On the spectrum of stiffness matrices arising from isogeometric analysis

Carlo Garoni *Carla Manni* *Francesca Pelosi*
Stefano Serra-Capizzano *Hendrik Speleers*

Report TW 632, July 2013

Katholieke Universiteit Leuven
Department of Computer Science
Celestijnenlaan 200A – B-3001 Heverlee (Belgium)

On the spectrum of stiffness matrices arising from isogeometric analysis

Carlo Garoni *Carla Manni* *Francesca Pelosi*
Stefano Serra-Capizzano *Hendrik Speleers*

Report TW 632, July 2013

Department of Computer Science, K.U.Leuven

Abstract

We study the spectral properties of stiffness matrices that arise in the context of isogeometric analysis for the numerical solution of classical second order elliptic problems. Motivated by the applicative interest in the fast solution of the related linear systems, we are looking for a spectral characterization of the involved matrices. In particular, we investigate non-singularity, conditioning (extremal behavior), spectral distribution in the Weyl sense, as well as clustering of the eigenvalues to a certain (compact) subset of $\mathbb{C}$. All the analysis is related to the notion of symbol in the Toeplitz setting and is carried out both for the cases of 1D and 2D problems.

Keywords : Spectral distribution, B-splines, isogeometric analysis, elliptic problems.

Noname manuscript No.
(will be inserted by the editor)

On the spectrum of stiffness matrices arising from isogeometric analysis

**Carlo Garoni · Carla Manni · Francesca Pelosi ·
Stefano Serra-Capizzano · Hendrik Speleers**

Received: date / Accepted: date

Abstract We study the spectral properties of stiffness matrices that arise in the context of isogeometric analysis for the numerical solution of classical second order elliptic problems. Motivated by the applicative interest in the fast solution of the related linear systems, we are looking for a spectral characterization of the involved matrices. In particular, we investigate non-singularity, conditioning (extremal behavior), spectral distribution in the Weyl sense, as well as clustering of the eigenvalues to a certain (compact) subset of $\mathbb{C}$. All the analysis is related to the notion of symbol in the Toeplitz setting and is carried out both for the cases of 1D and 2D problems.

Keywords Spectral distribution · B-splines · Isogeometric analysis · Elliptic problems

1 Introduction

We consider the stiffness matrices that are encountered when approximating the solution of a classical second order elliptic problem, by using the Isogeometric Analysis (IgA) approach. More precisely, we are interested in studying

1. the eigenvalue of minimal modulus and the eigenvalue of maximal modulus,
2. the conditioning,
3. the localization of the spectrum,
4. the global behavior of the spectrum,

C. Garoni, S. Serra-Capizzano
University of Insubria, Department of Science and High Technology
Via Valleggio 11, 22100 Como, Italy
E-mail: carlo.garoni@uninsubria.it, stefano.serrac@uninsubria.it

C. Manni, F. Pelosi
University of Roma 'Tor Vergata', Department of Mathematics
Via della Ricerca Scientifica, 00133 Roma, Italy
E-mail: manni@mat.uniroma2.it, pelosi@mat.uniroma2.it

H. Speleers
University of Leuven, Department of Computer Science
Celestijnenlaan 200A, B-3001 Leuven, Belgium
E-mail: hendrik.speleers@cs.kuleuven.be

as the finesse parameter h tends to zero, and, in the case of item 2 and item 3 also for fixed h. Regarding the global behavior, we mean the asymptotic eigenvalue distribution in the sense of Weyl (see e.g. [15]), as reported in Definition 1.

The task of evaluating the asymptotic conditioning has a plain numerical motivation in understanding the numerical intrinsic difficulty of the problem, while the motivation of evaluating extremal eigenvalues and the localization of the spectrum is evident for obtaining reasonable bounds for the number of iterations when Krylov methods – such as the Conjugate Gradient (CG) in the Hermitian positive definite setting or GMRES (see [3,35,45]) – are employed. In particular, it is of paramount interest to find localization areas up to a small number of outliers, for estimating the convergence speed of such techniques (see the seminal paper by Axelsson and Lindskog [3] and subsequent results).

