C}$ be continuous and let f be a trigonometric polynomial of degree $\leq r$. Then, we have*

$$\|S_n(a) \circ T_n(f) - D_n(a)T_n(f)\| \leq (2r+1)\|f\|_\infty\, \omega_a\Big(\frac{r}{n}\Big) \tag{3.7}$$

for every $n \in \mathbb{N}$,

$$\|S_n(a) \circ T_n(f)\| \leq C \tag{3.8}$$

for every $n \in \mathbb{N}$ and for some constant C independent of n, and

$$\{S_n(a) \circ T_n(f)\}_n \sim_{\rm GLT} a(x)f(\theta). \tag{3.9}$$

Proof. For all $i, j = 1, \ldots, n$,

- if $|i-j| > r$, then the Fourier coefficient f_{i-j} is zero and, consequently,

$$(S_n(a) \circ T_n(f))_{ij} = (S_n(a))_{ij}(T_n(f))_{ij} = a\Big(\frac{\min(i,j)}{n}\Big) f_{i-j} = 0,$$
$$(D_n(a)T_n(f))_{ij} = (D_n(a))_{ii}(T_n(f))_{ij} = a\Big(\frac{i}{n}\Big) f_{i-j} = 0;$$

- if $|i-j| \le r$, then, using (3.2) and **T 3**, we obtain

$$\begin{aligned}|(S_n(a) \circ T_n(f))_{ij} - (D_n(a)T_n(f))_{ij}| &= |(S_n(a))_{ij}(T_n(f))_{ij} - (D_n(a))_{ii}(T_n(f))_{ij}| \\ &= |(S_n(a))_{ij} - (D_n(a))_{ii}|\,|(T_n(f))_{ij}| \\ &\le \Big|a\Big(\frac{\min(i,j)}{n}\Big) - a\Big(\frac{i}{n}\Big)\Big|\,\|T_n(f)\| \\ &\le \|f\|_\infty\, \omega_a\Big(\Big|\frac{\min(i,j)}{n} - \frac{i}{n}\Big|\Big).\end{aligned}$$

  Since $|i-j| \le r$, we have
$$\Big|\frac{\min(i,j)}{n} - \frac{i}{n}\Big| \le \frac{|j-i|}{n} \le \frac{r}{n},$$
  hence
$$\big|(S_n(a) \circ T_n(f))_{ij} - (D_n(a)T_n(f))_{ij}\big| \le \|f\|_\infty\, \omega_a\Big(\frac{r}{n}\Big).$$

It follows from the first item that the nonzero entries in each row and column of $S_n(a) \circ T_n(f) - D_n(a)T_n(f)$ are at most $2r+1$. Hence, from the second item we infer that the 1-norm and the ∞-norm of $S_n(a) \circ T_n(f) - D_n(a)T_n(f)$ are bounded by $(2r+1)\|f\|_\infty\, \omega_a(\frac{r}{n})$. The application of (3.1) yields (3.7). Using (3.7) we obtain

$$\|S_n(a) \circ T_n(f)\| \le \|S_n(a) \circ T_n(f) - D_n(a)T_n(f)\| + \|D_n(a)\|\,\|T_n(f)\| \le (2r+1)\|f\|_\infty\, \omega_a\Big(\frac{r}{n}\Big) + \|a\|_\infty \|f\|_\infty,$$

which implies (3.8). Finally, since $\omega_a(\frac{r}{n}) \to 0$ as $n \to \infty$, the matrix-sequence $\{S_n(a) \circ T_n(f) - D_n(a)T_n(f)\}_n$ is zero-distributed by (3.7) and **Z 1** (or **Z 2**). Thus, (3.9) follows from **GLT 3 – GLT 4**. ☐

3.2 FD Discretization of Differential Equations

3.2.1 FD Discretization of Diffusion Equations

Consider the following second-order differential problem:

$$\begin{cases} -(a(x)u'(x))' = f(x), & x \in (0,1), \\ u(0) = \alpha, \quad u(1) = \beta, \end{cases} \tag{3.10}$$

where $a \in C([0,1])$ and f is a given function. To ensure the well-posedness of this problem, further conditions on a and f should be imposed; for example, $f \in L^2([0,1])$ and $a \in C^1([0,1])$ with $a(x) > 0$ for every $x \in [0,1]$, so that problem (3.10) is elliptic (see Chapter 8 of [14], especially the Sturm-Liouville problem on page 223). However, we here only assume that $a \in C([0,1])$ as the GLT analysis presented herein does not require any other assumption.

FD discretization We consider the discretization of (3.10) by the classical second-order central FD scheme on a uniform grid. In the case where $a(x)$ is constant, this is also known as the $(-1,2,-1)$ scheme. Let us describe it shortly; for more details on FD methods, we refer the reader to the available literature (see, e.g., [53] or any good book on FDs). Choose a discretization parameter $n \in \mathbb{N}$, set $h = \frac{1}{n+1}$ and $x_j = jh$ for all $j \in [0, n+1]$. For $j = 1, \ldots, n$ we approximate

$-(a(x)u'(x))'|_{x=x_j}$ by the classical second-order central FD formula:

$$\begin{aligned}
-(a(x)u'(x))'|_{x=x_j} &\approx -\frac{a(x_{j+\frac{1}{2}})u'(x_{j+\frac{1}{2}}) - a(x_{j-\frac{1}{2}})u'(x_{j-\frac{1}{2}})}{h} \\
&\approx -\frac{a(x_{j+\frac{1}{2}})\dfrac{u(x_{j+1})-u(x_j)}{h} - a(x_{j-\frac{1}{2}})\dfrac{u(x_j)-u(x_{j-1})}{h}}{h} \\
&= \frac{-a(x_{j+\frac{1}{2}})u(x_{j+1}) + \big(a(x_{j+\frac{1}{2}})+a(x_{j-\frac{1}{2}})\big)u(x_j) - a(x_{j-\frac{1}{2}})u(x_{j-1})}{h^2}.
\end{aligned} \tag{3.11}$$

This means that the nodal values of the solution u satisfy (approximately) the following linear system:

$$-a(x_{j+\frac{1}{2}})u(x_{j+1}) + \big(a(x_{j+\frac{1}{2}})+a(x_{j-\frac{1}{2}})\big)u(x_j) - a(x_{j-\frac{1}{2}})u(x_{j-1}) = h^2 f(x_j), \qquad j=1,\ldots,n.$$

We then approximate the solution by the piecewise linear function that takes the value u_j in x_j for $j=0,\ldots,n+1$, where $u_0=\alpha$, $u_{n+1}=\beta$, and $\mathbf{u}=(u_1,\ldots,u_n)^T$ solves

$$-a(x_{j+\frac{1}{2}})u_{j+1} + \big(a(x_{j+\frac{1}{2}})+a(x_{j-\frac{1}{2}})\big)u_j - a(x_{j-\frac{1}{2}})u_{j-1} = h^2 f(x_j), \qquad j=1,\ldots,n. \tag{3.12}$$

The matrix of the linear system (3.12) is the $n\times n$ tridiagonal symmetric matrix given by

$$A_n = \begin{bmatrix}
a_{\frac{1}{2}}+a_{\frac{3}{2}} & -a_{\frac{3}{2}} & & & \\
-a_{\frac{3}{2}} & a_{\frac{3}{2}}+a_{\frac{5}{2}} & -a_{\frac{5}{2}} & & \\
 & -a_{\frac{5}{2}} & \ddots & \ddots & \\
 & & \ddots & \ddots & -a_{n-\frac{1}{2}} \\
 & & & -a_{n-\frac{1}{2}} & a_{n-\frac{1}{2}}+a_{n+\frac{1}{2}}
\end{bmatrix}, \tag{3.13}$$

where $a_i = a(x_i)$ for all $i\in[0,n+1]$.

GLT analysis of the FD discretization matrices We are going to see that the theory of GLT sequences allows one to compute the singular value and spectral distribution of the sequence of FD discretization matrices $\{A_n\}_n$. Actually, this is the fundamental example that led to the birth of the theory of LT sequences and, subsequently, of GLT sequences.

Theorem 3.3. *If $a\in C([0,1])$ then*

$$\{A_n\}_n \sim_{\rm GLT} a(x)(2-2\cos\theta) \tag{3.14}$$

and

$$\{A_n\}_n \sim_{\sigma,\lambda} a(x)(2-2\cos\theta). \tag{3.15}$$

Proof. It suffices to prove (3.14) because (3.15) follows from (3.14) and **GLT 1** as the matrices A_n are symmetric. Consider the matrix

$$D_n(a)T_n(2-2\cos\theta) = \begin{bmatrix}
2a(\frac{1}{n}) & -a(\frac{1}{n}) & & & \\
-a(\frac{2}{n}) & 2a(\frac{2}{n}) & -a(\frac{2}{n}) & & \\
 & -a(\frac{3}{n}) & \ddots & \ddots & \\
 & & \ddots & \ddots & -a(\frac{n-1}{n}) \\
 & & & -a(1) & 2a(1)
\end{bmatrix}. \tag{3.16}$$

In view of the inequalities $\left|x_j - \frac{j}{n}\right| \le \frac{1}{n+1} = h,\ j=1,\ldots,n$, a direct comparison between (3.16) and (3.13) shows that the modulus of each diagonal entry of the matrix $A_n - D_n(a)T_n(2-2\cos\theta)$ is bounded by $2\,\omega_a(3h/2)$, and the modulus of each off-diagonal entry of $A_n - D_n(a)T_n(2-2\cos\theta)$ is bounded by $\omega_a(3h/2)$. Therefore, the 1-norm and the ∞-norm of $A_n - D_n(a)T_n(2-2\cos\theta)$ are bounded by $4\,\omega_a(3h/2)$, and so, by (3.1),

$$\|A_n - D_n(a)T_n(2-2\cos\theta)\| \le 4\,\omega_a(3h/2) \to 0 \ \text{ as } n\to\infty.$$

Setting $Z_n = A_n - D_n(a)T_n(2-2\cos\theta)$, we have $\{Z_n\}_n \sim_\sigma 0$ by **Z 1** (or **Z 2**). Since

$$A_n = D_n(a)T_n(2-2\cos\theta) + Z_n,$$

GLT 3 and **GLT 4** yield (3.14). □

Remark 3.1 (formal structure of the symbol). From a formal viewpoint (i.e., disregarding the regularity of $a(x)$ and $u(x)$), problem (3.10) can be rewritten in the form

$$\begin{cases} -a(x)u''(x) - a'(x)u'(x) = f(x), & x\in(0,1),\\ u(0)=\alpha, \quad u(1)=\beta. \end{cases}$$

From this reformulation, it appears more clearly that the symbol $a(x)(2-2\cos\theta)$ consists of the two 'ingredients':

- The coefficient of the higher-order differential operator, namely $a(x)$, in the physical variable x. To make a parallelism with Hörmander's theory [38], the higher-order differential operator $-a(x)u''(x)$ is the so-called principal symbol of the complete differential operator $-a(x)u''(x)-a'(x)u'(x)$ and $a(x)$ is then the coefficient of the principal symbol.
- The trigonometric polynomial associated with the FD formula $(-1,2,-1)$ used to approximate the higher-order derivative $-u''(x)$, namely $2-2\cos\theta = -\mathrm{e}^{\mathrm{i}\theta}+2-\mathrm{e}^{-\mathrm{i}\theta}$, in the Fourier variable θ. To see that $(-1,2,-1)$ is precisely the FD formula used to approximate $-u''(x)$, simply imagine $a(x)=1$ and note that in this case the FD scheme (3.11) becomes
$$-u''(x_j) \approx \frac{-u(x_{j+1})+2u(x_j)-u(x_{j-1})}{h^2},$$
i.e., the FD formula $(-1,2,-1)$ to approximate $-u''(x_j)$.

We observe that the term $-a'(x)u'(x)$, which only depends on lower-order derivatives of $u(x)$, does not enter the expression of the symbol.

Remark 3.2 (nonnegativity and order of the zero at $\theta = 0$). The trigonometric polynomial $2-2\cos\theta$ is nonnegative on $[-\pi,\pi]$ and it has a unique zero of order 2 at $\theta=0$, because

$$\lim_{\theta\to 0}\frac{2-2\cos\theta}{\theta^2} = 1.$$

This reflects the fact that the associated FD formula $(-1,2,-1)$ approximates $-u''(x)$, which is a differential operator of order 2 (it is also nonnegative on the space of functions $v\in C^2([0,1])$ such that $v(0)=v(1)=0$, in the sense that $\int_0^1 -v''(x)v(x)\mathrm{d}x = \int_0^1 (v'(x))^2\mathrm{d}x \ge 0$ for all such v).

3.2.2 FD Discretization of Convection-Diffusion-Reaction Equations

1st Part

Suppose we add to the diffusion equation (3.10) a convection and a reaction term. In this way, we obtain the following convection-diffusion-reaction equation in divergence form with Dirichlet boundary conditions:

$$\begin{cases} -(a(x)u'(x))' + b(x)u'(x) + c(x)u(x) = f(x), & x\in(0,1),\\ u(0)=\alpha, \quad u(1)=\beta, \end{cases} \tag{3.17}$$

where $a:[0,1]\to\mathbb{R}$ is continuous as before and we assume that $b,c:[0,1]\to\mathbb{R}$ are bounded. Based on Remark 3.1, we expect that the term $b(x)u'(x)+c(x)u(x)$, which only involves lower-order derivatives of $u(x)$, does not enter the expression of the symbol. In other words, if we discretize the higher-order term $-(a(x)u'(x))'$ as in (3.11), the symbol of the resulting FD discretization matrices B_n should be again $a(x)(2-2\cos\theta)$. We are going to show that this is in fact the case.

FD discretization Let $n \in \mathbb{N}$, set $h = \frac{1}{n+1}$ and $x_j = jh$ for all $j \in [0, n+1]$. Consider the discretization of (3.17) by the FD scheme defined as follows.

- To approximate the higher-order (diffusion) term $-(a(x)u'(x))'$, use again the FD formula (3.11), i.e.,

$$-(a(x)u'(x))'|_{x=x_j} \approx \frac{-a(x_{j+\frac{1}{2}})u(x_{j+1}) + \big(a(x_{j+\frac{1}{2}}) + a(x_{j-\frac{1}{2}})\big)u(x_j) - a(x_{j-\frac{1}{2}})u(x_{j-1})}{h^2}. \tag{3.18}$$

- To approximate the convection term $b(x)u'(x)$, use any (consistent) FD formula; to fix the ideas, here we use the second-order central formula

$$b(x)u'(x)|_{x=x_j} \approx b(x_j)\frac{u(x_{j+1}) - u(x_{j-1})}{2h}. \tag{3.19}$$

- To approximate the reaction term $c(x)u(x)$, use the obvious equation

$$c(x)u(x)|_{x=x_j} = c(x_j)u(x_j). \tag{3.20}$$

The resulting FD discretization matrix B_n admits a natural decomposition as

$$B_n = A_n + Z_n, \tag{3.21}$$

where A_n is the matrix coming from the discretization of the higher-order (diffusion) term $-(a(x)u'(x))$, while Z_n is the matrix coming from the discretization of the lower-order (convection and reaction) terms $b(x)u'(x)$ and $c(x)u(x)$. Note that A_n is given by (3.13) and Z_n is given by

$$Z_n = \frac{h}{2}\begin{bmatrix} 0 & b_1 & & & \\ -b_2 & 0 & b_2 & & \\ & \ddots & \ddots & \ddots & \\ & & -b_{n-1} & 0 & b_{n-1} \\ & & & -b_n & 0 \end{bmatrix} + h^2 \begin{bmatrix} c_1 & & & & \\ & c_2 & & & \\ & & \ddots & & \\ & & & c_{n-1} & \\ & & & & c_n \end{bmatrix}, \tag{3.22}$$

where $b_i = b(x_i)$ and $c_i = c(x_i)$ for all $i = 1, \ldots, n$.

GLT analysis of the FD discretization matrices We now prove that Theorem 3.3 holds unchanged with B_n in place of A_n. This highlights a general aspect: *lower-order terms such as $b(x)u'(x) + c(x)u(x)$ do not enter the expression of the symbol and do not affect in any way the asymptotic singular value and spectral distribution of DE discretization matrices.*

Theorem 3.4. *If $a \in C([0,1])$ and $b, c : [0,1] \to \mathbb{R}$ are bounded then*

$$\{B_n\}_n \sim_{\rm GLT} a(x)(2 - 2\cos\theta) \tag{3.23}$$

and

$$\{B_n\}_n \sim_{\sigma,\lambda} a(x)(2 - 2\cos\theta). \tag{3.24}$$

Proof. By (3.1), the matrix Z_n in (3.22) satisfies

$$\|Z_n\| \le h\|b\|_\infty + h^2\|c\|_\infty \le C/n \tag{3.25}$$

for some constant C independent of n. As a consequence, $\{Z_n\}_n$ is zero-distributed by **Z 1** (or **Z 2**), hence $\{Z_n\}_n \sim_{\rm GLT} 0$ by **GLT 3**. Since $\{A_n\}_n \sim_{\rm GLT} a(x)(2 - 2\cos\theta)$ by Theorem 3.3, the decomposition (3.21) and **GLT 4** imply (3.23).

Now, if the convection term is not present, i.e., $b(x) = 0$ identically, then B_n is symmetric and (3.24) follows from (3.23) and **GLT 1**. If $b(x)$ is not identically 0, then B_n is not symmetric in general and so (3.23) and **GLT 1** only imply the singular value distribution $\{B_n\}_n \sim_\sigma a(x)(2 - 2\cos\theta)$. Nevertheless, in view of the decomposition (3.21), since A_n is symmetric and $\{A_n\}_n \sim_\lambda a(x)(2 - 2\cos\theta)$ by Theorem 3.3, since $\|Z_n\|_1 = O(1)$ by the inequalities (3.25) and (3.3), and since $\|A_n\| \le 4\|a\|_\infty$ by (3.1), the spectral distribution $\{B_n\}_n \sim_\lambda a(x)(2 - 2\cos\theta)$ holds (by **GLT 2**) even if $b(x)$ is an arbitrary bounded function. ☐

2nd Part

So far, we only considered differential equations with Dirichlet boundary conditions. A natural question is the following: if we change the boundary conditions in (3.17), does the expression of the symbol change? The answer is 'no': *boundary conditions do not affect the singular value and eigenvalue distribution because they only produce a small-rank perturbation in the resulting discretization matrices*. To understand better this point, we consider problem (3.17) with Neumann boundary conditions:

$$\begin{cases} -(a(x)u'(x))' + b(x)u'(x) + c(x)u(x) = f(x), & x \in (0,1), \\ u'(0) = \alpha, \quad u'(1) = \beta. \end{cases} \tag{3.26}$$

FD discretization We discretize (3.26) by the same FD scheme considered in the 1st part, which is defined by the FD formulas (3.18)–(3.20). In this way, we arrive at the linear system

$$\begin{aligned} &-a(x_{j+\frac{1}{2}})u_{j+1} + \big(a(x_{j+\frac{1}{2}}) + a(x_{j-\frac{1}{2}})\big)u_j - a(x_{j-\frac{1}{2}})u_{j-1} \\ &+ \frac{h}{2}\big(b(x_j)u_{j+1} - b(x_j)u_{j-1}\big) + h^2 c(x_j)u_j = h^2 f(x_j), \qquad j = 1,\ldots,n, \end{aligned} \tag{3.27}$$

which is formed by n equations in the $n+2$ unknowns $u_0, u_1, \ldots, u_n, u_{n+1}$. Note that u_0 and u_{n+1} should now be considered as unknowns, because they are not specified by the Dirichlet boundary conditions. However, as it is common in the FD context, u_0 and u_{n+1} are expressed in terms of $u_1, \ldots, u_n$ by exploiting the Neumann boundary conditions. The simplest choice is to express u_0 and u_{n+1} as a function of u_1 and u_n, respectively, by imposing the conditions

$$\frac{u_1 - u_0}{h} = \alpha, \qquad \frac{u_{n+1} - u_n}{h} = \beta, \tag{3.28}$$

which yield $u_0 = u_1 - \alpha h$ and $u_{n+1} = u_n + \beta h$. Substituting into (3.27), we obtain a linear system with n equations and n unknowns $u_1, \ldots, u_n$. Setting $a_i = a(x_i)$, $b_i = b(x_i)$, $c_i = c(x_i)$ for all $i \in [0, n+1]$, the matrix of this system is

$$C_n = B_n + R_n = A_n + Z_n + R_n, \tag{3.29}$$

where A_n, B_n, Z_n are given by (3.13), (3.21), (3.22), respectively, and

$$R_n = \begin{bmatrix} -a_{\frac{1}{2}} - \frac{h}{2}b_1 & & \\ & & \\ & & -a_{n+\frac{1}{2}} + \frac{h}{2}b_n \end{bmatrix}$$

is a small-rank correction coming from the discretization (3.28) of the boundary conditions.

GLT analysis of the FD discretization matrices We prove that Theorems 3.3 and 3.4 hold unchanged with C_n in place of A_n and B_n, respectively.

Theorem 3.5. *If $a \in C([0,1])$ and $b, c : [0,1] \to \mathbb{R}$ are bounded then*

$$\{C_n\}_n \sim_{\rm GLT} a(x)(2 - 2\cos\theta) \tag{3.30}$$

and

$$\{C_n\}_n \sim_{\sigma,\lambda} a(x)(2 - 2\cos\theta). \tag{3.31}$$

Proof. Let C denote a generic constant independent of n. It is clear that $\|R_n\| \le \|a\|_\infty + (h/2)\|b\|_\infty \le C$. Moreover, since $\|R_n\|_1 \le \text{rank}(R_n)\|R_n\| \le C$, the matrix-sequence $\{R_n\}_n$ is zero-distributed by **Z 2**. Note that $\{Z_n\}_n$ is zero-distributed as well because $\|Z_n\| \le C/n$ by (3.25). In view of the decomposition (3.29), Theorem 3.3 and **GLT 3 – GLT 4** imply (3.30).

If the matrices C_n are symmetric (this happens if $b(x) = 0$), from (3.30) and **GLT 1** we immediately obtain (3.31). If the matrices C_n are not symmetric, from (3.30) and **GLT 1** we only obtain the singular value distribution in (3.31). However, in view of (3.29), since $\|R_n + Z_n\|_1 = o(n)$ and $\|R_n + Z_n\|, \|A_n\| \le C$, the spectral distribution in (3.31) holds (by **GLT 2**) even if the matrices C_n are not symmetric. □

3rd Part

Consider the following convection-diffusion-reaction problem:

$$\begin{cases} -a(x)u''(x)+b(x)u'(x)+c(x)u(x)=f(x), & x\in(0,1),\\ u(0)=\alpha, \quad u(1)=\beta, \end{cases} \tag{3.32}$$

where $a:[0,1]\to\mathbb{R}$ is continuous and $b,c:[0,1]\to\mathbb{R}$ are bounded. The difference with respect to problem (3.17) is that the higher-order differential operator now appears in non-divergence form, i.e., we have $-a(x)u''(x)$ instead of $-(a(x)u'(x))'$. Neverthelss, based on Remark 3.1, if we use again the FD formula $(-1,2,-1)$ to discretize the second derivative $-u''(x)$, the symbol of the resulting FD discretization matrices should be again $a(x)(2-2\cos\theta)$. We are going to show that this is in fact the case.