On the other hand, the task of finding the asymptotic eigenvalue distribution is motivated by the analysis of multigrid methods where the notion of symbol is crucial in the proof of optimality of the method [1] and by recent results on the (superlinear) convergence behavior for the CG method [5,6,7]. The CG method is a popular method for solving positive definite linear systems, and its convergence properties have been analyzed by many authors (see e.g. [3,45]). For instance, one has a simple upper bound for the CG error in energy norm in terms of the spectral condition number, that is, the ratio of the largest divided by the smallest eigenvalue, see, e.g., [35, eq. (6.106)]. In reality, the upper bound based on the condition number may be not very accurate, especially in the range of superlinear convergence of CG. This superlinear convergence behavior is observed numerically in the context of discretized elliptic problems in dimension $d \geq 2$, in particular for small step-sizes h. In this setting, the CG convergence is known to be governed by the distribution of the spectrum and has been quantified only recently in [5,6,7]. Here, for distribution of the spectrum we mean a precise limit relation reported in Definition 1. Similar results are also available for other Krylov methods, when the matrices are not Hermitian positive definite (see the book by Saad [35]): in such a case an additional actor is the conditioning of the eigenvector matrix, but all the other ingredients – such as conditioning, extremal eigenvalues, localization of the spectrum, spectral distribution results – are all important.

A discretization of our differential problem for some sequence of step-sizes h tending to zero leads to a sequence of systems of linear equations $A_m\mathbf{x}_m = \mathbf{b}_m$ with A_m some matrix of order m, where of course m depends on h, and tends to ∞ for $h \to 0$.

A very classical example of sequences of matrices having an asymptotic spectrum is given by Hermitian Toeplitz matrices $T_m(f) = [f_{j-k}]_{j,k=1,\ldots,m}$ obtained from the Fourier coefficients of the Lebesgue integrable generating function f defined over $[-\pi,\pi]$ (see for instance [15] and references therein). Here the sequence $\{T_m(f)\}$ is distributed as the symbol f and, informally speaking, this means that the eigenvalues of $T_m(f)$ behave as a sampling of f over an equi-spaced grid of $[-\pi,\pi]$, at least if f is smooth enough.

Furthermore, in the case of Finite Difference discretizations for differential operators, explicit formulas for the asymptotic spectrum have been given in [32,41,44] for the one-dimensional setting, and in [39,40] for the two-dimensional and multi-dimensional setting. Each time, the underlying symbol includes information on the coefficients and the domain of the PDE and information on the discretization schemes for the derivatives. The technique works also for Finite Elements, and with grading meshes (see [8]).

In the present paper, the matrices A_m arise from the IgA process and one might expect that the sequence of matrices $\{A_m\}$ has an asymptotic spectrum, as in the case of Finite Difference [10,22,38,41] and Finite Element [8,34] approximations: the answer is affirmative and, to our knowledge, our findings are the first concerning the spectral behavior of IgA approximations. More precisely, in our setting the matrix A_m is not Hermitian positive definite but it is close to it, at least for large m (i.e. small h), since the real part of A_m is positive definite and differs from A_m by a term of infinitesimal spectral norm as $h \to 0$. Hence, the sequences $\{A_m\}$ and $\{\mathrm{Re}A_m\}$ share the same spectral distribution symbol which is a real-valued, bounded, nonnegative function having a unique zero at zero (in analogy with the classical approaches related to Finite Differences and Finite Elements).

We finally emphasize that the analysis in this paper is a preliminary step for designing efficient preconditioners and efficient projectors, in the spirit of the theory that has been widely developed for Finite Difference and Finite Element approximations and which is heavily based on the knowledge of the symbol describing the main spectral features of the sequence $\{A_m\}$.

The paper is organized as follows. In the remaining part of the Introduction, namely Sections 1.1 and 1.2, we present the considered differential problem and the main basics on IgA methods. In Section 2 we summarize some tools for dealing with the spectral analysis of sequences of matrices. Section 3 provides the definition and some properties of cardinal B-splines. Then Section 4 is devoted to the analysis of matrices arising from the IgA discretization based on B-splines in the 1D case, and Section 5 addresses the 2D case. We characterize the spectrum in a precise way, and no difficulties are expected for treating the higher dimensional case. A final Section 6 is devoted to conclusions and future lines of research.

1.1 Problem setting

As our model problem we consider the following second order linear elliptic differential equation with constant coefficients and homogeneous Dirichlet boundary conditions:

$$\begin{cases} -\Delta u + \beta \cdot \nabla u + \gamma u = \mathrm{f}, & \text{in } \Omega, \\ u = 0, & \text{on } \partial\Omega, \end{cases} \tag{1}$$

where $\Omega \subset \mathbb{R}^d$ is a domain with Lipschitz boundary, $\mathrm{f} \in L_2(\Omega)$, $\beta \in \mathbb{R}^d$ and $\gamma \geq 0$. The weak form of problem (1) reads as follows: find $u \in \mathscr{V} := H_0^1(\Omega)$ such that

$$a(u,v) = \mathrm{F}(v), \quad \forall v \in \mathscr{V}, \tag{2}$$

where

$$a(u,v) := \int_\Omega (\nabla u \cdot \nabla v + \beta \cdot \nabla u \; v + \gamma u v)\,\mathrm{d}\Omega, \qquad \mathrm{F}(v) := \int_\Omega \mathrm{f} v \,\mathrm{d}\Omega. \tag{3}$$