FD discretization Let $n\in\mathbb{N}$, set $h=\frac{1}{n+1}$ and $x_j=jh$ for all $j=0,\ldots,n+1$. We discretize again (3.32) by the central second-order FD scheme, which in this case is defined by the following formulas:

$$-a(x)u''(x)|_{x=x_j}\approx a(x_j)\frac{-u(x_{j+1})+2u(x_j)-u(x_{j-1})}{h^2}, \qquad j=1,\ldots,n,$$
$$b(x)u'(x)|_{x=x_j}\approx b(x_j)\frac{u(x_{j+1})-u(x_{j-1})}{2h}, \qquad j=1,\ldots,n,$$
$$c(x)u(x)|_{x=x_j}=c(x_j)u(x_j), \qquad j=1,\ldots,n.$$

Then, we approximate the solution of (3.32) by the piecewise linear function that takes the value u_j at the point x_j for $j=0,\ldots,n+1$, where $u_0=\alpha$, $u_{n+1}=\beta$, and $\mathbf{u}=(u_1,\ldots,u_n)^T$ solves the linear system

$$a(x_j)(-u_{j+1}+2u_j-u_{j-1})+\frac{h}{2}b(x_j)(u_{j+1}-u_{j-1})+h^2c(x_j)u_j=h^2f(x_j), \qquad j=1,\ldots,n.$$

The matrix E_n of this linear system can be decomposed according to the diffusion, convection and reaction term, as follows:

$$E_n=K_n+Z_n, \tag{3.33}$$

where Z_n is the sum of the convection and reaction matrix and is given by (3.22), while

$$K_n=\begin{bmatrix} 2a_1 & -a_1 & & & \\ -a_2 & 2a_2 & -a_2 & & \\ & \ddots & \ddots & \ddots & \\ & & -a_{n-1} & 2a_{n-1} & -a_{n-1}\\ & & & -a_n & 2a_n \end{bmatrix} \tag{3.34}$$

is the diffusion matrix ($a_i=a(x_i)$ for all $i=1,\ldots,n$).

GLT analysis of the FD discretization matrices Despite the nonsymmetry of the diffusion matrix, which is due to the non-divergence form of the higher-order (diffusion) operator $-a(x)u''(x)$, we will prove that Theorems 3.3, 3.4, 3.5 hold unchanged with E_n in place of A_n, B_n, C_n, respectively.

Theorem 3.6. *If $a\in C([0,1])$ and $b,c:[0,1]\to\mathbb{R}$ are bounded then*

$$\{E_n\}_n\sim_{\rm GLT} a(x)(2-2\cos\theta) \tag{3.35}$$

and

$$\{E_n\}_n\sim_{\sigma,\lambda} a(x)(2-2\cos\theta). \tag{3.36}$$

Proof. Throughout this proof, the letter C will denote a generic constant independent of n. By (3.25),

$$\|Z_n\|\le C/n,$$

hence $\{Z_n\}_n$ is zero-distributed. We prove that

$$\{K_n\}_n \sim_{\rm GLT} a(x)(2-2\cos\theta), \tag{3.37}$$

after which (3.35) will follow from **GLT 3 – GLT 4** and the decomposition (3.33). It is clear from (3.34) that

$$K_n = \operatorname*{diag}_{i=1,\ldots,n}(a_i)\, T_n(2-2\cos\theta).$$

By **T 3** applied with $p=\infty$, we obtain

$$\|K_n - D_n(a)T_n(2-2\cos\theta)\| \le \Big\|\operatorname*{diag}_{i=1,\ldots,n}(a_i) - D_n(a)\Big\|\,\|T_n(2-2\cos\theta)\| \le \omega_a(h)\,\|2-2\cos\theta\|_\infty = 4\,\omega_a(h),$$

which tends to 0 as $n\to\infty$. We conclude that $\{K_n - D_n(a)T_n(2-2\cos\theta)\}_n$ is zero-distributed, and so (3.37) follows from **GLT 3 – GLT 4**.

From (3.35) and **GLT 1** we obtain the singular value distribution in (3.36). To obtain the spectral distribution, the idea is to exploit the fact that K_n is 'almost' symmetric, because $a(x)$ varies continuously when x ranges in $[0,1]$, and so $a(x_j)\approx a(x_{j+1})$ for all $j=1,\ldots,n-1$ (when n is large enough). Therefore, by replacing K_n with one of its symmetric approximations $\tilde K_n$, we can write

$$E_n = \tilde K_n + (K_n - \tilde K_n) + Z_n, \tag{3.38}$$

and in view of the decomposition (3.38) we want to obtain the spectral distribution in (3.36) from **GLT 2** applied with $X_n = \tilde K_n$ and $Y_n = (K_n-\tilde K_n)+Z_n$. Let

$$\tilde K_n = \begin{bmatrix} 2a_1 & -a_1 & & & \\ -a_1 & 2a_2 & -a_2 & & \\ & \ddots & \ddots & \ddots & \\ & & -a_{n-2} & 2a_{n-1} & -a_{n-1} \\ & & & -a_{n-1} & 2a_n \end{bmatrix}. \tag{3.39}$$

Since

$$\|K_n-\tilde K_n\| \le \sqrt{|K_n-\tilde K_n|_1\,|K_n-\tilde K_n|_\infty} \le \max_{i=1,\ldots,n-1}|a_{i+1}-a_i| \le \omega_a(h)\to 0,$$

$$\|K_n\| \le \sqrt{|K_n|_1|K_n|_\infty} \le 4\|a\|_\infty \le C,$$

$$\|Z_n\|\to 0,$$

it follows from **GLT 2** that $\{E_n\}_n \sim_\lambda a(x)(2-2\cos\theta)$. ☐

Remark 3.3. In the proof of Theorem 3.6 we could also choose

$$\tilde K_n = S_n(a)\circ T_n(2-2\cos\theta) = \begin{bmatrix} 2\tilde a_1 & -\tilde a_1 & & & \\ -\tilde a_1 & 2\tilde a_2 & -\tilde a_2 & & \\ & \ddots & \ddots & \ddots & \\ & & -\tilde a_{n-2} & 2\tilde a_{n-1} & -\tilde a_{n-1} \\ & & & -\tilde a_{n-1} & 2\tilde a_n \end{bmatrix},$$

where $\tilde a_i = a(\frac{i}{n})$ for all $i=1,\ldots,n$ and $S_n(a)$ is the arrow-shaped sampling matrix defined in (3.6). With this choice of $\tilde K_n$, nothing changes in the proof of Theorem 3.6 except for the bound of $\|K_n-\tilde K_n\|$, which becomes $\|K_n-\tilde K_n\|\le 4\,\omega_a(h)$.

4th Part

Based on Remark 3.1, if we change the FD scheme to discretize the differential problem (3.32), the symbol should become $a(x)p(\theta)$, where $p(\theta)$ is the trigonometric polynomial associated with the new FD formula used to approximate the second derivative $-u''(x)$ (the higher-order differential operator). We are going to show through an example that this is indeed the case.

FD discretization Consider the convection-diffusion-reaction problem (3.32). Instead of the second-order central FD scheme $(-1,2,-1)$, this time we use the fourth-order central FD scheme $\frac{1}{12}(1,-16,30,-16,1)$ to approximate the second derivative $-u''(x)$. In other words, for $j=2,\ldots,n-1$ we approximate the higher-order term $-a(x)u''(x)$ by the FD formula

$$-a(x)u''(x)|_{x=x_j} \approx a(x_j)\frac{u(x_{j+2})-16u(x_{j+1})+30u(x_j)-16u(x_{j-1})+u(x_{j-2})}{12h^2},$$

while for $j=1,n$ we use again the FD scheme $(-1,2,-1)$,

$$-a(x)u''(x)|_{x=x_j} \approx a(x_j)\frac{-u(x_{j+1})+2u(x_j)-u(x_{j-1})}{h^2}.$$

From a numerical viewpoint, this is not a good choice because the FD formula $\frac{1}{12}(1,-16,30,-16,1)$ is a very accurate fourth-order formula, and in order not to destroy the accuracy one would gain from this formula, one should use a fourth-order scheme also for $j=1,n$ instead of the classical $(-1,2,-1)$. However, in this work we are not concerned with this kind of issues and we use the classical $(-1,2,-1)$ because it is simpler and allows us to better illustrate the GLT analysis without introducing useless technicalities. As already observed before, the FD schemes used to approximate the lower-order terms $b(x)u'(x)$ and $c(x)u(x)$ do not affect the symbol, as well as the singular value and eigenvalue distribution, of the resulting sequence of discretization matrices. To illustrate once again this point, in this example we assume to approximate $b(x)u'(x)$ and $c(x)u(x)$ by the following 'strange' FD formulas: for $j=1,\ldots,n$,

$$b(x)u'(x)|_{x=x_j} \approx b(x_j)\frac{u(x_j)-u(x_{j-1})}{h},$$
$$c(x)u(x)|_{x=x_j} \approx c(x_j)\frac{u(x_{j+1})+u(x_j)+u(x_{j-1})}{3}.$$

Setting $a_i=a(x_i)$, $b_i=b(x_i)$, $c_i=c(x_i)$ for all $i=1,\ldots,n$, the resulting FD discretization matrix P_n can be decomposed according to the diffusion, convection and reaction term, as follows:

$$P_n = K_n + Z_n,$$

where Z_n is the sum of the convection and reaction matrix,

$$Z_n = h\begin{bmatrix} b_1 & & & & \\ -b_2 & b_2 & & & \\ & \ddots & \ddots & & \\ & & -b_{n-1} & b_{n-1} & \\ & & & -b_n & b_n \end{bmatrix} + \frac{h^2}{3}\begin{bmatrix} c_1 & c_1 & & & \\ c_2 & c_2 & c_2 & & \\ & \ddots & \ddots & \ddots & \\ & & c_{n-1} & c_{n-1} & c_{n-1} \\ & & & c_n & c_n \end{bmatrix},$$

while K_n is the diffusion matrix,

$$K_n = \frac{1}{12}\begin{bmatrix} 24a_1 & -12a_1 & & & & & \\ -16a_2 & 30a_2 & -16a_2 & a_2 & & & \\ a_3 & -16a_3 & 30a_3 & -16a_3 & a_3 & & \\ & \ddots & \ddots & \ddots & \ddots & \ddots & \\ & & a_{n-2} & -16a_{n-2} & 30a_{n-2} & -16a_{n-2} & a_{n-2} \\ & & & a_{n-1} & -16a_{n-1} & 30a_{n-1} & -16a_{n-1} \\ & & & & & -12a_n & 24a_n \end{bmatrix}.$$

GLT analysis of the FD discretization matrices Let $p(\theta)$ be the trigonometric polynomial associated with the FD formula $\frac{1}{12}(1,-16,30,-16,1)$ used to approximate the second derivative $-u''(x)$, i.e.,

$$p(\theta) = \frac{1}{12}(\mathrm{e}^{-2\mathrm{i}\theta}-16\mathrm{e}^{-\mathrm{i}\theta}+30-16\mathrm{e}^{\mathrm{i}\theta}+\mathrm{e}^{2\mathrm{i}\theta}) = \frac{1}{12}(30-32\cos\theta+2\cos(2\theta)).$$

Based on Remark 3.1, the following result is not unexpected.

Theorem 3.7. *If $a \in C([0,1])$ and $b, c : [0,1] \to \mathbb{R}$ are bounded then*

$$\{P_n\}_n \sim_{\rm GLT} a(x)p(\theta) \tag{3.40}$$

and

$$\{P_n\}_n \sim_{\sigma,\lambda} a(x)p(\theta). \tag{3.41}$$

Proof. Throughout this proof, the letter C will denote a generic constant independent of n. To simultaneously obtain (3.40) and (3.41), we consider the following decomposition of P_n:

$$P_n = \tilde{K}_n + (K_n - \tilde{K}_n) + Z_n,$$

where $\tilde{K}_n$ is the symmetric approximation of K_n given by

$$\tilde{K}_n = S_n(a) \circ T_n(p) = \frac{1}{12}\begin{bmatrix} 30\tilde{a}_1 & -16\tilde{a}_1 & \tilde{a}_1 & & & & \\ -16\tilde{a}_1 & 30\tilde{a}_2 & -16\tilde{a}_2 & \tilde{a}_2 & & & \\ \tilde{a}_1 & -16\tilde{a}_2 & 30\tilde{a}_3 & -16\tilde{a}_3 & \tilde{a}_3 & & \\ & \ddots & \ddots & \ddots & \ddots & \ddots & \\ & & \tilde{a}_{n-4} & -16\tilde{a}_{n-3} & 30\tilde{a}_{n-2} & -16\tilde{a}_{n-2} & \tilde{a}_{n-2} \\ & & & \tilde{a}_{n-3} & -16\tilde{a}_{n-2} & 30\tilde{a}_{n-1} & -16\tilde{a}_{n-1} \\ & & & & \tilde{a}_{n-2} & -16\tilde{a}_{n-1} & 30\tilde{a}_n \end{bmatrix}$$

($\tilde{a}_i = a(\frac{i}{n})$ for all $i = 1,\ldots,n$). We show that:

(a) $\{\tilde{K}_n\}_n \sim_{\rm GLT} a(x)p(\theta)$;

(b) $\|K_n\|, \|\tilde{K}_n\| \le C$ and $\|Z_n\| \to 0$;

(c) $\|K_n - \tilde{K}_n\|_1 = o(n)$.

Note that (b)–(c) imply that $\{(K_n - \tilde{K}_n) + Z_n\}_n \sim_\sigma 0$ by **Z 2**. Once we have proved (a)–(c), the GLT relation (3.40) follows from **GLT 4**, the singular value distribution in (3.41) follows from (3.40) and **GLT 1**, and the spectral distribution in (3.41) follows from **GLT 2** applied with $X_n = \tilde{K}_n$ and $Y_n = (K_n - \tilde{K}_n) + Z_n$.

Proof of (a). See Theorem 3.2.

Proof of (b). We have

$$\|Z_n\| \le \sqrt{|Z_n|_1\,|Z_n|_\infty} \le 2h\|b\|_\infty + h^2\|c\|_\infty \to 0,$$

$$\|K_n\| \le \sqrt{|K_n|_1\,|K_n|_\infty} \le \frac{64}{12}\|a\|_\infty,$$

$$\|\tilde{K}_n\| \le \sqrt{|\tilde{K}_n|_1\,|\tilde{K}_n|_\infty} \le \frac{64}{12}\|a\|_\infty.$$

Note that the uniform boundedness of $\|\tilde{K}_n\|$ with respect to n was already known from Theorem 3.2.

Proof of (c). A direct comparison between K_n and $\tilde{K}_n$ shows that

$$K_n = \tilde{K}_n + R_n + N_n,$$

where $N_n = K_n - \tilde{K}_n - R_n$ and R_n is the matrix whose rows are all zeros except for the first and the last one, which are given by

$$[24a_1 - 30\tilde{a}_1 \quad -12a_1 + 16\tilde{a}_1 \quad -\tilde{a}_1 \quad 0 \quad \cdots \quad 0]$$

and

$$[0 \quad \cdots \quad 0 \quad -\tilde{a}_{n-2} \quad -12a_n + 16\tilde{a}_{n-1} \quad 24a_n - 30\tilde{a}_n],$$

respectively. We have

$$\|R_n\| \le \frac{83}{12}\|a\|_\infty, \qquad \mathrm{rank}(R_n) \le 2, \qquad \|N_n\| \le \frac{64}{12}\,\omega_a\Big(\frac{2}{n}\Big)$$

and

$$\|K_n - \tilde{K}_n\|_1 \le \|R_n\|_1 + \|N\|_1 \le \mathrm{rank}(R_n)\|R_n\| + n\|N_n\|,$$

hence $\|K_n - \tilde{K}_n\|_1 = o(n)$. □

Remark 3.4 (nonnegativity and order of the zero at $\theta = 0$). Despite we have changed the FD scheme to approximate the second derivative $-u''(x)$, the resulting trigonometric polynomial $p(\theta)$ retains some properties of $2-2\cos\theta$. In particular, $p(\theta)$ is nonnegative over $[-\pi,\pi]$ and it has a unique zero of order 2 at $\theta = 0$, because

$$\lim_{\theta\to 0}\frac{p(\theta)}{\theta^2} = 1 = \lim_{\theta\to 0}\frac{2-2\cos\theta}{\theta^2}.$$

This reflects the fact the the associated FD formula $\frac{1}{12}(1,-16,30,-16,1)$ approximates $-u''(x)$, which is a differential operator of order 2 and it is also nonnegative on $\{v \in C^2([0,1]) :\ v(0) = v(1) = 0\}$; cf. Remark 3.2.

3.2.3 FD Discretization of Higher-Order Equations

So far we only considered the FD discretization of second-order differential equations. In order to show that the GLT analysis is not limited to second-order equations, in this section we deal with an higher-order problem. For simplicity, we focus on the following fourth-order problem with homogeneous Dirichlet–Neumann boundary conditions:

$$\begin{cases} a(x)u^{(4)}(x) = f(x), & x \in (0,1),\\ u(0) = 0, \quad u(1) = 0, & \\ u'(0) = 0, \quad u'(1) = 0, & \end{cases} \tag{3.42}$$

where $a \in C([0,1])$ and f is a given function. We do not consider more complicated boundary conditions, and we do not include terms with lower-order derivatives, because we know from Remark 3.1 and the experience gained from the previous section that both these ingredients only serve to complicate things, but ultimately they do not affect the symbol, as well as the singular value and eigenvalue distribution, of the resulting discretization matrices. Based on Remark 3.1, the symbol of the matrix-sequence arising from the FD discretization of (3.42) should be $a(x)q(\theta)$, where $q(\theta)$ is the trigonometric polynomial associated with the FD formula used to discretize $u^{(4)}(x)$. We will see that this is in fact the case.

FD discretization We approximate the fourth derivative $u^{(4)}(x)$ by the second-order central FD scheme $(1,-4,6,-4,1)$, which yields the approximation

$$a(x)u^{(4)}(x)|_{x=x_j} \approx a(x_j)\frac{u(x_{j+2}) - 4u(x_{j+1}) + 6u(x_j) - 4u(x_{j-1}) + u(x_{j-2})}{h^4},$$

for all $j = 2,\dots,n+1$; here, $h = \frac{1}{n+3}$ and $x_j = jh$ for $j = 0,\dots,n+3$. Taking into account the homogeneous boundary conditions, we approximate the solution of (3.42) by the piecewise linear function that takes the value u_j in x_j for $j = 0,\dots,n+3$, where $u_0 = u_1 = u_{n+2} = u_{n+3} = 0$ and $\mathbf{u} = (u_2,\dots,u_{n+1})^T$ is the solution of the linear system

$$a(x_j)(u_{j+2} - 4u_{j+1} + 6u_j - 4u_{j-1} + u_{j-2}) = h^4 f(x_j), \qquad j = 2,\dots,n+1.$$

The matrix A_n of this linear system is given by

$$A_n = \begin{bmatrix} 6a_2 & -4a_2 & a_2 & & & & \\ -4a_3 & 6a_3 & -4a_3 & a_3 & & & \\ a_4 & -4a_4 & 6a_4 & -4a_4 & a_4 & & \\ & \ddots & \ddots & \ddots & \ddots & \ddots & \\ & & a_{n-1} & -4a_{n-1} & 6a_{n-1} & -4a_{n-1} & a_{n-1} \\ & & & a_n & -4a_n & 6a_n & -4a_n \\ & & & & a_{n+1} & -4a_{n+1} & 6a_{n+1} \end{bmatrix},$$

where $a_i = a(x_i)$ for all $i = 2,\dots,n+1$.

GLT analysis of the FD discretization matrices Let $q(\theta)$ be the trigonometric polynomial associated with the FD formula $(1,-4,6,-4,1)$, i.e.,

$$q(\theta)=\mathrm{e}^{-2\mathrm{i}\theta}-4\mathrm{e}^{-\mathrm{i}\theta}+6-4\mathrm{e}^{\mathrm{i}\theta}+\mathrm{e}^{2\mathrm{i}\theta}=6-8\cos\theta+2\cos(2\theta).$$

Theorem 3.8. *If $a\in C([0,1])$ then*

$$\{A_n\}_n\sim_{\mathrm{GLT}} a(x)q(\theta) \tag{3.43}$$

and

$$\{A_n\}_n\sim_{\sigma,\lambda} a(x)q(\theta), \tag{3.44}$$

Proof. We show that

$$\|A_n-S_n(a)\circ T_n(q)\|\to 0. \tag{3.45}$$

Once this is proved, since $\{S_n(a)\circ T_n(q)\}_n\sim_{\mathrm{GLT}} a(x)q(\theta)$ and $\|S_n(a)\circ T_n(q)\|$ is uniformly bounded with respect to n (by Theorem 3.2), and since $\|A_n\|\le 16\|a\|_\infty$ by (3.1), the relations (3.43)–(3.44) follow from the decomposition

$$A_n=S_n(a)\circ T_n(q)+(A_n-S_n(a)\circ T_n(q))$$

and from **GLT 1 – GLT 4**, taking into account that $S_n(a)\circ T_n(p)$ is symmetric and $\{A_n-S_n(a)\circ T_n(q)\}_n$ is zero-distributed by (3.45) and **Z 1** (or **Z 2**). Let us then prove (3.45). The matrices A_n and $S_n(a)\circ T_n(q)$ are banded (pentadiagonal) and, for all $i,j=1,\ldots,n$ with $|i-j|\le 2$, a crude estimates gives

$$\begin{aligned}\big|(A_n)_{ij}-(S_n(a)\circ T_n(q))_{ij}\big| &= \Big|a_{i+1}(T_n(q))_{ij}-a\Big(\frac{\min(i,j)}{n}\Big)(T_n(q))_{ij}\Big|\\ &= \Big|a\Big(\frac{i+1}{n+3}\Big)-a\Big(\frac{\min(i,j)}{n}\Big)\Big|\,|(T_n(q))_{ij}|\\ &\le 6\,\omega_a\Big(\frac{6}{n}\Big).\end{aligned}$$

Hence, by (3.1), $\|A_n-S_n(a)\circ T_n(q)\|\le 5\cdot 6\,\omega_a(\frac{6}{n})\to 0.$ □

Remark 3.5 (nonnegativity and order of the zero at $\theta=0$). The polynomial $q(\theta)$ is nonnegative over $[-\pi,\pi]$ and has a unique zero of order 4 at $\theta=0$, because

$$\lim_{\theta\to 0}\frac{q(\theta)}{\theta^4}=1.$$

This reflects the fact that the FD formula $(1,-4,6,-4,1)$ associated with $q(\theta)$ approximates the fourth derivative $u^{(4)}(x)$, which is a differential operator of order 4 (it is also nonnegative on the space of functions $v\in C^4([0,1])$ such that $v(0)=v(1)=0$ and $v'(0)=v'(1)=0$, in the sense that $\int_0^1 v^{(4)}(x)v(x)\mathrm{d}x=\int_0^1(v''(x))^2\mathrm{d}x\ge 0$ for all such v); see also Remarks 3.2 and 3.4.

3.2.4 Non-uniform FD Discretizations

All the FD discretizations considered in the previous sections are based on uniform grids. It is natural to ask whether the theory of GLT sequences finds applications also in the context of non-uniform FD discretizations. The answer to this question is affirmative, at least in the case where the non-uniform grid is obtained as the mapping of a uniform grid through a fixed function G, independent of the mesh size. In this section we illustrate this claim by means of a simple example.