There exists a unique solution u of (2), called the weak solution of (1), see e.g. [17]. In the standard Galerkin method we find an approximation of u in the following way: we choose a finite dimensional subspace $\mathscr{W} \subset \mathscr{V}$ and we look for a function $u_{\mathscr{W}} \in \mathscr{W}$ such that

$$a(u_{\mathscr{W}},v) = \mathrm{F}(v), \quad \forall v \in \mathscr{W}. \tag{4}$$

If $\dim \mathscr{W} = N$ and we fix a basis $\{\varphi_1, \dots, \varphi_N\}$ for $\mathscr{W}$, then each $v \in \mathscr{W}$ can be written as

$$v = \sum_{j=1}^{N} v_j \varphi_j,$$

and, by linearity, equation (4) is satisfied for all test functions $v \in \mathscr{W}$ if and only if it is satisfied for the basis functions $\varphi_1, \dots, \varphi_N$. Thus, the Galerkin problem (4) is equivalent to the problem of finding a vector $\mathbf{u} = [u_1 \; u_2 \; \cdots \; u_N]^T \in \mathbb{R}^N$ such that

$$A\mathbf{u} = \mathbf{f}, \tag{5}$$

where $A = [a(\varphi_j, \varphi_i)]_{i,j=1}^N \in \mathbb{R}^{N\times N}$ is the stiffness matrix and $\mathbf{f} = [\mathrm{F}(\varphi_i)]_{i=1}^N$. Once we find $\mathbf{u}$, we know $u_{\mathscr{W}} = \sum_{j=1}^N u_j\varphi_j$. It can be proved that A is a positive definite matrix in the sense that $\mathbf{v}^T A\mathbf{v} > 0$, $\forall \mathbf{v} \in \mathbb{R}^N\backslash\{\mathbf{0}\}$. In particular, A is non-singular and so there exists a unique solution $\mathbf{u}$ of (5).

In classical Finite Element Methods (FEM) the approximation space $\mathscr{W}$ is usually a space of C^0 piecewise linear polynomials vanishing at the boundary of Ω, whereas in IgA $\mathscr{W}$ is a space of polynomial splines with higher degree and higher continuity, or some of their generalizations. In this paper we are going to construct the matrix A in the case where $\mathscr{W}$ is the space spanned by B-spline functions. After the construction of A, we will study its spectral properties.

1.2 Isogeometric analysis based on B-splines

Isogeometric analysis is a recent, but well established and successful, paradigm for the analysis of problems governed by partial differential equations [20,27]. Its main goal is to improve the connection between numerical simulation and Computer Aided Design (CAD) systems.

In its original formulation, the main idea in IgA is to use directly the geometry provided by CAD systems – which is usually expressed in terms of tensor-product B-splines or their rational version, the so-called NURBS – and to approximate the unknown solutions of differential equations by the same type of functions. This results in some principal advantages of IgA with respect to classical FEM.

- Complicated geometries are represented more accurately, and some common profiles as conic sections are exactly described. This exact or accurate description of the geometry has a beneficial influence on the numerical solution of the addressed differential problem.
- The description of the geometry is incorporated exactly at the coarsest mesh level and mesh refinement does not modify the geometry. This greatly simplifies the refinement process because it eliminates any interaction with the CAD system, whereas such interaction is an unavoidable bottleneck in the classical CAD/FEM procedure.
- B-spline and NURBS representations allow an easy treatment and refinement of spaces with high approximation order and an inherent higher smoothness than those in classical FEM. This has been proved to be superior in various applications, see [20,27], and references therein.

Despite its name, the use of discretization spaces consisting of functions with high global smoothness (like tensor-product B-splines, NURBS, or some of their generalizations as T-splines, B-splines over triangulations, generalized B-splines, etc.) is as relevant as the accurate/exact description of the geometry in the context of IgA. Indeed, focusing for instance on the simpler and elegant structure of B-spline spaces, the use of B-splines of maximal smoothness allows to deal with spaces of high approximation power but lower dimension compared with standard low smoothness FEM. Moreover, the high smoothness of discretization spaces coupled with the variation diminishing property of the B-spline basis is, somehow unexpectedly, very fruitful in the numerical treatment of challenging problems as advection/reaction-dominated advective-reactive-diffusive equations and some eigenvalue problems as vibration of a finite elastic rod with fixed ends, see [20,27] and references therein. These appealing features are maintained by the above mentioned generalizations of B-splines, see [4,9,21,29,42].

Finally, the well known properties of the B-spline basis – convex partition of unity, minimum support, local linear independence, optimality of the basis, etc., see e.g. [14] – offer some relevant advantages from the numerical point of view and result in fast and robust evaluation algorithms for the basis functions and their derivatives.

Therefore, as a first step in the investigation of the properties of matrices arising in IgA, in this paper we present a detailed spectral analysis of the matrices obtained by the Galerkin method based on B-splines with equally spaced knots for problem (1) defined on the unit interval and on the unit square. Generalizations to higher dimensions are straightforward but more involved from the notational point of view. This topic has not yet been addressed in the literature. Some related results can be found in [18,23].

2 Preliminaries on spectral analysis

In this section we present the tools that will be employed in subsequent sections for performing the spectral analysis of the matrices arising from the approximation of problem (1) in the context of IgA. We will be interested in the conditioning, localization, extremal behavior, and global behavior of the spectrum in particular when the size tends to infinity: in such a setting we need to work in the framework of matrix sequences. As already recalled in the introduction, such a spectral information is important for understanding the numerical difficulty of the involved linear systems and represents a prerequisite for designing efficient preconditioners for Krylov methods and efficient projectors for multigrid techniques.

Before starting, let us introduce some notation and recall some basic results that will be used throughout this paper.

For any vector $\mathbf{x}$, the 2-norm (Euclidean norm) of $\mathbf{x}$ will be denoted by $\|\mathbf{x}\|$. Given a matrix $X \in \mathbb{C}^{m\times m}$, $\|X\|$ is the 2-norm of X, i.e. $\|X\| = \sqrt{\rho(X^*X)} = s_1(X)$, where $s_1(X)$ is the maximum singular value of X and $\rho(X)$ is the spectral radius of X. Denote by $\|X\|_1$ the trace norm of X, i.e. the sum of all the singular values of X: $\|X\|_1 = \sum_{j=1}^m s_j(X)$. Since the number of nonzero singular values of X is precisely $\mathrm{rank}(X)$, it follows that, for all $X \in \mathbb{C}^{m\times m}$, $\|X\|_1 \leq \mathrm{rank}(X)\|X\| \leq m\|X\|$. Recall that, if X is a normal matrix, i.e. $X^*X = XX^*$, then $\|X\| = \rho(X)$ and $\|X\|_1 = \sum_{j=1}^m |\lambda_j(X)|$, where $\lambda_j(X)$ is an eigenvalue of X. Note that, if X is Hermitian ($X = X^*$) or skew-Hermitian ($X = -X^*$), then X is normal. For any matrix $X \in \mathbb{C}^{m\times m}$, we will denote by $\mathrm{Re}X$ and $\mathrm{Im}X$ the real and imaginary part of X, respectively. Recall that $\mathrm{Re}X$ and $\mathrm{Im}X$ are the Hermitian matrices defined by

$$\mathrm{Re}X := \frac{X+X^*}{2}, \qquad \mathrm{Im}X := \frac{X-X^*}{2\mathrm{i}},$$

and $X = \mathrm{Re}X + \mathrm{i}\,\mathrm{Im}X$. If λ is an eigenvalue of X and $\mathbf{x} \in \mathbb{C}^m$ is a corresponding eigenvector, then, by the minimax principle [11,13], we have

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

which implies that the spectrum $\sigma(X)$ of X can be bounded as

$$\sigma(X) \subseteq [\lambda_{\min}(\mathrm{Re}X), \lambda_{\max}(\mathrm{Re}X)] \times [\lambda_{\min}(\mathrm{Im}X), \lambda_{\max}(\mathrm{Im}X)], \quad \forall X \in \mathbb{C}^{m\times m}. \tag{6}$$

Since many of the matrices appearing in Section 5 will be formed by a tensor-product of matrices defined in Section 4, we recall that, for every $X \in \mathbb{C}^{m_1 \times m_1}$ and $Y \in \mathbb{C}^{m_2 \times m_2}$, the tensor-product $X \otimes Y$ is the matrix in $\mathbb{C}^{m_1 m_2 \times m_1 m_2}$ given by:

$$X \otimes Y = \begin{bmatrix} x_{11}Y & x_{12}Y & \cdots & x_{1m_1}Y \\ x_{21}Y & x_{22}Y & \cdots & x_{2m_1}Y \\ \vdots & \vdots & & \vdots \\ x_{m_1 1}Y & x_{m_1 2}Y & \cdots & x_{m_1 m_1}Y \end{bmatrix}.$$

The next lemma, see e.g. [11], contains basic results concerning tensor-products.