FD discretization Consider the diffusion equation (3.10) with $a\in C([0,1])$. Take a discretization parameter $n\in\mathbb{N}$, fix a set of grid points $0=x_0<x_1<\ldots<x_{n+1}=1$ and define the corresponding stepsizes $h_j=x_j-x_{j-1},\ j=1,\ldots,n+1$. For each $j=1,\ldots,n$, we approximate $-(a(x)u'(x))'|_{x=x_j}$ by the FD formula

$$\begin{aligned}-(a(x)u'(x))'|_{x=x_j} &\approx -\frac{a(x_j+\frac{h_{j+1}}{2})u'(x_j+\frac{h_{j+1}}{2})-a(x_j-\frac{h_j}{2})u'(x_j-\frac{h_j}{2})}{\frac{h_{j+1}}{2}+\frac{h_j}{2}}\\ &\approx -\frac{a(x_j+\frac{h_{j+1}}{2})\dfrac{u(x_{j+1})-u(x_j)}{h_{j+1}}-a(x_j-\frac{h_j}{2})\dfrac{u(x_j)-u(x_{j-1})}{h_j}}{\frac{h_{j+1}}{2}+\frac{h_j}{2}}\\ &= \frac{2}{h_j+h_{j+1}}\left[-\frac{a(x_j-\frac{h_j}{2})}{h_j}u(x_{j-1})+\left(\frac{a(x_j-\frac{h_j}{2})}{h_j}+\frac{a(x_j+\frac{h_{j+1}}{2})}{h_{j+1}}\right)u(x_j)-\frac{a(x_j+\frac{h_{j+1}}{2})}{h_{j+1}}u(x_{j+1})\right].\end{aligned}$$

This means that the nodal values of the solution u satisfy (approximately) the following linear system:

$$-\frac{a(x_j-\frac{h_j}{2})}{h_j}u(x_{j-1})+\left(\frac{a(x_j-\frac{h_j}{2})}{h_j}+\frac{a(x_j+\frac{h_{j+1}}{2})}{h_{j+1}}\right)u(x_j)-\frac{a(x_j+\frac{h_{j+1}}{2})}{h_{j+1}}u(x_{j+1})=\frac{h_j+h_{j+1}}{2}f(x_j),\qquad j=1,\ldots,n.$$

We then approximate the solution by the piecewise linear function that takes the value u_j in x_j for $j=0,\ldots,n+1$, where $u_0=\alpha$, $u_{n+1}=\beta$, and $\mathbf{u}=(u_1,\ldots,u_n)^T$ solves

$$-\frac{a(x_j-\frac{h_j}{2})}{h_j}u_{j-1}+\left(\frac{a(x_j-\frac{h_j}{2})}{h_j}+\frac{a(x_j+\frac{h_{j+1}}{2})}{h_{j+1}}\right)u_j-\frac{a(x_j+\frac{h_{j+1}}{2})}{h_{j+1}}u_{j+1}=\frac{h_j+h_{j+1}}{2}f(x_j),\qquad j=1,\ldots,n.$$

The matrix of this linear system is the $n\times n$ tridiagonal symmetric matrix given by

$$\mathrm{tridiag}_n\left[-\frac{a(x_j-\frac{h_j}{2})}{h_j},\ \frac{a(x_j-\frac{h_j}{2})}{h_j}+\frac{a(x_j+\frac{h_{j+1}}{2})}{h_{j+1}},\ -\frac{a(x_j+\frac{h_{j+1}}{2})}{h_{j+1}}\right].\tag{3.46}$$

GLT analysis of the FD discretization matrices Let $h=\frac{1}{n+1}$ and $\hat{x}_j=jh$, $j=0,\ldots,n+1$. In the following, we assume that the set of points $\{x_0,x_1,\ldots,x_{n+1}\}$ is obtained as the mapping of the uniform grid $\{\hat{x}_0,\hat{x}_1,\ldots,\hat{x}_{n+1}\}$ through a fixed function G, i.e., $x_j=G(\hat{x}_j)$ for $j=0,\ldots,n+1$, where $G:[0,1]\to[0,1]$ is an increasing and bijective map, independent of the mesh parameter n. The resulting FD discretization matrix (3.46) will be denoted by $A_{G,n}$ in order to emphasize its dependence on G. In formulas,

$$A_{G,n}=\mathrm{tridiag}_n\left[-\frac{a(G(\hat{x}_j)-\frac{h_j}{2})}{h_j},\ \frac{a(G(\hat{x}_j)-\frac{h_j}{2})}{h_j}+\frac{a(G(\hat{x}_j)+\frac{h_{j+1}}{2})}{h_{j+1}},\ -\frac{a(G(\hat{x}_j)+\frac{h_{j+1}}{2})}{h_{j+1}}\right]\tag{3.47}$$

with

$$h_j=G(\hat{x}_j)-G(\hat{x}_{j-1}),\qquad j=1,\ldots,n.$$

Theorem 3.9. *Let $a\in C([0,1])$. Suppose $G:[0,1]\to[0,1]$ is an increasing bijective map in $C^1([0,1])$ and there exist at most finitely many points $\hat{x}$ such that $G'(\hat{x})=0$. Then*

$$\left\{\frac{1}{n+1}A_{G,n}\right\}_n\sim_{\mathrm{GLT}}\frac{a(G(\hat{x}))}{G'(\hat{x})}(2-2\cos\theta)\tag{3.48}$$

and

$$\left\{\frac{1}{n+1}A_{G,n}\right\}_n\sim_{\sigma,\lambda}\frac{a(G(\hat{x}))}{G'(\hat{x})}(2-2\cos\theta).\tag{3.49}$$

Proof. We only prove (3.48) because (3.49) follows immediately from (3.48) and **GLT 1** as the matrices $A_{G,n}$ are symmetric. Since $G\in C^1([0,1])$, for every $j=1,\ldots,n$ there exist $\alpha_j\in[\hat{x}_{j-1},\hat{x}_j]$ and $\beta_j\in[\hat{x}_j,\hat{x}_{j+1}]$ such that

$$h_j=G(\hat{x}_j)-G(\hat{x}_{j-1})=G'(\alpha_j)h=(G'(\hat{x}_j)+\delta_j)h,\tag{3.50}$$
$$h_{j+1}=G(\hat{x}_{j+1})-G(\hat{x}_j)=G'(\beta_j)h=(G'(\hat{x}_j)+\varepsilon_j)h,\tag{3.51}$$

where

$$\delta_j=G'(\alpha_j)-G'(\hat{x}_j),$$
$$\varepsilon_j=G'(\beta_j)-G'(\hat{x}_j).$$

Note that

$$|\delta_j|,|\varepsilon_j|\le\omega_{G'}(h),\qquad j=1,\ldots,n,$$

where $\omega_{G'}$ is the modulus of continuity of G'. In view of (3.50) and (3.51), we have, for each $j=1,\ldots,n$,

$$a\Big(G(\hat{x}_j)-\frac{h_j}{2}\Big)=a\Big(G(\hat{x}_j)-\frac{h}{2}(G'(\hat{x}_j)+\delta_j)\Big)=a(G(\hat{x}_j))+\mu_j,\tag{3.52}$$
$$a\Big(G(\hat{x}_j)+\frac{h_{j+1}}{2}\Big)=a\Big(G(\hat{x}_j)+\frac{h}{2}(G'(\hat{x}_j)+\varepsilon_j)\Big)=a(G(\hat{x}_j))+\eta_j,\tag{3.53}$$

where

$$\mu_j = a\Big(G(\hat{x}_j) - \frac{h}{2}(G'(\hat{x}_j)+\delta_j)\Big) - a(G(\hat{x}_j)),$$
$$\eta_j = a\Big(G(\hat{x}_j) + \frac{h}{2}(G'(\hat{x}_j)+\varepsilon_j)\Big) - a(G(\hat{x}_j)).$$

This time

$$|\mu_j|, |\eta_j| \le C_G\omega_a(h), \qquad j = 1,\ldots,n,$$

where ω_a is the modulus of continuity of a and C_G is a constant depending only on G. Substituting (3.50)–(3.53) in (3.47), we obtain

$$\frac{1}{n+1}A_{G,n} = hA_{G,n} = \mathrm{tridiag}_n\left[-\frac{a(G(\hat{x}_j))+\mu_j}{G'(\hat{x}_j)+\delta_j},\ \frac{a(G(\hat{x}_j))+\mu_j}{G'(\hat{x}_j)+\delta_j} + \frac{a(G(\hat{x}_j))+\eta_j}{G'(\hat{x}_j)+\varepsilon_j},\ -\frac{a(G(\hat{x}_j))+\eta_j}{G'(\hat{x}_j)+\varepsilon_j}\right]. \tag{3.54}$$

Consider the matrix

$$D_n\Big(\frac{a(G(\hat{x}))}{G'(\hat{x})}\Big)T_n(2-2\cos\theta) = \mathrm{tridiag}_n\left[-\frac{a(G(\hat{x}_j))}{G'(\hat{x}_j)},\ 2\,\frac{a(G(\hat{x}_j))}{G'(\hat{x}_j)},\ -\frac{a(G(\hat{x}_j))}{G'(\hat{x}_j)}\right]. \tag{3.55}$$

Note that this matrix seems to be an 'approximation' of $\frac{1}{n+1}A_{G,n}$; cf. (3.54) and (3.55). Since the function $a(G(\hat{x}))/G'(\hat{x})$ is continuous a.e., **GLT 3** and **GLT 4** yield

$$\left\{D_n\Big(\frac{a(G(\hat{x}))}{G'(\hat{x})}\Big)T_n(2-2\cos\theta)\right\}_n \sim_{\mathrm{GLT}} \frac{a(G(\hat{x}))}{G'(\hat{x})}(2-2\cos\theta).$$

We are going to show that

$$\left\{D_n\Big(\frac{a(G(\hat{x}))}{G'(\hat{x})}\Big)T_n(2-2\cos\theta)\right\}_n \xrightarrow{\text{a.c.s.}} \left\{\frac{1}{n+1}A_{G,n}\right\}_n. \tag{3.56}$$

Once this is proved, (3.48) follows immediately from **GLT 7**.

We first prove (3.56) in the case where $G'(\hat{x})$ does not vanish over $[0,1]$, so that

$$m_{G'} = \min_{\hat{x}\in[0,1]} G'(\hat{x}) > 0.$$

In this case, we will show directly that $\|Z_n\| \to 0$, where

$$Z_n = \frac{1}{n+1}A_{G,n} - D_n\Big(\frac{a(G(\hat{x}))}{G'(\hat{x})}\Big)T_n(2-2\cos\theta). \tag{3.57}$$

The matrix Z_n in (3.57) is tridiagonal and a straightforward computation based on (3.54)–(3.55) shows that all its components are bounded in modulus by a quantity that depends only on n,G,a and that converges to 0 as $n\to\infty$. For example, if $j = 2,\ldots,n$, then

$$\begin{aligned}|(Z_n)_{j,j-1}| &= \left|\frac{a(G(\hat{x}_j))+\mu_j}{G'(\hat{x}_j)+\delta_j} - \frac{a(G(\hat{x}_j))}{G'(\hat{x}_j)}\right| \le \left|\frac{a(G(\hat{x}_j))+\mu_j}{G'(\hat{x}_j)+\delta_j} - \frac{a(G(\hat{x}_j))}{G'(\hat{x}_j)+\delta_j}\right| + \left|\frac{a(G(\hat{x}_j))}{G'(\hat{x}_j)+\delta_j} - \frac{a(G(\hat{x}_j))}{G'(\hat{x}_j)}\right| \\ &= \left|\frac{\mu_j}{G'(\hat{x}_j)+\delta_j}\right| + \left|\frac{a(G(\hat{x}_j))\delta_j}{G'(\hat{x}_j)(G'(\hat{x}_j)+\delta_j)}\right| \le \frac{C_G\omega_a(h)}{m_{G'}-\omega_{G'}(h)} + \frac{\|a\|_\infty\omega_{G'}(h)}{m_{G'}(m_{G'}-\omega_{G'}(h))},\end{aligned} \tag{3.58}$$

which tends to 0 as $n\to\infty$. Thus, $\|Z_n\|\to 0$ as $n\to\infty$ by (3.1).

Now we consider the case where G has a finite number of points $\hat{x}$ where $G'(\hat{x}) = 0$. In this case, the previous argument does not work because $m_{G'} = 0$. However, we can still prove (3.56) in the following way. Let $\hat{x}^{(1)},\ldots,\hat{x}^{(s)}$ be the points where G' vanishes, and consider the balls (intervals) $B(\hat{x}^{(k)},\frac{1}{m}) = \{\hat{x}\in[0,1] : |\hat{x}-\hat{x}^{(k)}| < \frac{1}{m}\}$. The function G' is continuous and positive on the complement of the union $\bigcup_{k=1}^s B(\hat{x}^{(k)},\frac{1}{m})$, so

$$m_{G',m} = \min_{\hat{x}\in[0,1]\setminus\bigcup_{k=1}^s B(\hat{x}^{(k)},\frac{1}{m})} G'(\hat{x}) > 0.$$

For all indices $j = 1,\ldots,n$ such that $\hat{x}_j \in [0,1]\setminus\bigcup_{k=1}^s B(\hat{x}^{(k)},\frac{1}{m})$, the components in the jth row of the matrix (3.57) are bounded in modulus by a quantity that depends only on n,m,G,a and that converges to 0 as $n\to\infty$. This becomes immediately clear if we note that, for such indices j, the inequality (3.58) holds unchanged with $m_{G'}$ replaced by $m_{G',m}$. The number of remaining rows of Z_n (the rows corresponding to indices j such that $\hat{x}_j \in \bigcup_{k=1}^s B(\hat{x}^{(k)}, \frac{1}{m})$) is at most $2s(n+1)/m+s$. Indeed, each interval $B(\hat{x}^{(k)}, \frac{1}{m})$ has length $2/m$ (at most) and can contain at most $2(n+1)/m+1$ grid points $\hat{x}_j$. Thus, for every n,m we can split the matrix Z_n into the sum of two terms, i.e.,

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

where $N_{n,m}$ is obtained from Z_n by setting to zero all the rows corresponding to indices j such that $\hat{x}_j \in \bigcup_{k=1}^s B(\hat{x}^{(k)}, \frac{1}{m})$ and $R_{n,m} = Z_n - N_{n,m}$ is obtained from Z_n by setting to zero all the rows corresponding to indices j such that $\hat{x}_j \in [0,1]\backslash \bigcup_{k=1}^s B(\hat{x}^{(k)}, \frac{1}{m})$. From the above discussion we have

$$\lim_{n\to\infty} \|N_{n,m}\| = 0$$

for all m, and

$$\mathrm{rank}(R_{n,m}) \le \frac{2s(n+1)}{m} + s$$

for all m,n. In particular, for each m we can choose n_m such that, for $n \ge n_m$, $\mathrm{rank}(R_{n,m}) \le 3sn/m$ and $\|N_{n,m}\| \le 1/m$. The convergence (3.56) now follows from the definition of a.c.s. □

An increasing bijective map $G : [0,1] \to [0,1]$ in $C^1([0,1])$ is said to be regular if $G'(\hat{x}) \neq 0$ for all $\hat{x} \in [0,1]$ and is said to be singular otherwise, i.e., if $G'(\hat{x}) = 0$ for some $\hat{x} \in [0,1]$. If G is singular, any point $\hat{x} \in [0,1]$ such that $G'(\hat{x}) = 0$ is referred to as a singularity point (or simply a singularity) of G. The choice of a map G with one or more singularity points corresponds to adopting a local refinement strategy, according to which the grid points x_j rapidly accumulate at the G-images of the singularities as n increases. For example, if

$$G(\hat{x}) = \hat{x}^q, \qquad q > 1, \tag{3.59}$$

then 0 is a singularity of G (because $G'(0) = 0$) and the grid points

$$x_j = G(\hat{x}_j) = \Big(\frac{j}{n+1}\Big)^q, \qquad j = 0,\dots,n+1,$$

rapidly accumulate at $G(0) = 0$ as $n \to \infty$. We note that, whenever G is singular, the symbol in (3.48) is unbounded (except in some rare cases where $a(G(\hat{x}))$ and $G'(\hat{x})$ vanish simultaneously).

3.3 FE Discretization of Differential Equations

3.3.1 FE Discretization of Convection-Diffusion-Reaction Equations

Consider the following convection-diffusion-reaction problem in divergence form with Dirichlet boundary conditions:

$$\begin{cases} -(a(x)u'(x))' + b(x)u'(x) + c(x)u(x) = f(x), & x \in (0,1), \\ u(0) = u(1) = 0, \end{cases} \tag{3.60}$$

where $f \in L^2([0,1])$ and the coefficients a,b,c are only assumed to be in $L^\infty([0,1])$. These sole assumptions are enough to perform the GLT analysis of the matrices arising from the FE discretization of (3.60). In this sense, we are going to see that the theory of GLT sequences allows one to derive the singular value and spectral distribution of DE discretization matrices under very weak hypotheses on the DE coefficients.

FE discretization We consider the approximation of (3.60) by classical linear FEs on a uniform mesh in $[0,1]$ with stepsize $h = \frac{1}{n+1}$. We briefly describe here this approximation technique and for more details we refer the reader to [45, Chapter 4] or to any other good book on FEs. We first recall from [14, Chapter 8] that, if $\Omega \subset \mathbb{R}$ is a bounded interval whose endpoints are, say, α and β, $H^1(\Omega)$ denotes the (Sobolev) space of functions $v \in L^2(\Omega)$ possessing a weak (Sobolev) derivative in $L^2(\Omega)$. We also recall that each $v \in H^1(\Omega)$ coincides a.e. with a continuous function in $C(\overline{\Omega})$, and $H^1(\Omega)$ can also be defined as the following subspace of $C(\overline{\Omega})$:

$$H^1(\Omega) = \Big\{ v \in C(\overline{\Omega}) : \ v \text{ is differentiable a.e. with } v' \in L^2(\Omega), \quad v(x) = v(\alpha) + \int_\alpha^x v'(y)\mathrm{d}y \ \text{ for all } \ x \in \overline{\Omega} \Big\}.$$

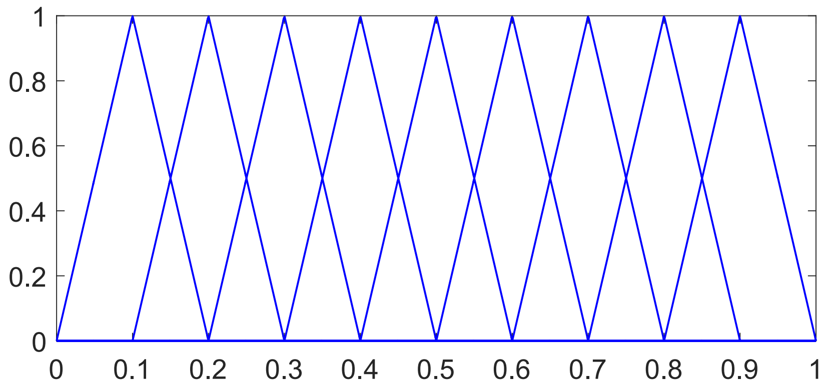

Figure 3.1: Graph of the hat-functions $\varphi_1,\ldots,\varphi_n$ for $n=9$.

In this definition, the weak derivative of a $v\in H^1(\Omega)$ is just the classical derivative v' (which exists a.e.). Let

$$H_0^1(\Omega)=\{v\in H^1(\Omega):\ v(\alpha)=v(\beta)=0\}.$$

The weak form of (3.60) reads as follows [14, Chapter 8]: find $u\in H_0^1([0,1])$ such that

$$\mathrm{a}(u,w)=\mathrm{f}(w),\qquad \forall w\in H_0^1([0,1]),$$

where

$$\mathrm{a}(u,w)=\int_0^1 a(x)u'(x)w'(x)\mathrm{d}x+\int_0^1 b(x)u'(x)w(x)\mathrm{d}x+\int_0^1 c(x)u(x)w(x)\mathrm{d}x,$$
$$\mathrm{f}(w)=\int_0^1 f(x)w(x)\mathrm{d}x.$$

Let $h=\frac{1}{n+1}$ and $x_i=ih,\ i=0,\ldots,n+1$. In the linear FE approach based on the uniform mesh $\{x_0,\ldots,x_{n+1}\}$, we fix the subspace $\mathscr{W}_n=\mathrm{span}(\varphi_1,\ldots,\varphi_n)\subset H_0^1([0,1])$, where $\varphi_1,\ldots,\varphi_n$ are the so-called hat-functions:

$$\varphi_i(x)=\frac{x-x_{i-1}}{x_i-x_{i-1}}\chi_{[x_{i-1},x_i)}(x)+\frac{x_{i+1}-x}{x_{i+1}-x_i}\chi_{[x_i,x_{i+1})}(x),\qquad i=1,\ldots,n;\tag{3.61}$$

see Figure 3.1. Note that $\mathscr{W}_n$ is the space of piecewise linear functions corresponding to the sequence of points $0=x_0<x_1<\ldots<x_{n+1}=1$ and vanishing on the boundary of the domain $[0,1]$. In formulas,

$$\mathscr{W}_n=\left\{s:[0,1]\to\mathbb{R}:\ s_{|[\frac{i}{n+1},\frac{i+1}{n+1})}\in\mathbb{P}_1,\ \ i=0,\ldots,n,\ \ s(0)=s(1)=0\right\},$$

where $\mathbb{P}_1$ is the space of polynomials of degree less than or equal to 1. We look for an approximation $u_{\mathscr{W}_n}$ of u by solving the following (Galerkin) problem: find $u_{\mathscr{W}_n}\in\mathscr{W}_n$ such that

$$\mathrm{a}(u_{\mathscr{W}_n},w)=\mathrm{f}(w),\qquad \forall w\in\mathscr{W}_n.$$

Since $\{\varphi_1,\ldots,\varphi_n\}$ is a basis of $\mathscr{W}_n$, we can write $u_{\mathscr{W}_n}=\sum_{j=1}^n u_j\varphi_j$ for a unique vector $\mathbf{u}=(u_1,\ldots,u_n)^T$. By linearity, the computation of $u_{\mathscr{W}_n}$ (i.e., of $\mathbf{u}$) reduces to solving the linear system

$$A_n\mathbf{u}=\mathbf{f},$$

where $\mathbf{f}=(\mathrm{f}(\varphi_1),\ldots,\mathrm{f}(\varphi_n))^T$ and A_n is the stiffness matrix,

$$A_n=[\mathrm{a}(\varphi_j,\varphi_i)]_{i,j=1}^n.$$

Note that A_n admits the following decomposition:

$$A_n=K_n+Z_n,\tag{3.62}$$

where

$$K_n = \left[\int_0^1 a(x)\varphi_j'(x)\varphi_i'(x)\mathrm{d}x \right]_{i,j=1}^n \tag{3.63}$$

is the (symmetric) diffusion matrix and

$$Z_n = \left[\int_0^1 b(x)\varphi_j'(x)\varphi_i(x)\mathrm{d}x \right]_{i,j=1}^n + \left[\int_0^1 c(x)\varphi_j(x)\varphi_i(x)\mathrm{d}x \right]_{i,j=1}^n \tag{3.64}$$

is the sum of the convection and reaction matrix.

GLT analysis of the FE discretization matrices Using the theory of GLT sequences we now derive the spectral and singular value distribution of the sequence of normalized stiffness matrices $\{\frac{1}{n+1}A_n\}_n$.

Theorem 3.10. *If $a,b,c \in L^\infty([0,1])$ then*

$$\left\{ \frac{1}{n+1}A_n \right\}_n \sim_{\mathrm{GLT}} a(x)(2-2\cos\theta) \tag{3.65}$$

and

$$\left\{ \frac{1}{n+1}A_n \right\}_n \sim_{\sigma,\lambda} a(x)(2-2\cos\theta). \tag{3.66}$$

Proof. The proof consists of the following steps. Throughout the proof, the letter C will denote a generic constant independent of n.

Step 1. We show that

$$\left\| \frac{1}{n+1}K_n \right\| \le C \tag{3.67}$$

and

$$\left\| \frac{1}{n+1}Z_n \right\| \le C/n. \tag{3.68}$$

To prove (3.67), we note that K_n is a banded (tridiagonal) matrix, due to the local support property $\mathrm{supp}(\varphi_i) = [x_{i-1}, x_{i+1}]$, $i = 1,\dots,n$. Moreover, by the inequality $|\varphi_i'(x)| \le n+1$, for all $i,j = 1,\dots,n$ we have

$$|(K_n)_{ij}| = \left| \int_0^1 a(x)\varphi_j'(x)\varphi_i'(x)\mathrm{d}x \right| = \left| \int_{x_{i-1}}^{x_{i+1}} a(x)\varphi_j'(x)\varphi_i'(x)\mathrm{d}x \right| \le (n+1)^2 \|a\|_{L^\infty} \int_{x_{i-1}}^{x_{i+1}} \mathrm{d}x = 2(n+1)\|a\|_{L^\infty}.$$

Thus, the components of the tridiagonal matrix $\frac{1}{n+1}K_n$ are bounded (in modulus) by $2\|a\|_{L^\infty}$, and (3.67) follows from (3.1).

To prove (3.68), we follow the same argument as for the proof of (3.67). Due to the local support property of the hat-functions, Z_n is tridiagonal. Moreover, by the inequalities $|\varphi_i(x)| \le 1$ and $|\varphi_i'(x)| \le n+1$, for all $i,j = 1,\dots,n$ we have

$$|(Z_n)_{ij}| = \left| \int_{x_{i-1}}^{x_{i+1}} b(x)\varphi_j'(x)\varphi_i(x)\mathrm{d}x + \int_{x_{i-1}}^{x_{i+1}} c(x)\varphi_j(x)\varphi_i(x)\mathrm{d}x \right| \le 2\|b\|_{L^\infty} + \frac{2\|c\|_{L^\infty}}{n+1},$$

and (3.68) follows from (3.1).