Lemma 1 *Suppose that $X \in \mathbb{C}^{m_1 \times m_1}$ and $Y \in \mathbb{C}^{m_2 \times m_2}$ are normal matrices with eigenvalues given by $\lambda_1(X), \ldots, \lambda_{m_1}(X)$ and $\lambda_1(Y), \ldots, \lambda_{m_2}(Y)$. Then*

1. *$X \otimes Y$ is normal and $(X \otimes Y)^* = X^* \otimes Y^*$.*
2. *$\sigma(X \otimes Y) = \{\lambda_i(X)\lambda_j(Y) : i = 1, \ldots, m_1,\ j = 1, \ldots, m_2\}$.*
3. *$\mathrm{rank}(X \otimes Y) = \mathrm{rank}(X)\mathrm{rank}(Y)$.*
4. *$\|X \otimes Y\| = \|X\|\,\|Y\|$ and $\|X \otimes Y\|_1 = \|X\|_1\,\|Y\|_1$.*

In particular, from statements 1 and 2 it follows that if X, Y are Hermitian then $X \otimes Y$ is Hermitian, and if X, Y are Hermitian and positive definite then $X \otimes Y$ is Hermitian and positive definite.

To conclude this paragraph about notation and preliminary results, whenever $X, Y \in \mathbb{C}^{m \times m}$ are Hermitian, we write $X \geq Y$ if and only if $X - Y$ is non-negative definite.

Now we introduce the fundamental definitions for developing our spectral analysis, see [24, Definitions 1.1 and 1.2]. We denote by μ_d the Lebesgue measure in $\mathbb{R}^d$.

Definition 1 (Spectral distribution of a sequence of matrices) Let $\{X_n\}$ be a sequence of matrices with increasing dimension ($X_n \in \mathbb{C}^{d_n \times d_n}$ with $d_n < d_{n+1}$ for every n), and let $f : D \to \mathbb{C}$ be a measurable function defined on the measurable set $D \subset \mathbb{R}^d$ with $0 < \mu_d(D) < \infty$. We say that $\{X_n\}$ is distributed like f in the sense of the eigenvalues, and we write $\{X_n\} \overset{\lambda}{\sim} f$, if

$$\lim_{n\to\infty} \frac{1}{d_n} \sum_{j=1}^{d_n} F(\lambda_j(X_n)) = \frac{1}{\mu_d(D)} \int_D F(f(x_1, \ldots, x_d))\, \mathrm{d}x_1 \ldots \mathrm{d}x_d, \qquad \forall F \in C_c(\mathbb{C}, \mathbb{C}).$$

Here, $C_c(\mathbb{C}, \mathbb{C})$ is the space of continuous functions $F : \mathbb{C} \to \mathbb{C}$ with compact support.

Definition 2 (Clustering of a sequence of matrices at a subset of $\mathbb{C}$) Let $\{X_n\}$ be a sequence of matrices with increasing dimension ($X_n \in \mathbb{C}^{d_n \times d_n}$ with $d_n < d_{n+1}$ for every n), and let $S \subseteq \mathbb{C}$ be a non-empty closed subset of $\mathbb{C}$. We say that $\{X_n\}$ is strongly clustered at S if the following condition is satisfied:

$$\forall \varepsilon > 0, \quad \exists C_\varepsilon \text{ and } \exists n_\varepsilon : \quad \forall n \geq n_\varepsilon, \quad q_n(\varepsilon) \leq C_\varepsilon,$$

where $q_n(\varepsilon)$ is the number of eigenvalues of X_n lying outside the ε-expansion S_ε of S, i.e.,

$$S_\varepsilon := \bigcup_{s \in S} [\mathrm{Re}\, s - \varepsilon, \mathrm{Re}\, s + \varepsilon] \times [\mathrm{Im}\, s - \varepsilon, \mathrm{Im}\, s + \varepsilon].$$

We also recall the following results, see [24, Theorems 3.4 and 3.5].

Theorem 1 *Let $\{X_n\}$ and $\{Y_n\}$ be two sequences of matrices with $X_n, Y_n \in \mathbb{C}^{d_n \times d_n}$, and $d_n < d_{n+1}$ for all n, such that*

- *X_n is Hermitian for all n and $\{X_n\} \overset{\lambda}{\sim} f$, where $f : D \subset \mathbb{R}^d \to \mathbb{R}$ is a measurable function defined on the measurable set D with $0 < \mu_d(D) < \infty$;*
- *there exists a constant C so that $\|X_n\|, \|Y_n\| \le C$ for all n;*
- *$\|Y_n\|_1 = o(d_n)$ as $n \to \infty$, i.e., $\lim_{n\to\infty} \frac{\|Y_n\|_1}{d_n} = 0$.*

Set $Z_n := X_n + Y_n$. Then $\{Z_n\} \overset{\lambda}{\sim} f$.