Step 2. Consider the linear operator $K_n(\cdot) : L^1([0,1]) \to \mathbb{R}^{n\times n}$,

$$K_n(g) = \left[\int_0^1 g(x)\varphi_j'(x)\varphi_i'(x)\mathrm{d}x \right]_{i,j=1}^n.$$

By (3.63), we have $K_n = K_n(a)$. The next three steps are devoted to show that

$$\left\{ \frac{1}{n+1}K_n(g) \right\}_n \sim_{\mathrm{GLT}} g(x)(2-2\cos\theta), \qquad \forall g \in L^1([0,1]). \tag{3.69}$$

Once this is done, the theorem is proved. Indeed, by applying (3.69) with $g = a$ we immediately get $\{\frac{1}{n+1}K_n\}_n \sim_{\mathrm{GLT}} a(x)(2-2\cos\theta)$. Since $\{\frac{1}{n+1}Z_n\}_n$ is zero-distributed by Step 1, (3.65) follows from the decomposition

$$\frac{1}{n+1}A_n = \frac{1}{n+1}K_n + \frac{1}{n+1}Z_n \tag{3.70}$$

and from **GLT 3 – GLT 4**; and the singular value distribution in (3.66) follows from **GLT 1**. If $b(x)=0$ identically, then $\frac{1}{n+1}A_n$ is symmetric and also the spectral distribution in (3.66) follows from **GLT 1**. If $b(x)$ is not identically 0, the spectral distribution in (3.66) follows from **GLT 2** applied to the decomposition (3.70), taking into account what we have proved in Step 1.

Step 3. We first prove (3.69) in the constant-coefficient case where $g=1$ identically. In this case, a direct computation based on (3.61) shows that

$$K_n(1)=\left[\int_0^1 \varphi_j'(x)\varphi_i'(x)\mathrm{d}x\right]_{i,j=1}^n=\frac{1}{h}\begin{bmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 2 \end{bmatrix}=\frac{1}{h}T_n(2-2\cos\theta),$$

and the desired relation $\{\frac{1}{n+1}K_n(1)\}_n\sim_{\rm GLT}2-2\cos\theta$ follows from **GLT 3**. Note that it is precisely the analysis of the constant-coefficient case considered in this step that allows one to realize what is the correct normalization factor. In our case, this is $\frac{1}{n+1}$, which removes the $\frac{1}{h}$ from $K_n(1)$ and yields a normalized matrix $\frac{1}{n+1}K_n(1)=T_n(2-2\cos\theta)$, whose components are *bounded away from* 0 *and* ∞ (actually, in the present case they are even constant).

Step 4. Now we prove (3.69) in the case where $g\in C([0,1])$. We first illustrate the idea, and then we go into the details. The proof is based on the fact that the hat-functions (3.61) are 'locally supported'. Indeed, the support $[x_{i-1},x_{i+1}]$ of the ith hat-function $\varphi_i(x)$ is located near the point $\frac{i}{n}\in[x_i,x_{i+1}]$, and the amplitude of the support tends to 0 as $n\to\infty$. In this sense, the linear FE method considered herein belongs to the family of the so-called 'local' methods. Since $g(x)$ varies continuously over $[0,1]$, the (i,j) entry of $K_n(g)$ can be approximated as follows, for every $i,j=1,\ldots,n$:

$$\begin{aligned}(K_n(g))_{ij}&=\int_0^1 g(x)\varphi_j'(x)\varphi_i'(x)\mathrm{d}x=\int_{x_{i-1}}^{x_{i+1}} g(x)\varphi_j'(x)\varphi_i'(x)\mathrm{d}x\\ &\approx g\Big(\frac{i}{n}\Big)\int_{x_{i-1}}^{x_{i+1}}\varphi_j'(x)\varphi_i'(x)\mathrm{d}x=g\Big(\frac{i}{n}\Big)\int_0^1\varphi_j'(x)\varphi_i'(x)\mathrm{d}x=g\Big(\frac{i}{n}\Big)(K_n(1))_{ij}.\end{aligned}$$

This approximation can be rewritten in matrix form as

$$K_n(g)\approx D_n(g)K_n(1). \tag{3.71}$$

We will see that (3.71) implies that $\{\frac{1}{n+1}K_n(g)-\frac{1}{n+1}D_n(g)K_n(1)\}_n\sim_\sigma 0$, and (3.69) will then follow from Step 3 and **GLT 3 – GLT 4**.

Let us now go into the details. Since $\mathrm{supp}(\varphi_i)=[x_{i-1},x_{i+1}]$ and $|\varphi_i'(x)|\le n+1$, for all $i,j=1,\ldots,n$ we have

$$\begin{aligned}\big|(K_n(g))_{ij}-(D_n(g)K_n(1))_{ij}\big|&=\left|\int_0^1\Big[g(x)-g\Big(\frac{i}{n}\Big)\Big]\varphi_j'(x)\varphi_i'(x)\mathrm{d}x\right|\\ &\le(n+1)^2\int_{x_{i-1}}^{x_{i+1}}\Big|g(x)-g\Big(\frac{i}{n}\Big)\Big|\mathrm{d}x\le 2(n+1)\,\omega_g\Big(\frac{2}{n+1}\Big).\end{aligned}$$

It follows that each entry of the matrix $Z_n=\frac{1}{n+1}K_n(g)-\frac{1}{n+1}D_n(g)K_n(1)$ is bounded in modulus by $2\,\omega_g(\frac{2}{n+1})$. Moreover, Z_n is banded (tridiagonal), because of the local support property of the hat-functions. Thus, both the 1-norm and the ∞-norm of Z_n are bounded by $C\,\omega_g(\frac{2}{n+1})$, and (3.1) yields $\|Z_n\|\le C\,\omega_g(\frac{2}{n+1})\to 0$ as $n\to\infty$. Hence, $\{Z_n\}_n\sim_\sigma 0$, which implies (3.69) by Step 3 and **GLT 3 – GLT 4**.

Step 5. Finally, we prove (3.69) in the general case where $g\in L^1([0,1])$. By the density of $C([0,1])$ in $L^1([0,1])$, there exist continuous functions $g_m\in C([0,1])$ such that $g_m\to g$ in $L^1([0,1])$. By Step 4,

$$\Big\{\frac{1}{n+1}K_n(g_m)\Big\}_n\sim_{\rm GLT} g_m(x)(2-2\cos\theta). \tag{3.72}$$

Moreover,

$$g_m(x)(2-2\cos\theta)\to g(x)(2-2\cos\theta)\ \text{ in measure.} \tag{3.73}$$

We show that

$$\Big\{\frac{1}{n+1}K_n(g_m)\Big\}_n\xrightarrow{\text{a.c.s.}}\Big\{\frac{1}{n+1}K_n(g)\Big\}_n. \tag{3.74}$$

Since $\sum_{i=1}^n |\varphi_i'(x)| \le 2(n+1)$ for all $x \in [0,1]$, by (3.4) we obtain

$$\begin{aligned}\|K_n(g)-K_n(g_m)\|_1 &\le \sum_{i,j=1}^n |(K_n(g))_{ij}-(K_n(g_m))_{ij}| = \sum_{i,j=1}^n \left|\int_0^1 \big[g(x)-g_m(x)\big]\,\varphi_j'(x)\varphi_i'(x)\mathrm{d}x\right| \\ &\le \int_0^1 |g(x)-g_m(x)| \sum_{i,j=1}^n |\varphi_j'(x)|\,|\varphi_i'(x)|\mathrm{d}x \le 4(n+1)^2\|g-g_m\|_{L^1}\end{aligned}$$

and

$$\left\|\frac{1}{n+1}K_n(g)-\frac{1}{n+1}K_n(g_m)\right\|_1 \le Cn\|g-g_m\|_{L^1}.$$

Thus, $\{\frac{1}{n+1}K_n(g_m)\}_n \xrightarrow{\text{a.c.s.}} \{\frac{1}{n+1}K_n(g)\}_n$ by **ACS 3**. In view of (3.72)–(3.74), the relation (3.69) follows from **GLT 7**. □

Remark 3.6 (formal structure of the symbol). Problem (3.60) can be formally rewritten as follows:

$$\begin{cases} -a(x)u''(x)+(b(x)-a'(x))u'(x)+c(x)u(x)=f(x), & x\in(0,1),\\ u(0)=u(1)=0. \end{cases} \tag{3.75}$$

It is then clear that the symbol $a(x)(2-2\cos\theta)$ has the same formal structure of the higher-order differential operator $-a(x)u''(x)$ associated with (3.75) (as in the FD case; see Remark 3.1). The formal analogy becomes even more evident if we note that $2-2\cos\theta$ is the trigonometric polynomial in the Fourier variable coming from the FE discretization of the (negative) second derivative $-u''(x)$. Indeed, as we have seen in Step 3 of the proof of Theorem 3.10, $2-2\cos\theta$ is the symbol of the sequence of FE diffusion matrices $\{\frac{1}{n+1}K_n(1)\}_n$, which arises from the FE approximation of the Poisson problem

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

that is, problem (3.60) in the case where $a(x)=1$ and $b(x)=c(x)=0$ identically.

3.3.2 FE Discretization of a System of Equations

In this section we consider the linear FE approximation of a system of differential equations, namely

$$\begin{cases} -(a(x)u'(x))'+v'(x)=f(x), & x\in(0,1),\\ \qquad -u'(x)-\rho v(x)=g(x), & x\in(0,1),\\ u(0)=0, \quad u(1)=0,\\ v(0)=0, \quad v(1)=0, \end{cases} \tag{3.76}$$

where ρ is a constant and a is only assumed to be in $L^1([0,1])$. As we shall see, the resulting discretization matrices appear in the so-called saddle point form [6, p. 3], and we will illustrate the way to compute the asymptotic spectral and singular value distribution of their Schur complements using the theory of GLT sequences. It is worth noting that the Schur complement is a key tool for the numerical treatment of the related linear systems [6, Section 5]. The analysis of this section is similar to the analysis in [25, Section 2], but the discretization technique considered herein is a pure FE approximation, whereas in [25, Section 2] the authors adopted a mixed FD/FE technique.

FE discretization We consider the approximation of (3.76) by linear FEs on a uniform mesh in $[0,1]$ with stepsize $h=\frac{1}{n+1}$. Let us describe it shortly. The weak form of (3.76) reads as follows:[1] find $u,v\in H_0^1([0,1])$ such that, for all $w\in H_0^1([0,1])$,

$$\begin{cases} \int_0^1 a(x)u'(x)w'(x)\mathrm{d}x+\int_0^1 v'(x)w(x)\mathrm{d}x=\int_0^1 f(x)w(x)\mathrm{d}x,\\ -\int_0^1 u'(x)w(x)\mathrm{d}x-\rho\int_0^1 v(x)w(x)\mathrm{d}x=\int_0^1 g(x)w(x)\mathrm{d}x. \end{cases} \tag{3.77}$$

Let $h=\frac{1}{n+1}$ and $x_i=ih,\ i=0,\ldots,n+1$. In the linear FE approach based on the mesh $\{x_0,\ldots,x_{n+1}\}$, we fix the subspace $\mathscr{W}_n=\mathrm{span}(\varphi_1,\ldots,\varphi_n)\subset H_0^1([0,1])$, where $\varphi_1,\ldots,\varphi_n$ are the hat-functions in (3.61) (see also Figure 3.1). Then, we look

[1] We are proceeding formally here, because the assumption $a\in L^1([0,1])$ is too weak to ensure that the weak form (3.77) is well-defined. Keep in mind, however, that our formal derivation is correct if $a\in L^\infty([0,1])$.

for approximations $u_{\mathscr{W}_n}, v_{\mathscr{W}_n}$ of u, v by solving the following (Galerkin) problem: find $u_{\mathscr{W}_n}, v_{\mathscr{W}_n} \in \mathscr{W}_n$ such that, for all $w \in \mathscr{W}_n$,

$$\begin{cases} \int_0^1 a(x)u'_{\mathscr{W}_n}(x)w'(x)\mathrm{d}x + \int_0^1 v'_{\mathscr{W}_n}(x)w(x)\mathrm{d}x = \int_0^1 f(x)w(x)\mathrm{d}x, \\ -\int_0^1 u'_{\mathscr{W}_n}(x)w(x)\mathrm{d}x - \rho\int_0^1 v_{\mathscr{W}_n}(x)w(x)\mathrm{d}x = \int_0^1 g(x)w(x)\mathrm{d}x. \end{cases}$$

Since $\{\varphi_1,\ldots,\varphi_n\}$ is a basis of $\mathscr{W}_n$, we can write $u_{\mathscr{W}_n} = \sum_{j=1}^n u_j\varphi_j$ and $v_{\mathscr{W}_n} = \sum_{j=1}^n v_j\varphi_j$ for unique vectors $\mathbf{u} = (u_1,\ldots,u_n)^T$ and $\mathbf{v} = (v_1,\ldots,v_n)^T$. By linearity, the computation of $u_{\mathscr{W}_n}, v_{\mathscr{W}_n}$ (i.e., of $\mathbf{u},\mathbf{v}$) reduces to solving the linear system

$$A_{2n}\begin{bmatrix}\mathbf{u}\\ \mathbf{v}\end{bmatrix} = \begin{bmatrix}\mathbf{f}\\ \mathbf{g}\end{bmatrix},$$

where $\mathbf{f} = \left[\int_0^1 f(x)\varphi_i(x)\mathrm{d}x\right]_{i=1}^n$, $\mathbf{g} = \left[\int_0^1 g(x)\varphi_i(x)\mathrm{d}x\right]_{i=1}^n$ and A_{2n} is the stiffness matrix, which possesses the following saddle point structure:

$$A_{2n} = \begin{bmatrix} K_n & H_n \\ H_n^T & -\rho M_n \end{bmatrix}.$$

Here, the blocks K_n, H_n, M_n are square matrices of size n, and precisely

$$K_n = \left[\int_0^1 a(x)\varphi_j'(x)\varphi_i'(x)\mathrm{d}x\right]_{i,j=1}^n,$$

$$H_n = \left[\int_0^1 \varphi_j'(x)\varphi_i(x)\mathrm{d}x\right]_{i,j=1}^n = \frac{1}{2}\begin{bmatrix} 0 & 1 & & & \\ -1 & 0 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 0 & 1 \\ & & & -1 & 0 \end{bmatrix} = -\mathrm{i}\,T_n(\sin\theta),$$

$$M_n = \left[\int_0^1 \varphi_j(x)\varphi_i(x)\mathrm{d}x\right]_{i,j=1}^n = \frac{h}{6}\begin{bmatrix} 4 & 1 & & & \\ 1 & 4 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & 4 & 1 \\ & & & 1 & 4 \end{bmatrix} = \frac{h}{3}T_n(2+\cos\theta).$$

Note that K_n is exactly the matrix appearing in (3.63). Note also that the matrices K_n, M_n are symmetric, while H_n is skew-symmetric: $H_n^T = -H_n = \mathrm{i}\,T_n(\sin\theta)$.

GLT analysis of the Schur complements of the FE discretization matrices Assume that the matrices K_n are invertible. This is satisfied, for example, if $a > 0$ a.e., in which case the matrices K_n are positive definite. The (negative) Schur complement of A_{2n} is the symmetric matrix given by

$$S_n = \rho M_n + H_n^T K_n^{-1} H_n = \frac{\rho h}{3}T_n(2+\cos\theta) + T_n(\sin\theta)K_n^{-1}T_n(\sin\theta). \tag{3.78}$$

In the following, we perform the GLT analysis of the sequence of normalized Schur complements $\{(n+1)S_n\}_n$, and we compute its asymptotic spectral and singular value distribution under the additional necessary assumption that $a \neq 0$ a.e.

Theorem 3.11. *Let $\rho \in \mathbb{R}$ and $a \in L^1([0,1])$. Suppose that the matrices K_n are invertible and that $a \neq 0$ a.e. Then*

$$\{(n+1)S_n\}_n \sim_{\mathrm{GLT}} \varsigma(x,\theta) \tag{3.79}$$

and

$$\{(n+1)S_n\}_n \sim_{\sigma,\lambda} \varsigma(x,\theta), \tag{3.80}$$

where

$$\varsigma(x,\theta) = \frac{\rho}{3}(2+\cos\theta) + \frac{\sin^2\theta}{a(x)(2-2\cos\theta)}.$$

Proof. In view of (3.78), we have

$$(n+1)S_n = \frac{\rho}{3}\,T_n(2+\cos\theta) + T_n(\sin\theta)\left(\frac{1}{n+1}K_n\right)^{-1} T_n(\sin\theta).$$

Moreover, by (3.69),

$$\left\{\frac{1}{n+1}K_n\right\}_n = \left\{\frac{1}{n+1}K_n(a)\right\}_n \sim_{\rm GLT} a(x)(2-2\cos\theta).$$

Therefore, under the assumption that $a \neq 0$ a.e., the GLT relation (3.79) follows from **GLT 3 – GLT 5**. The singular value and spectral distributions in (3.80) follow from (3.79) and **GLT 1** as the Schur complements S_n are symmetric. □

3.4 IgA Discretization of Differential Equations

Isogeometric Analysis (IgA) is a modern and successful paradigm introduced in [17, 39] for analyzing problems governed by DEs. Its goal is to improve the connection between numerical simulation and Computer-Aided Design (CAD) systems. The main idea in IgA is to use directly the geometry provided by CAD systems and to approximate the solutions of DEs by the same type of functions (usually, B-splines or NURBS). In this way, it is possible to save about 80% of the CPU time, which is normally employed in the translation between two different languages (e.g., between FEs and CAD or between FDs and CAD). In its original formulation [17, 39], IgA employs Galerkin discretizations, which are typical of the FE approach. In the Galerkin framework an efficient implementation requires special numerical quadrature rules when constructing the resulting system of equations; see, e.g., [42]. To avoid this issue, isogeometric collocation methods have been recently introduced in [1]. Detailed comparisons with IgA Galerkin have shown the advantages of IgA collocation in terms of accuracy versus computational cost, in particular when higher-order approximation degrees are adopted [49]. Within the framework of IgA collocation, many applications have been successfully tackled, showing its potential and flexibility. Interested readers are referred to the recent review [47] and references therein. Section 3.4.1 is devoted to the isogeometric collocation approach, whereas the more traditional isogoemetric Galerkin methods will be addressed in Sections 3.4.2–3.4.3.

3.4.1 B-Spline IgA Collocation Discretization of Convection-Diffusion-Reaction Equations

Consider the convection-diffusion-reaction problem

$$\begin{cases} -(a(x)u'(x))' + b(x)u'(x) + c(x)u(x) = f(x), & x \in \Omega, \\ u(x) = 0, & x \in \partial\Omega, \end{cases} \tag{3.81}$$

where Ω is a bounded open interval of $\mathbb{R}$, $a : \overline{\Omega} \to \mathbb{R}$ is a function in $C^1(\overline{\Omega})$ and $b, c, f : \overline{\Omega} \to \mathbb{R}$ are functions in $C(\overline{\Omega})$. We consider the isogeometric collocation approximation of (3.81) based on uniform B-splines of degree $p \geq 2$. Since this approximation technique is not as known as FDs or FEs, we describe it below in some detail. For more on IgA collocation methods, see [1, 47].

Isogeometric collocation approximation Problem (3.81) can be reformulated as follows:

$$\begin{cases} -a(x)u''(x) + s(x)u'(x) + c(x)u(x) = f(x), & x \in \Omega, \\ u(x) = 0, & x \in \partial\Omega, \end{cases} \tag{3.82}$$

where $s(x) = b(x) - a'(x)$. In the standard collocation method, we choose a finite dimensional vector space $\mathscr{W}$, consisting of sufficiently smooth functions defined on $\overline{\Omega}$ and vanishing on the boundary $\partial\Omega$; we call $\mathscr{W}$ the approximation space. Then, we introduce a set of $N = \dim\mathscr{W}$ collocation points $\{\tau_1, \ldots, \tau_N\} \subset \Omega$ and we look for a function $u_{\mathscr{W}} \in \mathscr{W}$ satisfying the differential equation (3.82) at the points τ_i, i.e.,

$$-a(\tau_i)u''_{\mathscr{W}}(\tau_i) + s(\tau_i)u'_{\mathscr{W}}(\tau_i) + c(\tau_i)u_{\mathscr{W}}(\tau_i) = f(\tau_i), \qquad i = 1, \ldots, N.$$

The function $u_{\mathscr{W}}$ is taken as an approximation to the solution u of (3.82). If $\{\varphi_1, \ldots, \varphi_N\}$ is a basis of $\mathscr{W}$, then we have $u_{\mathscr{W}} = \sum_{j=1}^N u_j\,\varphi_j$ for a unique vector $\mathbf{u} = (u_1, \ldots, u_N)^T$, and, by linearity, the computation of $u_{\mathscr{W}}$ (i.e., of $\mathbf{u}$) reduces to solving the linear system

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

where $\mathbf{f} = \left[f(\tau_i)\right]_{i=1}^N$ and

$$\begin{aligned} A &= \left[-a(\tau_i)\varphi_j''(\tau_i) + s(\tau_i)\varphi_j'(\tau_i) + c(\tau_i)\varphi_j(\tau_i)\right]_{i,j=1}^N \\ &= \Big(\operatorname*{diag}_{i=1,\ldots,N} a(\tau_i)\Big)\left[-\varphi_j''(\tau_i)\right]_{i,j=1}^N + \Big(\operatorname*{diag}_{i=1,\ldots,N} s(\tau_i)\Big)\left[\varphi_j'(\tau_i)\right]_{i,j=1}^N + \Big(\operatorname*{diag}_{i=1,\ldots,N} c(\tau_i)\Big)\left[\varphi_j(\tau_i)\right]_{i,j=1}^N \end{aligned} \tag{3.83}$$

is the collocation matrix.

Now, suppose that the physical domain Ω can be described by a global geometry function $G : [0,1] \to \overline{\Omega}$, which is invertible and satisfies $G(\partial([0,1])) = \partial\overline{\Omega}$. Let

$$\{\hat{\varphi}_1, \ldots, \hat{\varphi}_N\} \tag{3.84}$$

be a set of basis functions defined on the parametric (or reference) domain $[0,1]$ and vanishing on the boundary $\partial([0,1])$. Let

$$\{\hat{\tau}_1, \ldots, \hat{\tau}_N\} \tag{3.85}$$

be a set of N collocation points in $(0,1)$. In the isogeometric collocation approach, we find an approximation $u_{\mathscr{W}}$ of u by using the standard collocation method described above, in which

- the approximation space is chosen as $\mathscr{W} = \mathrm{span}(\varphi_1, \ldots \varphi_N)$, with

$$\varphi_i(x) = \hat{\varphi}_i(G^{-1}(x)) = \hat{\varphi}_i(\hat{x}), \qquad x = G(\hat{x}), \qquad i = 1, \ldots, N, \tag{3.86}$$

- the collocation points in the physical domain Ω are defined as

$$\tau_i = G(\hat{\tau}_i), \qquad i = 1, \ldots, N. \tag{3.87}$$

The resulting collocation matrix A is given by (3.83), with the basis functions φ_i and the collocation points τ_i defined as in (3.86)–(3.87).

Assuming that G and $\hat{\varphi}_i$, $i = 1, \ldots, N$, are sufficiently regular, we can apply standard differential calculus to express A in terms of G and $\hat{\varphi}_i$, $\hat{\tau}_i$, $i = 1, \ldots, N$. Let us work out this expression. For any $u : \overline{\Omega} \to \mathbb{R}$, consider the corresponding function $\hat{u} : [0,1] \to \mathbb{R}$, which is defined on the parametric domain by

$$\hat{u}(\hat{x}) = u(x), \qquad x = G(\hat{x}). \tag{3.88}$$

In other words, $\hat{u}(\hat{x}) = u(G(\hat{x}))$.[2] Then, u satisfies (3.82) if and only if $\hat{u}$ satisfies the corresponding transformed problem

$$\begin{cases} -a_G(\hat{x})\hat{u}''(\hat{x}) + s_G(\hat{x})\hat{u}'(\hat{x}) + c_G(\hat{x})\hat{u}(\hat{x}) = f(G(\hat{x})), & \hat{x} \in (0,1), \\ \hat{u}(\hat{x}) = 0, & \hat{x} \in \partial((0,1)), \end{cases} \tag{3.89}$$

where a_G, s_G, c_G are, respectively, the transformed diffusion, convection, reaction coefficient. They are given by

$$a_G(\hat{x}) = \frac{a(G(\hat{x}))}{(G'(\hat{x}))^2}, \tag{3.90}$$

$$s_G(\hat{x}) = \frac{a(G(\hat{x}))G''(\hat{x})}{(G'(\hat{x}))^3} + \frac{s(G(\hat{x}))}{G'(\hat{x})}, \tag{3.91}$$

$$c_G(\hat{x}) = c(G(\hat{x})), \tag{3.92}$$

for $\hat{x} \in [0,1]$. The collocation matrix A in (3.83) can be expressed in terms of G and $\hat{\varphi}_i$, $\hat{\tau}_i$, $i = 1, \ldots, N$, as follows:

$$\begin{aligned} A &= \left[-a_G(\hat{\tau}_i)\hat{\varphi}_j''(\hat{\tau}_i) + s_G(\hat{\tau}_i)\hat{\varphi}_j'(\hat{\tau}_i) + c_G(\hat{\tau}_i)\hat{\varphi}_j(\hat{\tau}_i)\right]_{i,j=1}^N \\ &= \Big(\operatorname*{diag}_{i=1,\ldots,N} a_G(\hat{\tau}_i)\Big)\left[-\hat{\varphi}_j''(\hat{\tau}_i)\right]_{i,j=1}^N + \Big(\operatorname*{diag}_{i=1,\ldots,N} s_G(\hat{\tau}_i)\Big)\left[\hat{\varphi}_j'(\hat{\tau}_i)\right]_{i,j=1}^N + \Big(\operatorname*{diag}_{i=1,\ldots,N} c_G(\hat{\tau}_i)\Big)\left[\hat{\varphi}_j(\hat{\tau}_i)\right]_{i,j=1}^N. \end{aligned} \tag{3.93}$$

In the IgA context, the geometry map G is expressed in terms of the functions $\hat{\varphi}_i$, in accordance with the isoparametric approach [17, Section 3.1]. Moreover, the functions $\hat{\varphi}_i$ themselves are usually B-splines or their rational versions, the so-called NURBS. In this section, the role of the $\hat{\varphi}_i$ will be played by B-splines over uniform knot sequences. Furthermore, we do not limit ourselves to the isoparametric approach, but we allow the geometry map G to be any sufficiently regular function from $[0,1]$ to $\overline{\Omega}$, not necessarily expressed in terms of B-splines. Finally, following [1], the collocation points $\hat{\tau}_i$ will be chosen as the Greville abscissae corresponding to the B-splines $\hat{\varphi}_i$.

[2] Note that $\hat{\varphi}_i(\hat{x}) = \varphi_i(G(\hat{x}))$ for $i = 1, \ldots, N$, so $\hat{\varphi}_1, \ldots, \hat{\varphi}_N$ are obtained from $\varphi_1, \ldots, \varphi_N$ by the rule (3.88). Moreover, the equation $\tau_i = G(\hat{\tau}_i)$ is the same as the relation $x = G(\hat{x})$ in (3.88).

B-splines and Greville abscissae For $p,n \geq 1$, consider the uniform knot sequence

$$t_1 = \cdots = t_{p+1} = 0 < t_{p+2} < \cdots < t_{p+n} < 1 = t_{p+n+1} = \cdots = t_{2p+n+1}, \tag{3.94}$$

where

$$t_{i+p+1} = \frac{i}{n}, \qquad i = 0,\ldots,n. \tag{3.95}$$

The B-splines of degree p on this knot sequence are denoted by

$$N_{i,[p]} : [0,1] \to \mathbb{R}, \qquad i = 1,\ldots,n+p, \tag{3.96}$$

and are defined recursively as follows [19]: for $1 \leq i \leq n+2p$,

$$N_{i,[0]}(t) = \chi_{[t_i,t_{i+1})}(t), \qquad t \in [0,1]; \tag{3.97}$$

for $1 \leq k \leq p$ and $1 \leq i \leq n+2p-k$,

$$N_{i,[k]}(t) = \frac{t-t_i}{t_{i+k}-t_i} N_{i,[k-1]}(t) + \frac{t_{i+k+1}-t}{t_{i+k+1}-t_{i+1}} N_{i+1,[k-1]}(t), \qquad t \in [0,1], \tag{3.98}$$

where we assume that a fraction with zero denominator is zero. The Greville abscissa $\xi_{i,[p]}$ associated with the B-spline $N_{i,[p]}$ is defined by

$$\xi_{i,[p]} = \frac{t_{i+1}+t_{i+2}+\ldots+t_{i+p}}{p}, \qquad i = 1,\ldots,n+p. \tag{3.99}$$

We know from [19] that the functions $N_{1,[p]},\ldots,N_{n+p,[p]}$ belong to $C^{p-1}([0,1])$ and form a basis for the spline space

$$\Big\{ s \in C^{p-1}([0,1]) : \ s_{|[\frac{i}{n},\frac{i+1}{n})} \in \mathbb{P}_p, \ i = 0,\ldots,n-1 \Big\},$$

where $\mathbb{P}_p$ is the space of polynomials of degree less than or equal to p. Moreover, $N_{1,[p]},\ldots,N_{n+p,[p]}$ possess the following properties [19].

- Local support property:

$$\mathrm{supp}(N_{i,[p]}) = [t_i, t_{i+p+1}], \qquad i = 1,\ldots,n+p. \tag{3.100}$$

- Vanishment on the boundary:

$$N_{i,[p]}(0) = N_{i,[p]}(1) = 0, \qquad i = 2,\ldots,n+p-1. \tag{3.101}$$

- Nonnegative partition of unity:

$$N_{i,[p]}(t) \geq 0, \qquad t \in [0,1], \qquad i = 1,\ldots,n+p, \tag{3.102}$$

$$\sum_{i=1}^{n+p} N_{i,[p]}(t) = 1, \qquad t \in [0,1]. \tag{3.103}$$

- Bounds for derivatives:

$$\sum_{i=1}^{n+p} |N'_{i,[p]}(t)| \leq 2pn, \qquad t \in [0,1], \tag{3.104}$$

$$\sum_{i=1}^{n+p} |N''_{i,[p]}(t)| \leq 4p(p-1)n^2, \qquad t \in [0,1]. \tag{3.105}$$

Note that the derivatives $N'_{1,[p]}(t),\ldots,N'_{n+p,[p]}(t)$ (resp., $N''_{1,[p]}(t),\ldots,N''_{n+p,[p]}(t)$) may not be defined at some of the points $\frac{1}{n},\ldots,\frac{n-1}{n}$ when $p = 1$ (resp., $p = 1,2$). In the summations (3.104)–(3.105), it is understood that the undefined values are counted as 0.

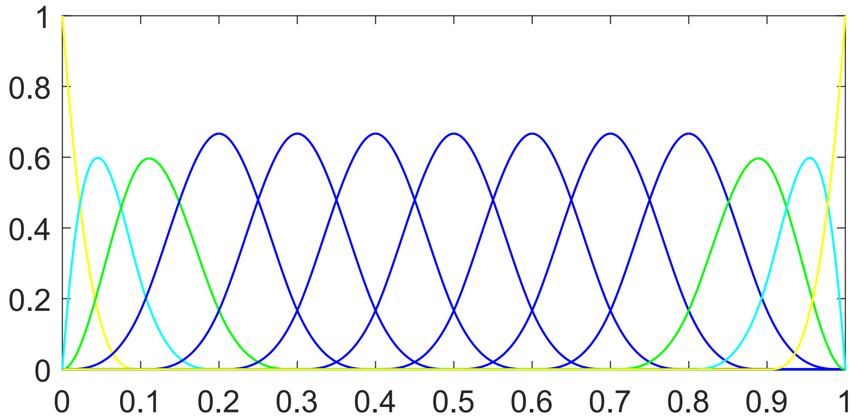

Figure 3.2: Graph of the B-splines $N_{i,[p]}$, $i = 1,\dots,n+p$, for $p = 3$ and $n = 10$; the central basis functions $N_{i,[p]}$, $i = p+1,\dots,n$, are depicted in blue.

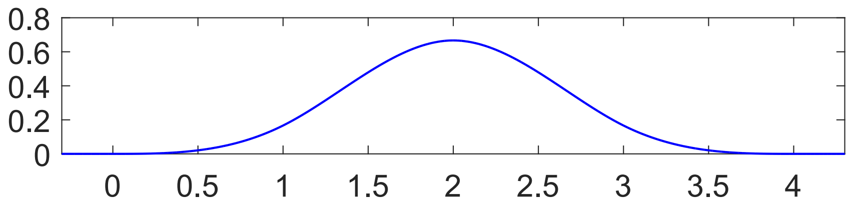

Figure 3.3: Graph of the cubic cardinal B-spline $\phi_{[3]}$.

Let $\phi_{[q]}$ be the cardinal B-spline of degree $q \geq 0$ over the uniform knot sequence $\{0,1,\dots,q+1\}$, which is defined recursively as follows [19]:

$$\phi_{[0]}(t) = \chi_{[0,1)}(t), \qquad t \in \mathbb{R}, \tag{3.106}$$

$$\phi_{[q]}(t) = \frac{t}{q}\phi_{[q-1]}(t) + \frac{q+1-t}{q}\phi_{[q-1]}(t-1), \qquad t \in \mathbb{R}, \qquad q \geq 1. \tag{3.107}$$

It is known from [15, 19] that $\phi_{[q]} \in C^{q-1}(\mathbb{R})$ and

$$\operatorname{supp}(\phi_{[q]}) = [0,q+1]. \tag{3.108}$$

Moreover, the following symmetry property holds by [29, Lemma 3] (see also [15, p. 86]):

$$\phi_{[q]}^{(r)}\Big(\frac{q+1}{2}+t\Big) = (-1)^r \phi_{[q]}^{(r)}\Big(\frac{q+1}{2}-t\Big), \qquad t \in \mathbb{R}, \qquad r,q \geq 0, \tag{3.109}$$

where $\phi_{[q]}^{(r)}$ is the rth derivative of $\phi_{[q]}$. Note that $\phi_{[q]}^{(r)}(t)$ is defined for all $t \in \mathbb{R}$ if $r < q$, and for all $t \in \mathbb{R}\backslash\{0,1,\dots,q+1\}$ if $r \geq q$. Nevertheless, (3.109) holds for all $t \in \mathbb{R}$, because when the left-hand side is not defined, the right-hand side is not defined as well. Concerning the L^2 inner products of derivatives of cardinal B-splines, it was proved in [29, Lemma 4] that

$$\int_{\mathbb{R}} \phi_{[q_1]}^{(r_1)}(t)\phi_{[q_2]}^{(r_2)}(t+\tau)\mathrm{d}t = (-1)^{r_1}\phi_{[q_1+q_2+1]}^{(r_1+r_2)}(q_1+1+\tau) = (-1)^{r_2}\phi_{[q_1+q_2+1]}^{(r_1+r_2)}(q_2+1-\tau) \tag{3.110}$$

for every $\tau \in \mathbb{R}$ and every $q_1,q_2,r_1,r_2 \geq 0$. Equation (3.110) is a property of the more general family of box splines [54] and generalizes the result appearing in [15, p. 89]. Cardinal B-splines are of interest herein, because the so-called central basis functions $N_{i,[p]}$, $i = p+1,\dots,n$, are uniformly shifted and scaled versions of the cardinal B-spline $\phi_{[p]}$. This is illustrated in Figures 3.2–3.3 for $p = 3$. In formulas, we have

$$N_{i,[p]}(t) = \phi_{[p]}(nt-i+p+1), \qquad t \in [0,1], \qquad i = p+1,\dots,n, \tag{3.111}$$

and, consequently,

$$N'_{i,[p]}(t) = n\,\phi'_{[p]}(nt-i+p+1), \qquad t \in [0,1], \qquad i = p+1,\dots,n, \tag{3.112}$$

$$N''_{i,[p]}(t) = n^2\phi''_{[p]}(nt-i+p+1), \qquad t \in [0,1], \qquad i = p+1,\dots,n. \tag{3.113}$$

Remark 3.7. For degree $p=1$, the central B-spline basis functions $N_{2,[1]},\ldots,N_{n,[1]}$ are the hat-functions $\varphi_1,\ldots,\varphi_{n-1}$ corresponding to the grid points

$$x_i = ih, \qquad i=0,\ldots,n, \qquad h=\frac{1}{n}.$$

To see this, simply write (3.98) for $p=1$ and compare it with (3.61). The graph of $N_{2,[1]},\ldots,N_{n,[1]}$ for $n=10$ is depicted in Figure 3.1.

In view of (3.99) and (3.100), the Greville abscissa $\xi_{i,[p]}$ lies in the support of $N_{i,[p]}$,

$$\xi_{i,[p]} \in \mathrm{supp}(N_{i,[p]}) = [t_i, t_{i+p+1}], \qquad i=1,\ldots,n+p. \tag{3.114}$$

The central Greville abscissae $\xi_{i,[p]}$, $i=p+1,\ldots,n$, which are the Greville abscissae associated with the central basis functions (3.111), simplify to

$$\xi_{i,[p]} = \frac{i}{n} - \frac{p+1}{2n}, \qquad i=p+1,\ldots,n. \tag{3.115}$$

The Greville abscissae are somehow equivalent, in an asymptotic sense, to the uniform knots in $[0,1]$. More precisely,

$$\left|\xi_{i,[p]} - \frac{i}{n+p}\right| \le \frac{C_p}{n}, \qquad i=1,\ldots,n+p, \tag{3.116}$$

where C_p depends only on p. The proof of (3.116) is a matter of straightforward computations; we leave the details to the reader.

B-spline IgA collocation matrices In the IgA collocation approach based on (uniform) B-splines, the basis functions $\hat\varphi_1,\ldots,\hat\varphi_N$ in (3.84) are chosen as the B-splines $N_{2,[p]},\ldots,N_{n+p-1,[p]}$ in (3.96), i.e.,

$$\hat\varphi_i = N_{i+1,[p]}, \qquad i=1,\ldots,n+p-2. \tag{3.117}$$

In this setting, $N=n+p-2$. Note that the boundary functions $N_{1,[p]}$ and $N_{n+p,[p]}$ are excluded because they do not vanish on the boundary $\partial([0,1])$; see also Figure 3.2. As for the collocation points $\hat\tau_1\ldots,\hat\tau_N$ in (3.85), they are chosen as the Greville abscissae $\xi_{2,[p]},\ldots,\xi_{n+p-1,[p]}$ in (3.99), i.e.,

$$\hat\tau_i = \xi_{i+1,[p]}, \qquad i=1,\ldots,n+p-2. \tag{3.118}$$

In what follows we assume $p\ge 2$, so as to ensure that $N''_{j+1,[p]}(\xi_{i+1,[p]})$ is defined for all $i,j=1,\ldots,n+p-2$. The collocation matrix (3.93) resulting from the choices of $\hat\varphi_i,\hat\tau_i$ as in (3.117)–(3.118) will be denoted by $A^{[p]}_{G,n}$, in order to emphasize its dependence on the geometry map G and the parameters n,p:

$$\begin{aligned} A^{[p]}_{G,n} &= \Big[-a_G(\xi_{i+1,[p]})N''_{j+1,[p]}(\xi_{i+1,[p]}) + s_G(\xi_{i+1,[p]})N'_{j+1,[p]}(\xi_{i+1,[p]}) + c_G(\xi_{i+1,[p]})N_{j+1,[p]}(\xi_{i+1,[p]})\Big]_{i,j=1}^{n+p-2} \\ &= D^{[p]}_n(a_G)\,K^{[p]}_n + D^{[p]}_n(s_G)\,H^{[p]}_n + D^{[p]}_n(c_G)\,M^{[p]}_n, \end{aligned}$$

where

$$D^{[p]}_n(v) = \operatorname*{diag}_{i=1,\ldots,n+p-2} v(\xi_{i+1,[p]})$$

is the diagonal sampling matrix containing the samples of the function $v:[0,1]\to\mathbb{R}$ at the Greville abscissae, and

$$\begin{aligned} K^{[p]}_n &= \big[-N''_{j+1,[p]}(\xi_{i+1,[p]})\big]_{i,j=1}^{n+p-2}, \\ H^{[p]}_n &= \big[N'_{j+1,[p]}(\xi_{i+1,[p]})\big]_{i,j=1}^{n+p-2}, \\ M^{[p]}_n &= \big[N_{j+1,[p]}(\xi_{i+1,[p]})\big]_{i,j=1}^{n+p-2}. \end{aligned}$$

Note that $A^{[p]}_{G,n}$ can be decomposed as follows:

$$A^{[p]}_{G,n} = K^{[p]}_{G,n} + Z^{[p]}_{G,n},$$

where

$$K_{G,n}^{[p]} = \Big[-a_G(\xi_{i+1,[p]})N''_{j+1,[p]}(\xi_{i+1,[p]})\Big]_{i,j=1}^{n+p-2} = D_n^{[p]}(a_G)\,K_n^{[p]}$$

is the diffusion matrix, i.e., the matrix resulting from the discretization of the higher-order (diffusion) term in (3.82), and

$$Z_{G,n}^{[p]} = \Big[s_G(\xi_{i+1,[p]})N'_{j+1,[p]}(\xi_{i+1,[p]}) + c_G(\xi_{i+1,[p]})N_{j+1,[p]}(\xi_{i+1,[p]})\Big]_{i,j=1}^{n+p-2} = D_n^{[p]}(s_G)\,H_n^{[p]} + D_n^{[p]}(c_G)\,M_n^{[p]}$$

is the matrix resulting from the discretization of the terms in (3.82) with lower-order derivatives (i.e., the convection and reaction terms). As already noticed in the previous sections about FD and FE discretizations, the matrix $Z_{G,n}^{[p]}$ can be regarded as a ‘residual term’, since it comes from the discretization of the lower-order differential operators. Indeed, we shall see that the norm of $Z_{G,n}^{[p]}$ is negligible with respect to the norm of the diffusion matrix $K_{G,n}^{[p]}$ when the discretization parameter n is large, because, after normalization by n^2, it will turn out that $\|n^{-2}Z_{G,n}^{[p]}\|$ tends to 0 as $n\to\infty$ (contrary to $\|n^{-2}K_{G,n}^{[p]}\|$, which remains bounded away from 0 and ∞).

Let us now provide an approximate construction of $K_n^{[p]}$, $M_n^{[p]}$, $H_n^{[p]}$. This is necessary for the GLT analysis of this section. We only construct the submatrices

$$\big[(K_n^{[p]})_{ij}\big]_{i,j=p}^{n-1}, \qquad \big[(H_n^{[p]})_{ij}\big]_{i,j=p}^{n-1}, \qquad \big[(M_n^{[p]})_{ij}\big]_{i,j=p}^{n-1}, \tag{3.119}$$

which are determined by the central basis functions (3.111) and by the central Greville abscissae (3.115). Note that the submatrix $[(K_n^{[p]})_{ij}]_{i,j=p}^{n-1}$, when embedded in any matrix of size $n+p-2$ at the right place (identified by the row and column indices $p,\ldots,n-1$), provides an approximation of $K_n^{[p]}$ up to a low-rank correction. A similar consideration also applies to the submatrices $[(H_n^{[p]})_{ij}]_{i,j=p}^{n-1}$ and $[(M_n^{[p]})_{ij}]_{i,j=p}^{n-1}$. A direct computation based on (3.109), (3.111)–(3.113) and (3.115) shows that, for $i,j=p,\ldots,n-1$,

$$\begin{aligned}
(K_n^{[p]})_{ij} &= -n^2\phi''_{[p]}\Big(\frac{p+1}{2}+i-j\Big) = -n^2\phi''_{[p]}\Big(\frac{p+1}{2}-i+j\Big),\\
(H_n^{[p]})_{ij} &= n\,\phi'_{[p]}\Big(\frac{p+1}{2}+i-j\Big) = -n\,\phi'_{[p]}\Big(\frac{p+1}{2}-i+j\Big),\\
(M_n^{[p]})_{ij} &= \phi_{[p]}\Big(\frac{p+1}{2}+i-j\Big) = \phi_{[p]}\Big(\frac{p+1}{2}-i+j\Big).
\end{aligned}$$

Since their entries depend only on the difference $i-j$, the submatrices (3.119) are Toeplitz matrices, and precisely

$$\big[(K_n^{[p]})_{ij}\big]_{i,j=p}^{n-1} = n^2\Big[-\phi''_{[p]}\Big(\frac{p+1}{2}-i+j\Big)\Big]_{i,j=p}^{n-1} = n^2\,T_{n-p}(f_p), \tag{3.120}$$

$$\big[(H_n^{[p]})_{ij}\big]_{i,j=p}^{n-1} = n\Big[-\phi'_{[p]}\Big(\frac{p+1}{2}-i+j\Big)\Big]_{i,j=p}^{n-1} = n\,\mathrm{i}\,T_{n-p}(g_p), \tag{3.121}$$

$$\big[(M_n^{[p]})_{ij}\big]_{i,j=p}^{n-1} = \Big[\phi_{[p]}\Big(\frac{p+1}{2}-i+j\Big)\Big]_{i,j=p}^{n-1} = T_{n-p}(h_p), \tag{3.122}$$

where

$$f_p(\theta) = \sum_{k\in\mathbb{Z}} -\phi''_{[p]}\Big(\frac{p+1}{2}-k\Big)\,\mathrm{e}^{\mathrm{i}k\theta} = -\phi''_{[p]}\Big(\frac{p+1}{2}\Big) - 2\sum_{k=1}^{\lfloor p/2\rfloor}\phi''_{[p]}\Big(\frac{p+1}{2}-k\Big)\cos(k\theta), \tag{3.123}$$

$$g_p(\theta) = -\mathrm{i}\sum_{k\in\mathbb{Z}} -\phi'_{[p]}\Big(\frac{p+1}{2}-k\Big)\,\mathrm{e}^{\mathrm{i}k\theta} = -2\sum_{k=1}^{\lfloor p/2\rfloor}\phi'_{[p]}\Big(\frac{p+1}{2}-k\Big)\sin(k\theta), \tag{3.124}$$

$$h_p(\theta) = \sum_{k\in\mathbb{Z}} \phi_{[p]}\Big(\frac{p+1}{2}-k\Big)\,\mathrm{e}^{\mathrm{i}k\theta} = \phi_{[p]}\Big(\frac{p+1}{2}\Big) + 2\sum_{k=1}^{\lfloor p/2\rfloor}\phi_{[p]}\Big(\frac{p+1}{2}-k\Big)\cos(k\theta); \tag{3.125}$$

note that we used (3.108)–(3.109) to simplify the expressions of $f_p(\theta)$, $g_p(\theta)$, $h_p(\theta)$. It follows from (3.120) that $T_{n-p}(f_p)$ is the principal submatrix of both $n^{-2}K_n^{[p]}$ and $T_{n+p-2}(f_p)$ corresponding to the set of indices $p,\ldots,n-1$. Similar results follow from (3.121)–(3.122), and so we obtain

$$n^{-2}K_n^{[p]} = T_{n+p-2}(f_p) + R_n^{[p]}, \qquad \operatorname{rank}(R_n^{[p]}) \le 4(p-1), \tag{3.126}$$

$$-\mathrm{i}n^{-1}H_n^{[p]} = T_{n+p-2}(g_p) + S_n^{[p]}, \qquad \operatorname{rank}(S_n^{[p]}) \le 4(p-1), \tag{3.127}$$

$$M_n^{[p]} = T_{n+p-2}(h_p) + V_n^{[p]}, \qquad \operatorname{rank}(V_n^{[p]}) \le 4(p-1). \tag{3.128}$$

To better appreciate the above construction, let us see two examples. We only consider the case of the matrix $K_n^{[p]}$ because for $H_n^{[p]}$ and $M_n^{[p]}$ the situation is the same. In the first example, we fix $p = 3$. The matrix $K_n^{[3]}$ is given by

$$K_n^{[3]} = \frac{n^2}{6}\left[\begin{array}{ccccccccc}
33 & -7 & -2 & & & & & & \\
-9 & 15 & -6 & & & & & & \\
 & -6 & 12 & -6 & & & & & \\
 & & -6 & 12 & -6 & & & & \\
 & & & \ddots & \ddots & \ddots & & & \\
 & & & & -6 & 12 & -6 & & \\
 & & & & & -6 & 12 & -6 & \\
 & & & & & & -6 & 15 & -9 \\
 & & & & & & -2 & -7 & 33
\end{array}\right].$$

The submatrix $T_{n-2}(f_3)$ appears in correspondence of the highlighted box and we have

$$f_3(\theta) = \frac{1}{6}(-6\mathrm{e}^{\mathrm{i}\theta} + 12 - 6\mathrm{e}^{-\mathrm{i}\theta}) = 2 - 2\cos\theta,$$

as given by (3.123) for $p = 3$. In the second example, we fix $p = 4$. The matrix $K_n^{[4]}$ is given by

$$K_n^{[4]} = \frac{n^2}{96}\left[\begin{array}{ccccccccccccc}
855 & -133 & -71 & -3 & & & & & & & & & \\
-81 & 243 & -63 & -27 & & & & & & & & & \\
-36 & -36 & 132 & -48 & -12 & & & & & & & & \\
 & -16 & -44 & 120 & -48 & -12 & & & & & & & \\
 & & -12 & -48 & 120 & -48 & -12 & & & & & & \\
 & & & -12 & -48 & 120 & -48 & -12 & & & & & \\
 & & & & \ddots & \ddots & \ddots & \ddots & \ddots & & & & \\
 & & & & & -12 & -48 & 120 & -48 & -12 & & & \\
 & & & & & & -12 & -48 & 120 & -48 & -12 & & \\
 & & & & & & & -12 & -48 & 120 & -44 & -16 & \\
 & & & & & & & & -12 & -48 & 132 & -36 & -36 \\
 & & & & & & & & & -27 & -63 & 243 & -81 \\
 & & & & & & & & & -3 & -71 & -133 & 855
\end{array}\right].$$

The submatrix $T_{n-3}(f_4)$ appears in correspondence of the highlighted box and we have

$$f_4(\theta) = \frac{1}{96}(-12\mathrm{e}^{2\mathrm{i}\theta} - 48\mathrm{e}^{\mathrm{i}\theta} + 120 - 48\mathrm{e}^{-\mathrm{i}\theta} - 12\mathrm{e}^{-2\mathrm{i}\theta}) = \frac{5}{4} - \cos\theta - \frac{1}{4}\cos(2\theta),$$

as given by (3.123) for $p = 4$.