Theorem 2 *Let $\{X_n\}$ and $\{Y_n\}$ be two sequences of matrices with $X_n, Y_n \in \mathbb{C}^{d_n \times d_n}$, and $d_n < d_{n+1}$ for all n, such that*

- *X_n is Hermitian for all n and $\{X_n\} \overset{\lambda}{\sim} f$, where $f : D \subset \mathbb{R}^d \to \mathbb{R}$ is a measurable function defined on the measurable set D with $0 < \mu_d(D) < \infty$;*
- *there exists a constant C so that $\|X_n\|, \|Y_n\|_1 \le C$ for all n.*

Set $Z_n := X_n + Y_n$. Then $\{Z_n\} \overset{\lambda}{\sim} f$. Moreover, $\{Z_n\}$ is strongly clustered at the essential range of f.[1]

A (one-level) Toeplitz matrix is a square matrix whose entries are constant along each diagonal. Given a (univariate) function $f : [-\pi, \pi] \to \mathbb{R}$ belonging to $L_1([-\pi, \pi])$, we can associate to f a family (sequence) of Hermitian Toeplitz matrices $\{T_m(f)\}$ parameterized by the integer index m and defined for all $m \ge 1$ in the following way:

$$T_m(f) := \begin{bmatrix} f_0 & f_{-1} & \cdots & \cdots & f_{-(m-1)} \\ f_1 & \ddots & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & f_{-1} \\ f_{m-1} & \cdots & \cdots & f_1 & f_0 \end{bmatrix} \in \mathbb{C}^{m \times m},$$

where

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

are the Fourier coefficients of f. The next theorem is one of the most important results concerning sequences of Toeplitz matrices. In particular, the third statement in the theorem was originally proved by Szegö [25], see also [43] for a generalization.

Theorem 3 (Szegö) *Let $f \in L_1([-\pi, \pi])$ be a real-valued function, and let $m_f := \operatorname{ess\,inf} f$, $M_f := \operatorname{ess\,sup} f$, and suppose $m_f < M_f$. Then*

- *$\sigma(T_m(f)) \subset (m_f, M_f), \quad \forall m \ge 1$;*
- *$\lambda_{\min}(T_m(f)) \searrow m_f$ and $\lambda_{\max}(T_m(f)) \nearrow M_f$ as $m \to \infty$;*
- *$\{T_m(f)\} \overset{\lambda}{\sim} f$, that is*

$$\lim_{m\to\infty} \frac{1}{m} \sum_{j=1}^{m} F(\lambda_j(T_m(f))) = \frac{1}{2\pi} \int_{-\pi}^{\pi} F(f(\theta)) \,\mathrm{d}\theta, \qquad \forall F \in C_c(\mathbb{C}, \mathbb{C}).$$

[1] The essential range of f coincides exactly with the range of f whenever f is continuous. In this paper we will only deal with continuous functions f and the application of Theorem 2 will not involve any complication.

Another result concerns the asymptotics of the j-th smallest eigenvalue $\lambda_j(T_m(f))$, for j fixed and $m \to \infty$. This result is due to Parter [30], see also [31] for a generalization.

Theorem 4 (Parter) *Let $f : \mathbb{R} \to \mathbb{R}$ be continuous and 2π-periodic. Let $m_f := \min_{\theta\in\mathbb{R}} f(\theta) = f(\theta_{\min})$ and let $\theta_{\min}$ be the unique point in $(-\pi,\pi]$ such that $f(\theta_{\min}) = m_f$. Assume there exists $s \geq 1$ such that f has $2s$ continuous derivatives in $(\theta_{\min}-\varepsilon, \theta_{\min}+\varepsilon)$ for some $\varepsilon > 0$ and $f^{(2s)}(\theta_{\min}) > 0$ is the first non-vanishing derivative of f at $\theta_{\min}$. Finally, for every $m \geq 1$, let $\lambda_1(T_m(f)) \leq \ldots \leq \lambda_m(T_m(f))$ be the eigenvalues of $T_m(f)$ arranged in increasing order. Then, for each fixed $j \geq 1$,*

$$\lambda_j(T_m(f)) - m_f \overset{m\to\infty}{\sim} c_{s,j}\,\frac{f^{(2s)}(\theta_{\min})}{(2s)!}\,\frac{1}{m^{2s}}, \tag{7}$$