Before passing to the GLT analysis of the collocation matrices $A_{G,n}^{[p]}$, we prove the existence of an n-independent bound for the spectral norms of $n^{-2}K_n^{[p]}$, $n^{-1}H_n^{[p]}$, $M_n^{[p]}$. Actually, one could also prove that the components of $n^{-2}K_n^{[p]}$, $n^{-1}H_n^{[p]}$, $M_n^{[p]}$ do not depend on n as illustrated above for the matrix $n^{-2}K_n^{[p]}$ in the cases $p = 3, 4$. However, for our purposes it suffices to show that, for every $p \ge 2$, there exists a constant $C^{[p]}$ such that, for all n,

$$\|n^{-2}K_n^{[p]}\| \le C^{[p]}, \qquad \|n^{-1}H_n^{[p]}\| \le C^{[p]}, \qquad \|M_n^{[p]}\| \le C^{[p]}. \tag{3.129}$$

To prove (3.129), we note that $K_n^{[p]}$, $H_n^{[p]}$, $M_n^{[p]}$ are banded, with bandwidth bounded by $2p+1$. Indeed, if $|i-j| > p$, one can show that $(K_n^{[p]})_{ij} = (H_n^{[p]})_{ij} = (M_n^{[p]})_{ij} = 0$ by using (3.114), which implies that $\xi_{i+1,[p]}$ lies outside or on the border of $\mathrm{supp}(N_{j+1,[p]})$, whose intersection with $\mathrm{supp}(N_{i+1,[p]})$ consists of at most one of the knots t_k. Moreover, by (3.102)–(3.105), for all $i, j = 1, \dots, n+p-2$ we have

$$\begin{aligned}
|(K_n^{[p]})_{ij}| &= |N''_{j+1,[p]}(\xi_{i+1,[p]})| \le 4p(p-1)n^2, \\
|(H_n^{[p]})_{ij}| &= |N'_{j+1,[p]}(\xi_{i+1,[p]})| \le 2pn, \\
|(M_n^{[p]})_{ij}| &= |N_{j+1,[p]}(\xi_{i+1,[p]})| \le 1.
\end{aligned}$$

Hence, (3.129) follows from (3.1).

GLT analysis of the B-spline IgA collocation matrices Assuming that the geometry map G possesses some regularity properties, we show that, for any $p \ge 2$, the sequence of normalized IgA collocation matrices $\{n^{-2}A_{G,n}^{[p]}\}_n$ is a GLT sequence whose symbol describes both its singular value and spectral distribution.

Theorem 3.12. *Let Ω be a bounded open interval of $\mathbb{R}$, let $a \in C^1(\overline{\Omega})$ and $b, c \in C(\overline{\Omega})$. Let $p \ge 2$ and let $G : [0,1] \to \overline{\Omega}$ be such that $G \in C^2([0,1])$ and $G'(\hat{x}) \neq 0$ for all $\hat{x} \in [0,1]$. Then*

$$\{n^{-2}A_{G,n}^{[p]}\}_n \sim_{\mathrm{GLT}} f_{G,p} \tag{3.130}$$

and

$$\{n^{-2}A_{G,n}^{[p]}\}_n \sim_{\sigma,\lambda} f_{G,p}, \tag{3.131}$$

where

$$f_{G,p}(\hat{x},\theta) = a_G(\hat{x}) f_p(\theta) = \frac{a(G(\hat{x}))}{(G'(\hat{x}))^2} f_p(\theta) \tag{3.132}$$

and $f_p(\theta)$ is defined in (3.123).

Proof. The proof consists of the following steps. Throughout the proof, the letter C will denote a generic constant independent of n.

Step 1. We show that

$$\|n^{-2}K_{G,n}^{[p]}\| \le C \tag{3.133}$$

and

$$\|n^{-2}Z_{G,n}^{[p]}\| \le C/n. \tag{3.134}$$

To prove (3.133), it suffices to use the regularity of G and (3.129):

$$\|n^{-2}K_{G,n}^{[p]}\| = \|n^{-2}D_n^{[p]}(a_G)K_n^{[p]}\| \le \|a_G\|_\infty C^{[p]} \le \frac{C^{[p]}\|a\|_\infty}{\min_{\hat{x}\in[0,1]} |G'(\hat{x})|^2}.$$

The proof of (3.134) is similar. It suffices to use the fact that $G \in C^2([0,1])$ and (3.129):

$$\|n^{-2}Z_{G,n}^{[p]}\| = \|n^{-2}D_n^{[p]}(s_G)H_n^{[p]} + n^{-2}D_n^{[p]}(c_G)M_n^{[p]}\| \le n^{-1}C^{[p]}\left(\frac{\|a\|_\infty\|G''\|_\infty}{\min_{\hat{x}\in[0,1]}|G'(\hat{x})|^3} + \frac{\|a'\|_\infty + \|b\|_\infty}{\min_{\hat{x}\in[0,1]}|G'(\hat{x})|}\right) + n^{-2}C^{[p]}\|c\|_\infty.$$

Step 2. Define the symmetric matrix

$$\tilde{K}_{G,n}^{[p]} = S_{n+p-2}(a_G) \circ n^2\, T_{n+p-2}(f_p), \tag{3.135}$$

where we recall that $S_m(v)$ is the mth arrow-shaped sampling matrix generated by v (see (3.6)), and consider the following decomposition of $n^{-2}A_{G,n}^{[p]}$:

$$n^{-2}A_{G,n}^{[p]} = n^{-2}\tilde{K}_{G,n}^{[p]} + \big(n^{-2}K_{G,n}^{[p]} - n^{-2}\tilde{K}_{G,n}^{[p]}\big) + n^{-2}Z_{G,n}^{[p]}. \tag{3.136}$$

We know from Theorem 3.2 that $\|n^{-2}\tilde{K}_{G,n}^{[p]}\| \le C$ and $\{n^{-2}\tilde{K}_{G,n}^{[p]}\}_n \sim_{\mathrm{GLT}} f_{G,p}(\hat{x},\theta)$.

Step 3. We show that

$$\|n^{-2}K_{G,n}^{[p]} - n^{-2}\tilde{K}_{G,n}^{[p]}\|_1 = o(n). \tag{3.137}$$

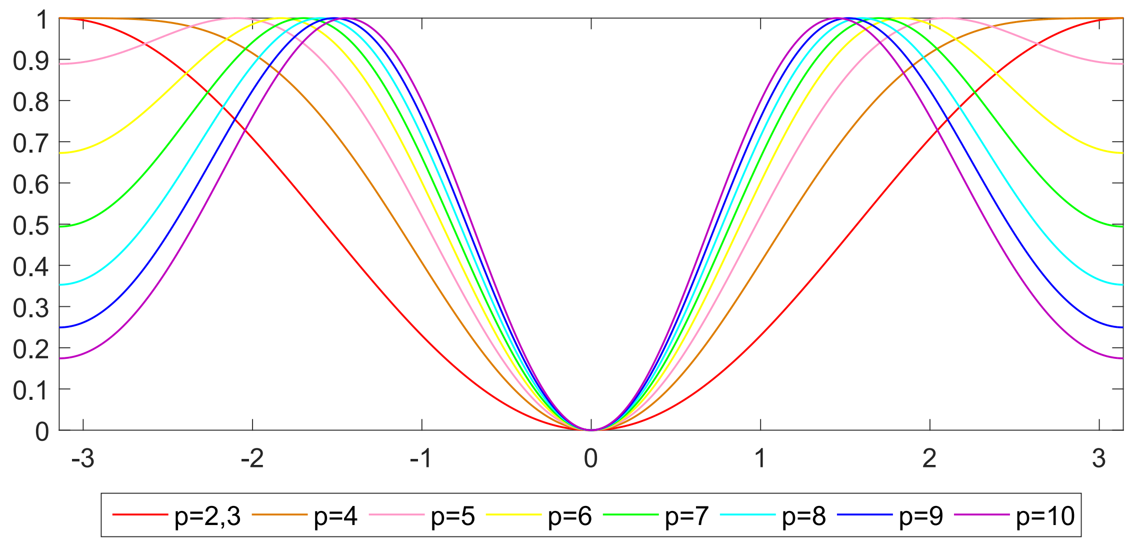

Figure 3.4: Graph of f_p/M_{f_p} for $p=2,\dots,10$.

Once this is done, the thesis is proved. Indeed, from (3.137) and (3.134) we obtain

$$\|(n^{-2}K_{G,n}^{[p]}-n^{-2}\tilde{K}_{G,n}^{[p]})+n^{-2}Z_{G,n}^{[p]}\|_1\le\|n^{-2}K_{G,n}^{[p]}-n^{-2}\tilde{K}_{G,n}^{[p]}\|_1+\|n^{-2}Z_{G,n}^{[p]}\|n=o(n),$$

hence $\{(n^{-2}K_{G,n}^{[p]}-n^{-2}\tilde{K}_{G,n}^{[p]})+n^{-2}Z_{G,n}^{[p]}\}_n$ is zero-distributed by **Z 2**. Thus, the GLT relation (3.130) follows from the decomposition (3.136) and **GLT 3 – GLT 4**, the singular value distribution in (3.131) follows from **GLT 1**, and the eigenvalue distribution in (3.131) follows from **GLT 2**.

To prove (3.137), we decompose the difference $n^{-2}K_{G,n}^{[p]}-n^{-2}\tilde{K}_{G,n}^{[p]}$ as follows:

$$\begin{aligned} n^{-2}K_{G,n}^{[p]}-n^{-2}\tilde{K}_{G,n}^{[p]}&=n^{-2}D_n^{[p]}(a_G)\,K_n^{[p]}-S_{n+p-2}(a_G)\circ T_{n+p-2}(f_p)\\ &=n^{-2}D_n^{[p]}(a_G)\,K_n^{[p]}-D_n^{[p]}(a_G)\,T_{n+p-2}(f_p) &&(3.138)\\ &\quad+D_n^{[p]}(a_G)\,T_{n+p-2}(f_p)-D_{n+p-2}(a_G)\,T_{n+p-2}(f_p) &&(3.139)\\ &\quad+D_{n+p-2}(a_G)\,T_{n+p-2}(f_p)-S_{n+p-2}(a_G)\circ T_{n+p-2}(f_p). &&(3.140)\end{aligned}$$

We consider separately the three matrices in (3.138)–(3.140) and we show that their trace-norms are $o(n)$.

- By (3.126), the rank of the matrix (3.138) is bounded by $4(p-1)$. By the regularity of G, the inequality (3.129) and **T 3**, the spectral norm of (3.138) is bounded by C. Thus, the trace-norm of (3.138) is $o(n)$ (actually, $O(1)$) by (3.3).
- By (3.116), the continuity of a_G and **T 3**, the spectral norm of the matrix (3.139) is bounded by $C\omega_{a_G}(n^{-1})$, so it tends to 0. Hence, the trace-norm of (3.139) is $o(n)$ by (3.3).
- By Theorem 3.2, the spectral norm of the matrix (3.140) is bounded by $C\omega_{a_G}(n^{-1})$, so it tends to 0. Hence, the trace-norm of (3.140) is $o(n)$ by (3.3).

In conclusion, $\|n^{-2}K_{G,n}^{[p]}-n^{-2}\tilde{K}_{G,n}^{[p]}\|_1=o(n)$. □

Remark 3.8 (formal structure of the symbol). We invite the reader to compare the symbol (3.132) with the transformed problem (3.89). It is clear that the higher-order operator $-a_G(\hat{x})\hat{u}''(\hat{x})$ has a discrete spectral counterpart $a_G(\hat{x})f_p(\theta)$ which looks formally the same (as in the FD and FE cases; see Remarks 3.1 and 3.6). To better appreciate the formal analogy, note that $f_p(\theta)$ is the trigonometric polynomial in the Fourier variable coming from the B-spline IgA collocation discretization of the second derivative $-\hat{u}''(\hat{x})$ on the parametric domain $[0,1]$. Indeed, $f_p(\theta)$ is the symbol of the sequence of B-spline IgA collocation diffusion matrices $\{n^{-2}K_n^{[p]}\}_n$, which arises from the B-spline IgA collocation approximation of (3.82) in the case where $a(x)=1$, $b(x)=c(x)=0$ identically, $\Omega=(0,1)$ and G is the identity map over $[0,1]$; note that in this case (3.82) is the same as (3.89), $x=G(\hat{x})=\hat{x}$ and $u=\hat{u}$.

Remark 3.9 (nonnegativity and order of the zero at $\theta = 0$). Figure 3.4 shows the graph of $f_p(\theta)$ normalized by its maximum $M_{f_p} = \max_{\theta\in[-\pi,\pi]} f_p(\theta)$ for $p = 2,\dots,10$. Note that $f_2(\theta) = f_3(\theta) = 2-2\cos\theta$. We see from the figure (and it was proved in [22]) that $f_p(\theta)$ is nonnegative over $[-\pi,\pi]$ and has a unique zero of order 2 at $\theta = 0$ because

$$\lim_{\theta\to 0}\frac{f_p(\theta)}{\theta^2} = 1.$$

This reflects the fact that, as observed in Remark 3.8, $f_p(\theta)$ arises from the B-spline IgA collocation discretization of the second derivative $-\hat{u}''(\hat{x})$ on the parametric domain $[0,1]$, which is a differential operator of order 2 (and it is nonnegative on $\{v\in C^2([0,1]) :\ v(0) = v(1) = 0\}$); see also Remarks 3.2, 3.4 and 3.5.

Further properties of the functions $f_p(\theta)$, $g_p(\theta)$, $h_p(\theta)$ can be found in [22, Section 3]. In particular, it was proved therein that $f_p(\pi)/M_{f_p}\to 0$ exponentially as $p\to\infty$. Moreover, observing that $h_p(\theta)$ is defined by (3.125) for all degrees $p\ge 0$ (and we have $h_0(\theta) = h_1(\theta) = 1$ identically) provided that we use the standard convention that an empty sum like $\sum_{k=1}^{0}\phi_{[1]}(1-k)\cos(k\theta)$ equals 0,[3] it was proved in [22] that, for all $p\ge 2$ and $\theta\in[-\pi,\pi]$,

$$f_p(\theta) = (2-2\cos\theta)h_{p-2}(\theta),\qquad \left(\frac{2}{\pi}\right)^{p-1}\le h_{p-2}(\theta)\le h_{p-2}(0) = 1. \tag{3.141}$$

3.4.2 Galerkin B-Spline IgA Discretization of Convection-Diffusion-Reaction Equations

Consider the convection-diffusion-reaction problem

$$\begin{cases} -(a(x)u'(x))' + b(x)u'(x) + c(x)u(x) = f(x), & x\in\Omega,\\ u(x) = 0, & x\in\partial\Omega, \end{cases} \tag{3.142}$$

where Ω is a bounded open interval of $\mathbb{R}$, $f\in L^2(\Omega)$ and $a,b,c\in L^\infty(\Omega)$. Problem (3.142) is the same as (3.81), except for the assumptions on a,b,c,f. We consider the isogeometric Galerkin approximation of (3.142) based on uniform B-splines of degree $p\ge 1$. This approximation technique is described below in some detail. For more on IgA Galerkin methods, see [17, 39].

Isogeometric Galerkin approximation The weak form of (3.142) reads as follows: find $u\in H_0^1(\Omega)$ such that

$$\mathrm{a}(u,v) = \mathrm{f}(v),\qquad \forall\, v\in H_0^1(\Omega),$$

where

$$\begin{aligned} \mathrm{a}(u,v) &= \int_\Omega \big(a(x)u'(x)v'(x) + b(x)u'(x)v(x) + c(x)u(x)v(x)\big)\mathrm{d}x,\\ \mathrm{f}(v) &= \int_\Omega f(x)v(x)\mathrm{d}x. \end{aligned}$$

In the standard Galerkin method, we look for an approximation $u_{\mathscr{W}}$ of u by choosing a finite dimensional vector space $\mathscr{W}\subset H_0^1(\Omega)$, the so-called approximation space, and by solving the following (Galerkin) problem: find $u_{\mathscr{W}}\in\mathscr{W}$ such that

$$\mathrm{a}(u_{\mathscr{W}},v) = \mathrm{f}(v),\qquad \forall\, v\in\mathscr{W}.$$

If $\{\varphi_1,\dots,\varphi_N\}$ is a basis of $\mathscr{W}$, then we can write $u_{\mathscr{W}} = \sum_{j=1}^N u_j\varphi_j$ for a unique vector $\mathbf{u} = (u_1,\dots,u_N)^T$, and, by linearity, the computation of $u_{\mathscr{W}}$ (i.e., of $\mathbf{u}$) reduces to solving the linear system

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

where $\mathbf{f} = \big[\mathrm{f}(\varphi_i)\big]_{i=1}^N$ and

$$A = [\mathrm{a}(\varphi_j,\varphi_i)]_{i,j=1}^N = \left[\int_\Omega \big(a(x)\varphi_j'(x)\varphi_i'(x) + b(x)\varphi_j'(x)\varphi_i(x) + c(x)\varphi_j(x)\varphi_i(x)\big)\mathrm{d}x\right]_{i,j=1}^N \tag{3.143}$$

[3] On the contrary, $f_p(\theta)$ and $g_p(\theta)$ are defined by (3.123) and (3.124) only for $p\ge 2$, because $\phi_{[1]}'(1)$ and $\phi_{[1]}''(1)$ do not exist.

is the stiffness matrix.

Now, suppose that the physical domain Ω can be described by a global geometry function $G : [0,1] \to \overline{\Omega}$, which is invertible and satisfies $G(\partial([0,1])) = \partial\overline{\Omega}$. Let $\{\hat{\varphi}_1,\ldots,\hat{\varphi}_N\}$ be a set of basis functions defined on the parametric (or reference) domain $[0,1]$ and vanishing on the boundary $\partial([0,1])$. In the isogeometric Galerkin approach, we find an approximation $u_{\mathscr{W}}$ of u by using the standard Galerkin method, in which the approximation space is chosen as $\mathscr{W} = \mathrm{span}(\varphi_1,\ldots,\varphi_N)$, where

$$\varphi_i(x) = \hat{\varphi}_i(G^{-1}(x)) = \hat{\varphi}_i(\hat{x}), \qquad x = G(\hat{x}). \tag{3.144}$$

The resulting stiffness matrix A is given by (3.143), with the basis functions φ_i defined as in (3.144). Assuming that G and $\hat{\varphi}_i$, $i = 1,\ldots,N$, are sufficiently regular, we can apply standard differential calculus to obtain the following expression for A in terms of G and $\hat{\varphi}_i$, $i = 1,\ldots,N$:

$$A = \left[\int_{[0,1]} \Big(a_G(\hat{x})\hat{\varphi}_j'(\hat{x})\hat{\varphi}_i'(\hat{x}) + \frac{b(G(\hat{x}))}{G'(\hat{x})}\hat{\varphi}_j'(\hat{x})\hat{\varphi}_i(\hat{x}) + c(G(\hat{x}))\hat{\varphi}_j(\hat{x})\hat{\varphi}_i(\hat{x})\Big)|G'(\hat{x})|\mathrm{d}\hat{x}\right]_{i,j=1}^{N}, \tag{3.145}$$

where $a_G(\hat{x})$ is the same as in (3.90),

$$a_G(\hat{x}) = \frac{a(G(\hat{x}))}{(G'(\hat{x}))^2}. \tag{3.146}$$

In the IgA framework, the functions $\hat{\varphi}_i$ are usually B-splines or NURBS. Here, the role of the $\hat{\varphi}_i$ will be played by B-splines over uniform knot sequences.

Galerkin B-spline IgA discretization matrices As in the IgA collocation framework considered in Section 3.4.1, in the Galerkin B-spline IgA based on (uniform) B-splines, the basis functions $\hat{\varphi}_1,\ldots,\hat{\varphi}_N$ are chosen as the B-splines $N_{2,[p]},\ldots,N_{n+p-1,[p]}$ defined in (3.96)–(3.98), i.e.,

$$\hat{\varphi}_i = N_{i+1,[p]}, \qquad i = 1,\ldots,n+p-2.$$

The boundary functions $N_{1,[p]}$ and $N_{n+p,[p]}$ are excluded because they do not vanish on $\partial([0,1])$; see also Figure 3.2. The stiffness matrix (3.145) resulting from this choice of the $\hat{\varphi}_i$ will be denoted by $A_{G,n}^{[p]}$:

$$A_{G,n}^{[p]} = \left[\int_{[0,1]} \Big(a_G(\hat{x})N_{j+1,[p]}'(\hat{x})N_{i+1,[p]}'(\hat{x}) + \frac{b(G(\hat{x}))}{G'(\hat{x})}N_{j+1,[p]}'(\hat{x})N_{i+1,[p]}(\hat{x}) + c(G(\hat{x}))N_{j+1,[p]}(\hat{x})N_{i+1,[p]}(\hat{x})\Big)|G'(\hat{x})|\mathrm{d}\hat{x}\right]_{i,j=1}^{n+p-2}. \tag{3.147}$$

Note that $A_{G,n}^{[p]}$ can be decomposed as follows:

$$A_{G,n} = K_{G,n}^{[p]} + Z_{G,n}^{[p]}, \tag{3.148}$$

where

$$K_{G,n}^{[p]} = \left[\int_{[0,1]} a_G(\hat{x})|G'(\hat{x})|N_{j+1,[p]}'(\hat{x})N_{i+1,[p]}'(\hat{x})\mathrm{d}\hat{x}\right]_{i,j=1}^{n+p-2} \tag{3.149}$$

is the diffusion matrix, resulting from the discretization of the higher-order (diffusion) term in (3.142), and

$$Z_{G,n}^{[p]} = \left[\int_{[0,1]} \Big(\frac{b(G(\hat{x}))}{G'(\hat{x})}N_{j+1,[p]}'(\hat{x})N_{i+1,[p]}(\hat{x}) + c(G(\hat{x}))N_{j+1,[p]}(\hat{x})N_{i+1,[p]}(\hat{x})\Big)|G'(\hat{x})|\mathrm{d}\hat{x}\right]_{i,j=1}^{n+p-2} \tag{3.150}$$

is the matrix resulting from the discretization of the lower-order (convection and reaction) terms. We will see that, as usual, the GLT analysis of a properly scaled version of the sequence $\{A_{G,n}^{[p]}\}_n$ reduces to the GLT analysis of its 'diffusion part' $\{K_{G,n}^{[p]}\}_n$, because $\|Z_{G,n}^{[p]}\|$ is negligible with respect to $\|K_{G,n}^{[p]}\|$ as $n \to \infty$.