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

Remark 1 The constant $c_{s,j}$ is the j-th smallest eigenvalue of the boundary value problem

$$\begin{cases} (-1)^s u^{(2s)}(x) = \mathrm{f}(x), \quad \text{for} \quad 0 < x < 1, \\ u(0) = u'(0) = \ldots = u^{(s-1)}(0) = 0. \quad u(1) = u'(1) = \ldots = u^{(s-1)}(1) = 0, \end{cases} \tag{8}$$

see [30, p. 191]. This means that $c_{s,j}$ is the j-th smallest number satisfying $(-1)^s u^{(2s)}(x) = c_{s,j}\,u(x)$ for some (nonzero) function u belonging to an 'appropriate functional space' associated with (8). In particular, $c_{s,1}$ is the minimum eigenvalue of (8). The sequence $\{c_{s,1}\}$ was investigated in [16], where it was shown that the numbers $c_{1,1}, c_{2,1}, c_{3,1}, \ldots$ appear in many situations and the following asymptotic formula holds:

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

Remark 2 When $s = 1$, the boundary value problem (8) becomes

$$\begin{cases} -u''(x) = \mathrm{f}(x), \quad 0 < x < 1, \\ u(0) = u(1) = 0, \end{cases} \tag{9}$$

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

$$-\frac{d^2}{dx^2} : H_0^2([0,1]) \subset L_2([0,1]) \to L_2([0,1]). \tag{10}$$

The mentioned 'appropriate functional space' is in this case $H_0^2([0,1])$. The eigenvalues of (10) are $j^2\pi^2$, $j = 1,2,\ldots$, and an eigenfunction corresponding to the j-th eigenvalue $j^2\pi^2$ is $u_j(x) = \sin(j\pi x)$: $-u_j''(x) = j^2\pi^2 u_j(x)$. Thus, by Remark 1, we find that $c_{1,j} = j^2\pi^2$ for all $j \geq 1$.

Remark 3 Parter's theorem applies to the function $f(\theta) = (2-2\cos\theta)^s$, $s \geq 1$. Indeed, it can be proved that this function satisfies all the hypotheses of Theorem 4 with $m_f = 0$, $\theta_{\min} = 0$, and the number s appearing in Theorem 4 being exactly the exponent s in the definition of f. Moreover, $f^{(2s)}(\theta_{\min}) = (2s)!$. Therefore, by (7) we obtain that, for each fixed $j \geq 1$,

$$\lambda_j(T_m((2-2\cos\theta)^s)) \overset{m\to\infty}{\sim} \frac{c_{s,j}}{m^{2s}}.$$

On the other hand, by using Theorem 13 below, for the case $s = 1$ we get

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

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

In view of Section 5, it is also important to recall some properties of two-level Toeplitz matrices. Given a bivariate function $g : [-\pi,\pi]^2 \to \mathbb{R}$ belonging to $L_1([-\pi,\pi]^2)$, we can associate to g a family of two-level Hermitian Toeplitz matrices $\{T_{m_1,m_2}(g)\}$ parameterized by two integer indices m_1, m_2 and defined for all $m_1, m_2 \geq 1$ in the following way:

$$T_{m_1,m_2}(g) := \begin{bmatrix} G_0 & G_{-1} & \cdots & \cdots & G_{-(m_1-1)} \\ G_1 & \ddots & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & G_{-1} \\ G_{m_1-1} & \cdots & \cdots & G_1 & G_0 \end{bmatrix} \in \mathbb{C}^{m_1m_2 \times m_1m_2},$$

where for every $k \in \mathbb{Z}$,

$$G_k := \begin{bmatrix} g_{k,0} & g_{k,-1} & \cdots & \cdots & g_{k,-(m_2-1)} \\ g_{k,1} & \ddots & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & g_{k,-1} \\ g_{k,m_2-1} & \cdots & \cdots & g_{k,1} & g_{k,0} \end{bmatrix} \in \mathbb{C}^{m_2 \times m_2},$$

and for every $k, l \in \mathbb{Z}$,

$$g_{k,l} := \frac{1}{(2\pi)^2}\int_{-\pi}^{\pi}\int_{-\pi}^{\pi} g(\theta_1,\theta_2)\mathrm{e}^{-\mathrm{i}(k\theta_1+l\theta_2)}\,\mathrm{d}\theta_1\mathrm{d}\theta_2$$

is the (k,l) Fourier coefficient of g. For sequences of two-level Hermitian Toeplitz matrices we have the following classical theorem analogous to Theorem 3, see [37] (and references therein) and again [43] for the distribution results.