Let

$$K_n^{[p]} = \left[\int_{[0,1]} N_{j+1,[p]}'(\hat{x})N_{i+1,[p]}'(\hat{x})\mathrm{d}\hat{x}\right]_{i,j=1}^{n+p-2}, \tag{3.151}$$

$$H_n^{[p]} = \left[\int_{[0,1]} N_{j+1,[p]}'(\hat{x})N_{i+1,[p]}(\hat{x})\mathrm{d}\hat{x}\right]_{i,j=1}^{n+p-2}, \tag{3.152}$$

$$M_n^{[p]} = \left[\int_{[0,1]} N_{j+1,[p]}(\hat{x})N_{i+1,[p]}(\hat{x})\mathrm{d}\hat{x}\right]_{i,j=1}^{n+p-2}. \tag{3.153}$$

These matrices will play an important role in the GLT analysis of this section. In particular, it is necessary to understand their approximate structure. This is achieved by (approximately) construct them. We only construct their central submatrices

$$\big[(K_n^{[p]})_{ij}\big]_{i,j=p}^{n-1}, \qquad \big[(H_n^{[p]})_{ij}\big]_{i,j=p}^{n-1}, \qquad \big[(M_n^{[p]})_{ij}\big]_{i,j=p}^{n-1}, \tag{3.154}$$

which are determined by the central basis functions in (3.111). For all $i,j=p,\ldots,n-1$, noting that $[-i+p,n-i+p]\supseteq \mathrm{supp}(\phi_{[p]})=[0,p+1]$ and using (3.109)–(3.110) and (3.112), we obtain

$$\begin{aligned}(K_n^{[p]})_{ij} &= \int_{[0,1]} N'_{j+1,[p]}(\hat{x})N'_{i+1,[p]}(\hat{x})\mathrm{d}\hat{x} = n^2\int_{[0,1]}\phi'_{[p]}(n\hat{x}-j+p)\phi'_{[p]}(n\hat{x}-i+p)\mathrm{d}\hat{x}\\ &= n\int_{[-i+p,n-i+p]}\phi'_{[p]}(t+i-j)\phi'_{[p]}(t)\mathrm{d}t = n\int_{\mathbb{R}}\phi'_{[p]}(t+i-j)\phi'_{[p]}(t)\mathrm{d}t\\ &= -n\,\phi''_{[2p+1]}(p+1+i-j) = -n\,\phi''_{[2p+1]}(p+1-i+j),\end{aligned}$$

and similarly

$$\begin{aligned}(H_n^{[p]})_{ij} &= \phi'_{[2p+1]}(p+1+i-j) = -\phi'_{[2p+1]}(p+1-i+j),\\ (M_n^{[p]})_{ij} &= \frac{1}{n}\phi_{[2p+1]}(p+1+i-j) = \frac{1}{n}\phi_{[2p+1]}(p+1-i+j).\end{aligned}$$

Since their entries depend only on the difference $i-j$, the submatrices (3.154) are Toeplitz matrices. More precisely,

$$\big[(K_n^{[p]})_{ij}\big]_{i,j=p}^{n-1} = n\big[-\phi''_{[2p+1]}(p+1-i+j)\big]_{i,j=p}^{n-1} = n\,T_{n-p}(f_p), \tag{3.155}$$

$$\big[(H_n^{[p]})_{ij}\big]_{i,j=p}^{n-1} = \big[-\phi'_{[2p+1]}(p+1-i+j)\big]_{i,j=p}^{n-1} = \mathrm{i}\,T_{n-p}(g_p), \tag{3.156}$$

$$\big[(M_n^{[p]})_{ij}\big]_{i,j=p}^{n-1} = \frac{1}{n}\big[\phi_{[2p+1]}(p+1-i+j)\big]_{i,j=p}^{n-1} = \frac{1}{n}T_{n-p}(h_p), \tag{3.157}$$

where

$$f_p(\theta) = \sum_{k\in\mathbb{Z}} -\phi''_{[2p+1]}(p+1-k)\,\mathrm{e}^{\mathrm{i}k\theta} = -\phi''_{[2p+1]}(p+1) - 2\sum_{k=1}^{p}\phi''_{[2p+1]}(p+1-k)\cos(k\theta), \tag{3.158}$$

$$g_p(\theta) = -\mathrm{i}\sum_{k\in\mathbb{Z}} -\phi'_{[2p+1]}(p+1-k)\,\mathrm{e}^{\mathrm{i}k\theta} = -2\sum_{k=1}^{p}\phi'_{[2p+1]}(p+1-k)\sin(k\theta), \tag{3.159}$$

$$h_p(\theta) = \sum_{k\in\mathbb{Z}} \phi_{[2p+1]}(p+1-k)\,\mathrm{e}^{\mathrm{i}k\theta} = \phi_{[2p+1]}(p+1) + 2\sum_{k=1}^{p}\phi_{[2p+1]}(p+1-k)\cos(k\theta); \tag{3.160}$$

note that we used (3.108)–(3.109) to simplify the expressions of $f_p(\theta), g_p(\theta), h_p(\theta)$. It follows from (3.155) that $T_{n-p}(f_p)$ is the principal submatrix of both $n^{-1}K_n^{[p]}$ and $T_{n+p-2}(f_p)$ corresponding to the set of indices $p,\ldots,n-1$. Similar results follow from (3.156)–(3.157), and so

$$n^{-1}K_n^{[p]} = T_{n+p-2}(f_p) + R_n^{[p]}, \qquad \mathrm{rank}(R_n^{[p]}) \le 4(p-1), \tag{3.161}$$

$$-\mathrm{i}H_n^{[p]} = T_{n+p-2}(g_p) + S_n^{[p]}, \qquad \mathrm{rank}(S_n^{[p]}) \le 4(p-1), \tag{3.162}$$

$$n M_n^{[p]} = T_{n+p-2}(h_p) + V_n^{[p]}, \qquad \mathrm{rank}(V_n^{[p]}) \le 4(p-1). \tag{3.163}$$

Let us see two examples. In the case $p=2$, the matrix $K_n^{[2]}$ is given by

$$K_n^{[2]} = \frac{n}{6}\left[\begin{array}{c|cccccccc|c} 8 & -1 & -1 & & & & & & & \\ \hline -1 & 6 & -2 & -1 & & & & & & \\ -1 & -2 & 6 & -2 & -1 & & & & & \\ & -1 & -2 & 6 & -2 & -1 & & & & \\ & & \ddots & \ddots & \ddots & \ddots & \ddots & & & \\ & & & -1 & -2 & 6 & -2 & -1 & & \\ & & & & -1 & -2 & 6 & -2 & -1 & \\ & & & & & -1 & -2 & 6 & -1 \\ \hline & & & & & & -1 & -1 & 8 \end{array}\right].$$

The submatrix $T_{n-2}(f_2)$ appears in correspondence of the highlighted box and we have

$$f_2(\theta) = \frac{1}{6}(-\mathrm{e}^{2\mathrm{i}\theta} - 2\mathrm{e}^{\mathrm{i}\theta} + 6 - 2\mathrm{e}^{-\mathrm{i}\theta} - \mathrm{e}^{-2\mathrm{i}\theta}) = 1 - \frac{2}{3}\cos\theta - \frac{1}{3}\cos(2\theta),$$

as given by (3.158) for $p = 2$. In the case $p = 3$, the matrix $K_n^{[3]}$ is given by

$$K_n^{[3]} = \frac{n}{240}\left[\begin{array}{rr|rrrrrrrrrr|rr}
360 & 9 & -60 & -3 & & & & & & & & & \\
9 & 162 & -8 & -47 & -2 & & & & & & & & \\
\hline
-60 & -8 & 160 & -30 & -48 & -2 & & & & & & & \\
-3 & -47 & -30 & 160 & -30 & -48 & -2 & & & & & & \\
 & -2 & -48 & -30 & 160 & -30 & -48 & -2 & & & & & \\
 & & -2 & -48 & -30 & 160 & -30 & -48 & -2 & & & & \\
 & & & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & & & \\
 & & & & -2 & -48 & -30 & 160 & -30 & -48 & -2 & & \\
 & & & & & -2 & -48 & -30 & 160 & -30 & -48 & -2 & \\
 & & & & & & -2 & -48 & -30 & 160 & -30 & -47 & -3 \\
 & & & & & & & -2 & -48 & -30 & 160 & -8 & -60 \\
\hline
 & & & & & & & & -2 & -47 & -8 & 162 & 9 \\
 & & & & & & & & & -3 & -60 & 9 & 360
\end{array}\right].$$

The submatrix $T_{n-3}(f_3)$ appears in correspondence of the highlighted box and we have

$$f_3(\theta) = \frac{1}{240}(-2\mathrm{e}^{3\mathrm{i}\theta} - 48\mathrm{e}^{2\mathrm{i}\theta} - 30\mathrm{e}^{\mathrm{i}\theta} + 160 - 30\mathrm{e}^{-\mathrm{i}\theta} - 48\mathrm{e}^{-2\mathrm{i}\theta} - 2\mathrm{e}^{-3\mathrm{i}\theta}) = \frac{2}{3} - \frac{1}{4}\cos\theta - \frac{2}{5}\cos(2\theta) - \frac{1}{60}\cos(3\theta),$$

as given by (3.158) for $p = 3$.

Remark 3.10. For every degree $q \geq 1$, the functions $f_q(\theta)$, $g_q(\theta)$, $h_q(\theta)$ defined by (3.158)–(3.160) for $p = q$ coincide with the functions $f_{2q+1}(\theta)$, $g_{2q+1}(\theta)$, $h_{2q+1}(\theta)$ defined by (3.123)–(3.125) for odd degree $p = 2q+1$.

GLT analysis of the Galerkin B-spline IgA discretization matrices Assuming that the geometry map G is regular, i.e., $G \in C^1([0,1])$ and $G'(\hat{x}) \neq 0$ for every $\hat{x} \in [0,1]$, we show that, for any $p \geq 1$, $\{n^{-1}A_{G,n}^{[p]}\}_n$ is a GLT sequence whose symbol describes both its singular value and spectral distribution.

Theorem 3.13. *Let Ω be a bounded open interval of $\mathbb{R}$ and let $a, b, c \in L^\infty(\Omega)$. Let $p \geq 1$ and let $G : [0,1] \to \overline{\Omega}$ be such that $G \in C^1([0,1])$ and $G'(\hat{x}) \neq 0$ for all $\hat{x} \in [0,1]$. Then*

$$\{n^{-1}A_{G,n}^{[p]}\}_n \sim_{\mathrm{GLT}} f_{G,p} \tag{3.164}$$

and

$$\{n^{-1}A_{G,n}^{[p]}\}_n \sim_{\sigma,\lambda} f_{G,p}, \tag{3.165}$$

where

$$f_{G,p}(\hat{x},\theta) = a_G(\hat{x})|G'(\hat{x})|f_p(\theta) = \frac{a(G(\hat{x}))}{|G'(\hat{x})|}f_p(\theta) \tag{3.166}$$

and $f_p(\theta)$ is defined in (3.158).

Proof. We follow the same argument as in the proof of Theorem 3.10. Throughout the proof, the letter C will denote a generic constant independent of n.

Step 1. We show that

$$\|n^{-1}K_{G,n}^{[p]}\| \leq C \tag{3.167}$$

and

$$\|n^{-1}Z_{G,n}^{[p]}\| \leq C/n. \tag{3.168}$$

To prove (3.167), we note that $K_{G,n}^{[p]}$ is a banded matrix, with bandwidth at most equal to $2p+1$. Indeed, due to the local support property (3.100), if $|i-j|>p$ then the supports of $N_{i+1,[p]}$ and $N_{j+1,[p]}$ intersect in at most one point, hence $(K_{G,n}^{[p]})_{ij}=0$. Moreover, by (3.100) and (3.104), for all $i,j=1,\ldots,n+p-2$ we have

$$|(K_{G,n}^{[p]})_{ij}| = \left|\int_{[0,1]} a_G(\hat{x})|G'(\hat{x})|N'_{j+1,[p]}(\hat{x})N'_{i+1,[p]}(\hat{x})\mathrm{d}\hat{x}\right| = \left|\int_{[t_{i+1},t_{i+p+2}]} \frac{a(G(\hat{x}))}{|G'(\hat{x})|}N'_{j+1,[p]}(\hat{x})N'_{i+1,[p]}(\hat{x})\mathrm{d}\hat{x}\right|$$
$$\le \frac{4p^2n^2\|a\|_{L^\infty}}{\min_{\hat{x}\in[0,1]}|G'(\hat{x})|}\int_{[t_{i+1},t_{i+p+2}]}\mathrm{d}\hat{x} \le \frac{4p^2(p+1)n\|a\|_{L^\infty}}{\min_{\hat{x}\in[0,1]}|G'(\hat{x})|},$$

where in the last inequality we used the fact that $t_{k+p+1}-t_k \le (p+1)/n$ for all $k=1,\ldots,n+p$; see (3.94)–(3.95). In conclusion, the components of the banded matrix $n^{-1}K_{G,n}^{[p]}$ are bounded (in modulus) by a constant independent of n, and (3.167) follows from (3.1).

To prove (3.168), we follow the same argument as for the proof of (3.167). Due to the local support property (3.100), $Z_{G,n}^{[p]}$ is banded and, precisely, $(Z_{G,n}^{[p]})_{ij}=0$ whenever $|i-j|>p$. Moreover, by (3.100) and (3.102)–(3.104), for all $i,j=1,\ldots,n+p-2$ we have

$$|(Z_{G,n}^{[p]})_{ij}| = \left|\int_{[t_{i+1},t_{i+p+2}]}\Big(\frac{b(G(\hat{x}))}{G'(\hat{x})}N'_{j+1,[p]}(\hat{x})N_{i+1,[p]}(\hat{x}) + c(G(\hat{x}))N_{j+1,[p]}(\hat{x})N_{i+1,[p]}(\hat{x})\Big)|G'(\hat{x})|\mathrm{d}\hat{x}\right|$$
$$\le 2p(p+1)\|b\|_{L^\infty} + \frac{(p+1)\|c\|_{L^\infty}\|G'\|_\infty}{n},$$

and (3.168) follows from (3.1).

Step 2. Consider the linear operator $K_n^{[p]}(\cdot): L^1([0,1]) \to \mathbb{R}^{(n+p-2)\times(n+p-2)}$,

$$K_n^{[p]}(g) = \left[\int_{[0,1]} g(\hat{x})N'_{j+1,[p]}(\hat{x})N'_{i+1,[p]}(\hat{x})\mathrm{d}\hat{x}\right]_{i,j=1}^{n+p-2}.$$

By (3.149), we have $K_{G,n}^{[p]} = K_n^{[p]}(a_G|G'|)$. The next three steps are devoted to show that

$$\{n^{-1}K_n^{[p]}(g)\}_n \sim_{\rm GLT} g(\hat{x})f_p(\theta), \qquad \forall g\in L^1([0,1]). \tag{3.169}$$

Once this is done, the theorem is proved. Indeed, by applying (3.169) with $g=a_G|G'|$ we immediately obtain the relation $\{n^{-1}K_{G,n}^{[p]}\}_n \sim_{\rm GLT} f_{G,p}(\hat{x},\theta)$. Since $\{n^{-1}Z_{G,n}^{[p]}\}_n$ is zero-distributed by Step 1, (3.164) follows from the decomposition

$$n^{-1}A_{G,n}^{[p]} = n^{-1}K_{G,n}^{[p]} + n^{-1}Z_{G,n}^{[p]} \tag{3.170}$$

and from **GLT 3 – GLT 4**; and the singular value distribution in (3.165) follows from **GLT 1**. If $b(x)=0$ identically, then $n^{-1}A_{G,n}^{[p]}$ is symmetric and also the spectral distribution in (3.165) follows from **GLT 1**. If $b(x)$ is not identically 0, the spectral distribution in (3.165) follows from **GLT 2** applied to the decomposition (3.170), taking into account what we have seen in Step 1.

Step 3. We first prove (3.169) in the constant-coefficient case $g(\hat{x})=1$. In this case, we note that $K_n^{[p]}(1)=K_n^{[p]}$. Hence, the desired GLT relation $\{n^{-1}K_n^{[p]}(1)\}_n \sim_{\rm GLT} f_p(\theta)$ follows from (3.161) and **GLT 3 – GLT 4**, taking into account that $\{R_n^{[p]}\}_n$ is zero-distributed by **Z 1**.

Step 4. Now we prove (3.169) in the case where $g\in C([0,1])$. As in Step 4 of Section 3.3.1, the proof is based on the fact that the B-spline basis functions $N_{2,[p]},\ldots,N_{n+p-1,[p]}$ are 'locally supported'. Indeed, the width of the support $[t_{i+1},t_{i+p+2}]$ of the ith basis function $N_{i+1,[p]}$ is bounded by $(p+1)/n$ and goes to 0 as $n\to\infty$. Moreover, the support itself is located near the point $\frac{i}{n+p-2}$, because

$$\max_{\hat{x}\in[t_{i+1},t_{i+p+2}]}\left|\hat{x}-\frac{i}{n+p-2}\right| \le \frac{C_p}{n} \tag{3.171}$$

for all $i=2,\ldots,n+p-1$ and for some constant C_p depending only on p. By (3.104) and (3.171), for all $i,j=1,\ldots,n+p-2$ we have

$$\left|(K_n^{[p]}(g))_{ij} - (D_{n+p-2}(g)K_n^{[p]}(1))_{ij}\right| = \left|\int_{[0,1]}\Big[g(\hat{x})-g\Big(\frac{i}{n+p-2}\Big)\Big]N'_{j+1,[p]}(\hat{x})N'_{i+1,[p]}(\hat{x})\mathrm{d}\hat{x}\right|$$
$$\le 4p^2n^2\int_{[t_{i+1},t_{i+p+2}]}\left|g(\hat{x})-g\Big(\frac{i}{n+p-2}\Big)\right|\mathrm{d}\hat{x} \le 4p^2(p+1)n\,\omega_g\Big(\frac{C_p}{n}\Big).$$

It follows that each entry of $Z_n = n^{-1}K_n^{[p]}(g) - n^{-1}D_{n+p-2}(g)K_n^{[p]}(1)$ is bounded in modulus by $C\omega_g(1/n)$. Moreover, Z_n is banded with bandwidth at most $2p+1$, due to the local support property of the B-spline basis functions $N_{i,[p]}$. By (3.1) we conclude that $\|Z_n\| \le C\omega_g(1/n) \to 0$ as $n \to \infty$. Thus, $\{Z_n\}_n \sim_\sigma 0$, which implies (3.169) by Step 3 and **GLT 3 – GLT 4**.

Step 5. Finally, we prove (3.169) in the general case where $g \in L^1([0,1])$. By the density of $C([0,1])$ in $L^1([0,1])$, there exist continuous functions $g_m \in C([0,1])$ such that $g_m \to g$ in $L^1([0,1])$. By Step 4,

$$\{n^{-1}K_n^{[p]}(g_m)\}_n \sim_{\mathrm{GLT}} g_m(\hat{x})f_p(\theta).$$

Moreover,

$$g_m(\hat{x})f_p(\theta) \to g(\hat{x})f_p(\theta) \text{ in measure.}$$

We show that

$$\{n^{-1}K_n^{[p]}(g_m)\}_n \xrightarrow{\text{a.c.s.}} \{n^{-1}K_n^{[p]}(g)\}_n.$$

Using (3.104) and (3.4), we obtain

$$\begin{aligned}
\|K_n^{[p]}(g) - K_n^{[p]}(g_m)\|_1 &\le \sum_{i,j=1}^{n+p-2} \left|(K_n^{[p]}(g))_{ij} - (K_n^{[p]}(g_m))_{ij}\right| \\
&= \sum_{i,j=1}^{n+p-2} \left|\int_{[0,1]} \big[g(\hat{x}) - g_m(\hat{x})\big] N'_{j+1,[p]}(\hat{x}) N'_{i+1,[p]}(\hat{x})\mathrm{d}\hat{x}\right| \\
&\le \int_{[0,1]} |g(\hat{x}) - g_m(\hat{x})| \sum_{i,j=1}^{n+p-2} |N'_{j+1,[p]}(\hat{x})|\,|N'_{i+1,[p]}(\hat{x})|\mathrm{d}\hat{x} \\
&\le 4p^2n^2 \int_{[0,1]} |g(\hat{x}) - g_m(\hat{x})|\mathrm{d}\hat{x}
\end{aligned}$$

and

$$\|n^{-1}K_n^{[p]}(g) - n^{-1}K_n^{[p]}(g_m)\|_1 \le 4p^2 n \|g - g_m\|_{L^1}.$$

Thus, $\{n^{-1}K_n^{[p]}(g_m)\}_n \xrightarrow{\text{a.c.s.}} \{n^{-1}K_n^{[p]}(g)\}_n$ by **ACS 3**. The relation (3.169) now follows from **GLT 7**. □

Remark 3.11 (formal structure of the symbol). Problem (3.142) can be formally rewritten as in (3.82). If, for any $u : \overline{\Omega} \to \mathbb{R}$, we define $\hat{u} : [0,1] \to \mathbb{R}$ as in (3.88), then u satisfies (3.82) if and only if $\hat{u}$ satisfies the corresponding transformed problem (3.89), in which the higher-order operator takes the form $-a_G(\hat{x})\hat{u}''(\hat{x})$. It is then clear that, similarly to the collocation case (see Remark 3.8), even in the Galerkin case the symbol $f_{G,p}(\hat{x},\theta) = a_G(\hat{x})|G'(\hat{x})|f_p(\theta)$ preserves the formal structure of the higher-order operator associated with the transformed problem (3.89). However, in this Galerkin context we notice the appearance of the factor $|G'(\hat{x})|$, which is not present in the collocation setting; cf. (3.166) with (3.132).

Remark 3.12 (the case $p = 1$). For $p = 1$, the symbol $f_p(\theta)$ in (3.158) is given by $f_1(\theta) = 2 - 2\cos\theta$. This should not come as a surprise, because the Galerkin B-spline IgA approximation with $p = 1$ (and G equal to the identity map over $[0,1]$) coincides precisely with the linear FE approximation considered in Section 3.3.1; the only (unessential) difference is that the discretization step in Section 3.3.1 was chosen as $h = \frac{1}{n+1}$, while in this section we have $h = \frac{1}{n}$. In particular, the B-spline basis functions of degree 1, namely $N_{2,[1]}, \ldots, N_{n,[1]}$, are the hat-functions; cf. (3.98) (with $p = 1$) and (3.61).

Remark 3.13. The matrix $A_{G,n}^{[p]}$ in (3.147), which we decomposed as in (3.148), can also be decomposed as follows, according to the diffusion, convection and reaction terms:

$$A_{G,n}^{[p]} = K_{G,n}^{[p]} + H_{G,n}^{[p]} + M_{G,n}^{[p]},$$

where the diffusion, convection and reaction matrices are given by

$$K_{G,n}^{[p]} = \left[\int_{[0,1]} \frac{a(G(\hat{x}))}{|G'(\hat{x})|} N'_{j+1,[p]}(\hat{x}) N'_{i+1,[p]}(\hat{x})\mathrm{d}\hat{x}\right]_{i,j=1}^{n+p-2}, \tag{3.172}$$

$$H_{G,n}^{[p]} = \left[\int_{[0,1]} \frac{b(G(\hat{x}))|G'(\hat{x})|}{G'(\hat{x})} N'_{j+1,[p]}(\hat{x}) N_{i+1,[p]}(\hat{x})\mathrm{d}\hat{x}\right]_{i,j=1}^{n+p-2}, \tag{3.173}$$

$$M_{G,n}^{[p]} = \left[\int_{[0,1]} c(G(\hat{x}))|G'(\hat{x})| N_{j+1,[p]}(\hat{x}) N_{i+1,[p]}(\hat{x})\mathrm{d}\hat{x}\right]_{i,j=1}^{n+p-2}; \tag{3.174}$$

note that the diffusion matrix is the same as in (3.149). Let Ω be a bounded open interval of $\mathbb{R}$ and let $p \geq 1$. Then, the following results hold.

(a) Suppose
$$\frac{a(G(\hat{x}))}{|G'(\hat{x})|} \in L^1([0,1]);$$
then $\{n^{-1}K_{G,n}^{[p]}\}_n \sim_{\mathrm{GLT}} f_{G,p}$ and $\{n^{-1}K_{G,n}^{[p]}\}_n \sim_{\sigma,\lambda} f_{G,p}$, where
$$f_{G,p}(\hat{x},\theta) = \frac{a(G(\hat{x}))}{|G'(\hat{x})|} f_p(\theta) \tag{3.175}$$
and $f_p(\theta)$ is defined in (3.158); note that $f_{G,p}(\hat{x},\theta)$ is the same as in (3.166).

(b) Suppose
$$\frac{b(G(\hat{x}))|G'(\hat{x})|}{G'(\hat{x})} \in C([0,1]);$$
then $\{-\mathrm{i}H_{G,n}^{[p]}\}_n \sim_{\mathrm{GLT}} g_{G,p}$ and $\{-\mathrm{i}H_{G,n}^{[p]}\}_n \sim_{\sigma,\lambda} g_{G,p}$, where
$$g_{G,p}(\hat{x},\theta) = \frac{b(G(\hat{x}))|G'(\hat{x})|}{G'(\hat{x})} g_p(\theta) \tag{3.176}$$
and $g_p(\theta)$ is defined in (3.159).

(c) Suppose
$$c(G(\hat{x}))|G'(\hat{x})| \in L^1([0,1]);$$
then $\{nM_{G,n}^{[p]}\}_n \sim_{\mathrm{GLT}} h_{G,p}$ and $\{nM_{G,n}^{[p]}\}_n \sim_{\sigma,\lambda} h_{G,p}$, where
$$h_{G,p}(\hat{x},\theta) = c(G(\hat{x}))|G'(\hat{x})| h_p(\theta) \tag{3.177}$$
and $h_p(\theta)$ is defined in (3.160).

While the proof of (b) requires some work, the proofs of (a) and (c) can be done by following the same argument as in the proof of Theorem 3.13. The proofs of (a)–(c) can be found in [33, solution to Exercise 10.5].

3.4.3 Galerkin B-Spline IgA Discretization of Second-Order Eigenvalue Problems

Let $\mathbb{R}^+$ be the set of positive real numbers. Consider the following second-order eigenvalue problem: find eigenvalues $\lambda_j \in \mathbb{R}^+$ and eigenfunctions u_j, for $j = 1,2,\ldots,\infty$, such that
$$\begin{cases} -(a(x)u_j'(x))' = \lambda_j c(x) u_j(x), & x \in \Omega, \\ u_j(x) = 0, & x \in \partial\Omega, \end{cases} \tag{3.178}$$
where Ω is a bounded open interval of $\mathbb{R}$ and we assume that $a,c \in L^1(\Omega)$ and $a,c > 0$ a.e. in Ω. It can be shown that the eigenvalues λ_j must necessarily be real and positive. This can be formally seen by multiplying (3.178) by $u_j(x)$ and integrating over Ω:
$$\lambda_j = \frac{-\int_\Omega (a(x)u_j'(x))' u_j(x)\mathrm{d}x}{\int_\Omega c(x)(u_j(x))^2\mathrm{d}x} = \frac{\int_\Omega a(x)(u_j'(x))^2\mathrm{d}x}{\int_\Omega c(x)(u_j(x))^2\mathrm{d}x} > 0.$$

Isogeometric Galerkin approximation The weak form of (3.178) reads as follows: find eigenvalues $\lambda_j \in \mathbb{R}^+$ and eigenfunctions $u_j \in H_0^1(\Omega)$, for $j = 1,2,\ldots,\infty$, such that
$$\mathrm{a}(u_j,w) = \lambda_j (c u_j, w), \qquad \forall w \in H_0^1(\Omega),$$
where
$$\mathrm{a}(u_j,w) = \int_\Omega a(x)u_j'(x)w'(x)\mathrm{d}x,$$
$$(c u_j, w) = \int_\Omega c(x)u_j(x)w(x)\mathrm{d}x.$$

In the standard Galerkin method, we choose a finite dimensional vector space $\mathscr{W} \subset H_0^1(\Omega)$, the so-called approximation space, we let $N = \dim \mathscr{W}$ and we look for approximations of the eigenpairs (λ_j, u_j), $j = 1, 2, \ldots, \infty$, by solving the following discrete (Galerkin) problem: find $\lambda_{j,\mathscr{W}} \in \mathbb{R}^+$ and $u_{j,\mathscr{W}} \in \mathscr{W}$, for $j = 1, \ldots, N$, such that

$$\mathrm{a}(u_{j,\mathscr{W}}, w) = \lambda_{j,\mathscr{W}}(c\, u_{j,\mathscr{W}}, w), \qquad \forall w \in \mathscr{W}. \tag{3.179}$$

Assuming that both the exact and numerical eigenvalues are arranged in non-decreasing order, the pair $(\lambda_{j,\mathscr{W}}, u_{j,\mathscr{W}})$ is taken as an approximation to the pair (λ_j, u_j) for all $j = 1, 2, \ldots, N$. The numbers $\lambda_{j,\mathscr{W}}/\lambda_j - 1$, $j = 1, \ldots, N$, are referred to as the (relative) eigenvalue errors. If $\{\varphi_1, \ldots, \varphi_N\}$ is a basis of $\mathscr{W}$, we can identify each $w \in \mathscr{W}$ with its coefficient vector relative to this basis. With this identification in mind, solving the discrete problem (3.179) is equivalent to solving the generalized eigenvalue problem

$$K\mathbf{u}_{j,\mathscr{W}} = \lambda_{j,\mathscr{W}} M\mathbf{u}_{j,\mathscr{W}}, \tag{3.180}$$

where $\mathbf{u}_{j,\mathscr{W}}$ is the coefficient vector of $u_{j,\mathscr{W}}$ with respect to $\{\varphi_1, \ldots, \varphi_N\}$ and

$$K = \left[\int_\Omega a(x)\varphi_j'(x)\varphi_i'(x)\mathrm{d}x\right]_{i,j=1}^N, \tag{3.181}$$

$$M = \left[\int_\Omega c(x)\varphi_j(x)\varphi_i(x)\mathrm{d}x\right]_{i,j=1}^N. \tag{3.182}$$

The matrices K and M are referred to as the stiffness and mass matrix, respectively. Due to our assumption that $a, c > 0$ a.e., both K and M are symmetric positive definite, regardless of the chosen basis functions $\varphi_1, \ldots, \varphi_N$. Moreover, it is clear from (3.180) that the numerical eigenvalues $\lambda_{j,\mathscr{W}}$, $j = 1, \ldots, N$, are just the eigenvalues of the matrix

$$L = M^{-1}K. \tag{3.183}$$

In the isogeometric Galerkin method, we assume that the physical domain Ω is described by a global geometry function $G : [0,1] \to \overline{\Omega}$, which is invertible and satisfies $G(\partial([0,1])) = \partial\overline{\Omega}$. We fix a set of basis functions $\{\hat{\varphi}_1, \ldots, \hat{\varphi}_N\}$ defined on the reference (parametric) domain $[0,1]$ and vanishing on the boundary $\partial([0,1])$, and we find approximations to the exact eigenpairs (λ_j, u_j), $j = 1, 2, \ldots, \infty$, by using the standard Galerkin method described above, in which the approximation space is chosen as $\mathscr{W} = \mathrm{span}(\varphi_1, \ldots, \varphi_N)$, where

$$\varphi_i(x) = \hat{\varphi}_i(G^{-1}(x)) = \hat{\varphi}_i(\hat{x}), \qquad x = G(\hat{x}). \tag{3.184}$$

The resulting stiffness and mass matrices K and M are given by (3.181)–(3.182), with the basis functions φ_i defined as in (3.184). If we assume that G and $\hat{\varphi}_i$, $i = 1, \ldots, N$, are sufficiently regular, we can apply standard differential calculus to obtain for K and M the following expressions:

$$K = \left[\int_{[0,1]} \frac{a(G(\hat{x}))}{|G'(\hat{x})|}\hat{\varphi}_j'(\hat{x})\hat{\varphi}_i'(\hat{x})\mathrm{d}\hat{x}\right]_{i,j=1}^N, \tag{3.185}$$

$$M = \left[\int_{[0,1]} c(G(\hat{x}))|G'(\hat{x})|\hat{\varphi}_j(\hat{x})\hat{\varphi}_i(\hat{x})\mathrm{d}\hat{x}\right]_{i,j=1}^N. \tag{3.186}$$

GLT analysis of the Galerkin B-spline IgA discretization matrices Following the approach of Sections 3.4.1–3.4.2, we choose the basis functions $\hat{\varphi}_i$, $i = 1, \ldots, N$, as the B-splines $N_{i+1,[p]}$, $i = 1, \ldots, n+p-2$. The resulting stiffness and mass matrices (3.185)–(3.186) are given by

$$K_{G,n}^{[p]} = \left[\int_{[0,1]} \frac{a(G(\hat{x}))}{|G'(\hat{x})|}N_{j+1,[p]}'(\hat{x})N_{i+1,[p]}'(\hat{x})\mathrm{d}\hat{x}\right]_{i,j=1}^{n+p-2},$$

$$M_{G,n}^{[p]} = \left[\int_{[0,1]} c(G(\hat{x}))|G'(\hat{x})|N_{j+1,[p]}(\hat{x})N_{i+1,[p]}(\hat{x})\mathrm{d}\hat{x}\right]_{i,j=1}^{n+p-2},$$

and it is immediately seen that they are the same as the diffusion and reaction matrices in (3.172) and (3.174). The numerical eigenvalues will be henceforth denoted by $\lambda_{j,n}$, $j = 1, \ldots, n+p-2$; as noted above, they are simply the eigenvalues of the matrix

$$L_{G,n}^{[p]} = (M_{G,n}^{[p]})^{-1}K_{G,n}^{[p]}.$$

Theorem 3.14. *Let Ω be a bounded open interval of $\mathbb{R}$ and let $a,c \in L^1(\Omega)$ with $a,c>0$ a.e. Let $p\geq 1$ and let $G:[0,1]\to\overline{\Omega}$ be such that*

$$\frac{a(G(\hat{x}))}{|G'(\hat{x})|}\in L^1([0,1]).$$

Then

$$\{n^{-2}L_{G,n}^{[p]}\}_n \sim_{\rm GLT} e_{G,p}(\hat{x},\theta) \tag{3.187}$$

and

$$\{n^{-2}L_{G,n}^{[p]}\}_n \sim_{\sigma,\lambda} e_{G,p}(\hat{x},\theta), \tag{3.188}$$

where

$$e_{G,p}(\hat{x},\theta) = (h_{G,p}(\theta))^{-1} f_{G,p}(\theta) = \frac{a(G(\hat{x}))}{c(G(\hat{x}))(G'(\hat{x}))^2}\, e_p(\theta), \tag{3.189}$$

$$e_p(\theta) = (h_p(\theta))^{-1} f_p(\theta), \tag{3.190}$$

and $f_p(\theta)$, $h_p(\theta)$, $f_{G,p}(\hat{x},\theta)$, $h_{G,p}(\hat{x},\theta)$ are given by (3.158), (3.160), (3.175), (3.177)*, respectively.*

Proof. We have $a(G(\hat{x}))/|G'(\hat{x})|\in L^1([0,1])$ by assumption and $c(G(\hat{x}))|G'(\hat{x})|\in L^1([0,1])$ because $c\in L^1(\Omega)$ by assumption and

$$\int_{[0,1]} c(G(\hat{x}))|G'(\hat{x})|\mathrm{d}\hat{x} = \int_\Omega c(x)\mathrm{d}x.$$

Hence, by Remark 3.13,

$$\{n^{-1}K_{G,n}^{[p]}\}_n \sim_{\rm GLT} f_{G,p}, \qquad \{nM_{G,n}^{[p]}\}_n \sim_{\rm GLT} h_{G,p},$$

and the relations (3.187)–(3.188) follow from Theorem 3.1, taking into account that $h_{G,p}(\hat{x},\theta)\neq 0$ a.e. by our assumption that $c(x)>0$ a.e. and by the positivity of $h_p(\theta)$; see (3.141) and Remark 3.10. ☐

For $p=1,2,3,4$, Eq. (3.190) gives

$$\begin{aligned}
e_1(\theta) &= \frac{6(1-\cos\theta)}{2+\cos\theta},\\
e_2(\theta) &= \frac{20(3-2\cos\theta-\cos(2\theta))}{33+26\cos\theta+\cos(2\theta)},\\
e_3(\theta) &= \frac{42(40-15\cos\theta-24\cos(2\theta)-\cos(3\theta))}{1208+1191\cos\theta+120\cos(2\theta)+\cos(3\theta)},\\
e_4(\theta) &= \frac{72(1225-154\cos\theta-952\cos(2\theta)-118\cos(3\theta)-\cos(4\theta))}{78095+88234\cos\theta+14608\cos(2\theta)+502\cos(3\theta)+\cos(4\theta)}.
\end{aligned}$$

These equations are the analogs of formulas (117), (130), (135), (140) obtained by engineers in [40]; see also formulas (32), (33) in [18], formulas (23), (56) in [41], and formulas (23), (24) in [46]. We may therefore conclude that (3.189) is a generalization of these formulas to any degree $p\geq 1$.

Remark 3.14. Contrary to the B-spline IgA discretizations investigated herein and in [40], the authors of [18, 41, 46] considered NURBS IgA discretizations. However, the same formulas are obtained in both cases. This can be easily explained in view of the results of [32], where it is shown that the symbols f_p, g_p, h_p in (3.158)–(3.160) are exactly the same in the B-spline and NURBS IgA frameworks.

For an extension of the results obtained in this section, we refer the reader to the engineering paper [28].

4 Bibliography

[1] 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.

[2] AVRAM F. *On bilinear forms in Gaussian random variables and Toeplitz matrices.* Probab. Theory Related Fields 79 (1988) 37–45.

[3] BARBARINO G. *Equivalence between GLT sequences and measurable functions.* Linear Algebra Appl. 529 (2017) 397–412.

[4] BECKERMANN B., KUIJLAARS A. B. J. *Superlinear convergence of conjugate gradients.* SIAM J. Numer. Anal. 39 (2001) 300–329.

[5] BECKERMANN B., SERRA-CAPIZZANO S. *On the asymptotic spectrum of finite element matrix sequences.* SIAM J. Numer. Anal. 45 (2007) 746–769.

[6] BENZI M., GOLUB G. H., LIESEN J. *Numerical solution of saddle point problems.* Acta Numerica 14 (2005) 1–137.

[7] BHATIA R. *Matrix Analysis.* Springer-Verlag, New York (1997).

[8] BINI D. A., CAPOVANI M., MENCHI O. *Metodi Numerici per l'Algebra Lineare.* Zanichelli, Bologna (1988).

[9] BÖTTCHER A., GARONI C., SERRA-CAPIZZANO S. *Exploration of Toeplitz-like matrices with unbounded symbols: not a purely academic journey.* Sb. Math. (in press).

[10] BÖTTCHER A., GRUDSKY S. M. *Toeplitz Matrices, Asymptotic Linear Algebra, and Functional Analysis.* Birkhäuser Verlag, Basel–Boston–Berlin (2000).

[11] BÖTTCHER A., GRUDSKY S. M. *Spectral Properties of Banded Toeplitz Matrices.* SIAM, Philadelphia (2005).

[12] BÖTTCHER A., SILBERMANN B. *Introduction to Large Truncated Toeplitz Matrices.* Springer-Verlag, New York (1999).

[13] BÖTTCHER A., SILBERMANN B. *Analysis of Toeplitz Operators.* Second Edition, Springer-Verlag, Berlin (2006).

[14] BREZIS H. *Functional Analysis, Sobolev Spaces and Partial Differential Equations.* Springer, New York (2011).

[15] CHUI C. K. *An Introduction to Wavelets.* Academic Press (1992).

[16] CICONE A., GARONI C., SERRA-CAPIZZANO S. *Spectral and convergence analysis of the discrete ALIF method.* In preparation.

[17] COTTRELL J. A., HUGHES T. J. R., BAZILEVS Y. *Isogeometric Analysis: Toward Integration of CAD and FEA.* John Wiley & Sons, Chichester (2009).

[18] COTTRELL J. A., REALI A., BAZILEVS Y., HUGHES T. J. R. *Isogeometric analysis of structural vibrations.* Comput. Methods Appl. Mech. Engrg. 195 (2006) 5257–5296.

[19] DE BOOR C. *A Practical Guide to Splines.* Revised Edition, Springer-Verlag, New York (2001).

[20] DONATELLI M., GARONI C., MANNI C., SERRA-CAPIZZANO S., SPELEERS H. *Robust and optimal multi-iterative techniques for IgA Galerkin linear systems.* Comput. Methods Appl. Mech. Engrg. 284 (2015) 230–264.

[21] DONATELLI M., GARONI C., MANNI C., SERRA-CAPIZZANO S., SPELEERS H. *Robust and optimal multi-iterative techniques for IgA collocation linear systems.* Comput. Methods Appl. Mech. Engrg. 284 (2015) 1120–1146.

[22] DONATELLI M., GARONI C., MANNI C., SERRA-CAPIZZANO S., SPELEERS H. *Spectral analysis and spectral symbol of matrices in isogeometric collocation methods.* Math. Comp. 85 (2016) 1639–1680.

[23] 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. 55 (2017) 31–62.

[24] DONATELLI M., MAZZA M., SERRA-CAPIZZANO S. *Spectral analysis and structure preserving preconditioners for fractional diffusion equations.* J. Comput. Phys. 307 (2016) 262–279.

[25] DOROSTKAR A., NEYTCHEVA M., SERRA-CAPIZZANO S. *Spectral analysis of coupled PDEs and of their Schur complements via the notion of generalized locally Toeplitz sequences.* Technical Report 2015-008 (2015), Department of Information Technology, Uppsala University. Full text available at: http://www.it.uu.se/research/publications/reports/2015-008/.

[26] DOROSTKAR A., NEYTCHEVA M., SERRA-CAPIZZANO S. *Spectral analysis of coupled PDEs and of their Schur complements via generalized locally Toeplitz sequences in 2D.* Comput. Methods Appl. Mech. Engrg. 309 (2016) 74–105.

[27] GARONI C. *Spectral distribution of PDE discretization matrices from isogeometric analysis: the case of L^1 coefficients and non-regular geometry.* J. Spectral Theory (in press).

[28] GARONI C., HUGHES T. J. R., REALI A., SERRA-CAPIZZANO S., SPELEERS H. *Finite element and isogeometric B-spline discretizations of eigenvalue problems: symbol-based analysis.* In preparation.

[29] 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.

[30] GARONI C., MANNI C., SERRA-CAPIZZANO S., SESANA D., SPELEERS H. *Spectral analysis and spectral symbol of matrices in isogeometric Galerkin methods.* Math. Comp. 86 (2017) 1343–1373.

[31] GARONI C., MANNI C., SERRA-CAPIZZANO S., SESANA D., SPELEERS H. *Lusin theorem, GLT sequences and matrix computations: an application to the spectral analysis of PDE discretization matrices.* J. Math. Anal. Appl. 446 (2017) 365–382.

[32] GARONI C., MANNI C., SERRA-CAPIZZANO S., SPELEERS H. *NURBS versus B-splines in isogeometric discretization methods: a spectral analysis.* Submitted.

[33] GARONI C., SERRA-CAPIZZANO S. *Generalized Locally Toeplitz Sequences: Theory and Applications, Volume I.* Springer (2017).

[34] GARONI C., SERRA-CAPIZZANO S. *Generalized Locally Toeplitz Sequences: Theory and Applications.* Technical Report 2017-002, Uppsala University (2017). Preliminary version of: GARONI C., SERRA-CAPIZZANO S. *Generalized Locally Toeplitz Sequences: Theory and Applications, Volume II.* In preparation for Springer.

[35] GOLUB G. H., VAN LOAN C. F. *Matrix Computations.* Fourth Edition, The Johns Hopkins University Press, Baltimore (2013).

[36] GRENANDER U., SZEGŐ G. *Toeplitz Forms and Their Applications.* Second Edition, AMS Chelsea Publishing, New York (1984).

[37] HIGHAM N. J. *Functions of Matrices: Theory and Computation.* SIAM, Philadelphia (2008).

[38] HÖRMANDER L. *Pseudo-differential operators and non-elliptic boundary problems.* Annals of Math. 83 (1966) 129–209.

[39] HUGHES T. J. R., COTTRELL J. A., BAZILEVS Y. *Isogeometric analysis: CAD, finite elements, NURBS, exact geometry and mesh refinement.* Comput. Methods Appl. Mech. Engrg. 194 (2005) 4135–4195.

[40] HUGHES T. J. R., EVANS J. A., REALI A. *Finite element and NURBS approximations of eigenvalue, boundary-value, and initial-value problems.* Comput. Methods Appl. Mech. Engrg. 272 (2014) 290–320.

[41] HUGHES T. J. R., REALI A., SANGALLI G. *Duality and unified analysis of discrete approximations in structural dynamics and wave propagation: comparison of p-method finite elements with k-method NURBS*. Comput. Methods Appl. Mech. Engrg. 197 (2008) 4104–4124.

[42] HUGHES T. J. R., REALI A., SANGALLI G. *Efficient quadrature for NURBS-based isogeometric analysis.* Comput. Methods Appl. Mech. Engrg. 199 (2010) 301–313.

[43] KELLEY J. L. *General Topology.* Van Nostrand (1955).

[44] PARTER S. V. *On the distribution of the singular values of Toeplitz matrices.* Linear Algebra Appl. 80 (1986) 115–130.

[45] QUARTERONI A. *Numerical models for differential problems.* Second Edition, Springer-Verlag Italia, Milan (2014).

[46] REALI A. *An isogeometric analysis approach for the study of structural vibrations.* J. Earthquake Engrg. 10 (2006) 1–30.

[47] REALI A., HUGHES T. J. R. *An introduction to isogeometric collocation methods.* Chapter 4 of the Springer Book 'Isogeometric Methods for Numerical Simulation', Edited by G. Beer and S. Bordas, CISM, Udine (2015).

[48] ROMAN F., MANNI C., SPELEERS H. *Spectral analysis of matrices in Galerkin methods based on generalized B-splines with high smoothness.* Numer. Math. 135 (2017) 169–216.

[49] 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. Methods Appl. Mech. Engrg. 267 (2013) 170–232.

[50] SERRA-CAPIZZANO S. *Generalized locally Toeplitz sequences: spectral analysis and applications to discretized partial differential equations.* Linear Algebra Appl. 366 (2003) 371–402.

[51] SERRA-CAPIZZANO S. *The GLT class as a generalized Fourier analysis and applications.* Linear Algebra Appl. 419 (2006) 180–233.

[52] SERRA-CAPIZZANO S., TABLINO-POSSIO C. *Analysis of preconditioning strategies for collocation linear systems.* Linear Algebra Appl. 369 (2003) 41–75.

[53] SMITH G. D. *Numerical Solution of Partial Differential Equations: Finite Difference Methods.* Third Edition, Oxford University Press, New York (1985).

[54] SPELEERS H. *Inner products of box splines and their derivatives.* BIT Numer. Math. 55 (2015) 559–567.

[55] TILLI P. *A note on the spectral distribution of Toeplitz matrices.* Linear and Multilinear Algebra 45 (1998) 147–159.

[56] TILLI P. *Locally Toeplitz sequences: spectral properties and applications.* Linear Algebra Appl. 278 (1998) 91–120.

[57] TILLI P. *Some results on complex Toeplitz eigenvalues.* J. Math. Anal. Appl. 239 (1999) 390–401.

[58] TYRTYSHNIKOV E. E. *A unifying approach to some old and new theorems on distribution and clustering.* Linear Algebra Appl. 232 (1996) 1–43.

[59] TYRTYSHNIKOV E. E., ZAMARASHKIN N. L. *Spectra of multilevel Toeplitz matrices: advanced theory via simple matrix relationships.* Linear Algebra Appl. 270 (1998) 15–27.

[60] ZAMARASHKIN N. L., TYRTYSHNIKOV E. E. *Distribution of eigenvalues and singular values of Toeplitz matrices under weakened conditions on the generating function.* Sb. Math. 188 (1997) 1191–1201.