Theorem 5 *Let $g \in L_1([-\pi,\pi]^2)$ be a real-valued function, and let $m_g := \operatorname{ess\,inf} g$, $M_g := \operatorname{ess\,sup} g$, and suppose $m_g < M_g$. Then*

- *$\sigma(T_{m_1,m_2}(g)) \subset (m_g, M_g)$, $\quad \forall m_1, m_2 \geq 1$;*
- *it holds that $\forall F \in C_c(\mathbb{C},\mathbb{C})$,*

$$\lim_{\substack{m_1\to\infty \\ m_2\to\infty}} \frac{1}{m_1m_2}\sum_{j=1}^{m_1m_2} F(\lambda_j(T_{m_1,m_2}(g))) = \frac{1}{(2\pi)^2}\int_{-\pi}^{\pi}\int_{-\pi}^{\pi} F(g(\theta_1,\theta_2))\,\mathrm{d}\theta_1\mathrm{d}\theta_2.$$

The last result relates tensor-products and Toeplitz matrices. Observe that, given two (univariate) functions $f,h:[-\pi,\pi]\to\mathbb{R}$ in $L_1([-\pi,\pi])$, we can construct the (bivariate) tensor-product function

$$f\otimes h:[-\pi,\pi]^2\to\mathbb{R},\quad (f\otimes h)(\theta_1,\theta_2):=f(\theta_1)h(\theta_2),$$

which belongs to $L_1([-\pi,\pi]^2)$. Hence, we can consider the three families of Hermitian Toeplitz matrices $\{T_{m_1}(f)\}$, $\{T_{m_2}(h)\}$ and $\{T_{m_1,m_2}(f\otimes h)\}$. A direct computation gives the following result.

Lemma 2 *Let $f,h\in L_1([-\pi,\pi])$ be real-valued functions. Then, for all $m_1,m_2\geq 1$,*

$$T_{m_1}(f)\otimes T_{m_2}(h)=T_{m_1,m_2}(f\otimes h).$$

3 Cardinal B-splines

Let us denote by $\phi_{[p]}$ the cardinal B-spline of degree p over the uniform knot sequence $\{0,1,\ldots,p+1\}$, which is defined recursively as follows [14]:

$$\phi_{[0]}(t):=\begin{cases}1, & \text{if } t\in[0,1),\\ 0, & \text{elsewhere,}\end{cases}\tag{11}$$

and

$$\phi_{[p]}(t):=\frac{t}{p}\phi_{[p-1]}(t)+\frac{p+1-t}{p}\phi_{[p-1]}(t-1),\quad p\geq 1.\tag{12}$$

The cardinal B-spline can also be expressed in terms of truncated powers [14],

$$\phi_{[p]}(t)=\frac{1}{p!}\sum_{i=0}^{p+1}(-1)^i\binom{p+1}{i}(t-i)_+^p,\tag{13}$$

where $(t)_+^r:=(\max(t,0))^r$. As usual in the literature, we will refer to cardinal B-splines of degree p as the set of integer translates of $\phi_{[p]}$, that is $\{\phi_{[p]}(\cdot-k),\ k\in\mathbb{Z}\}$. In the next subsections we collect some properties of cardinal B-splines and their Fourier transform that will be useful later on.

3.1 Properties of cardinal B-splines

Denoting by $\mathbb{P}_p$ the space of algebraic polynomials of degree less than or equal to p, it turns out that the cardinal B-spline $\phi_{[p]}$ belongs piecewisely to $\mathbb{P}_p$ and it is globally of class C^{p-1}.

It is well known that the cardinal B-spline possesses some fundamental properties. Some of them are briefly summarized below, see [14, 19].

- *Positivity:*
$$\phi_{[p]}(t)\geq 0,\quad t\in\mathbb{R}.$$
- *Minimal support:*
$$\phi_{[p]}(t)=0,\quad t\notin[0,p+1].\tag{14}$$
- *Symmetry:*
$$\phi_{[p]}\left(\frac{p+1}{2}+t\right)=\phi_{[p]}\left(\frac{p+1}{2}-t\right).\tag{15}$$

- *Partition of unity:*
$$\sum_{k\in\mathbb{Z}} \phi_{[p]}(t-k) = 1, \tag{16}$$
which gives in combination with the local support and smoothness,
$$\sum_{k=1}^{p} \phi_{[p]}(k) = 1, \quad p \geq 1. \tag{17}$$
- *Recurrence relation for derivatives:*
$$\dot{\phi}_{[p]}(t) = \phi_{[p-1]}(t) - \phi_{[p-1]}(t-1), \tag{18}$$
$$\phi_{[p]}^{(r)}(t) = \phi_{[p-1]}^{(r-1)}(t) - \phi_{[p-1]}^{(r-1)}(t-1), \tag{19}$$
where $\dot{\phi}_{[p]}(t)$ denotes the derivative of $\phi_{[p]}(t)$ with respect to its argument t.
- *Unimodal behavior:*
$$\dot{\phi}_{[p]}(t) = 0, \ \ t \in (0, p+1) \quad \text{if and only if} \quad t = \frac{p+1}{2}, \qquad p \geq 2. \tag{20}$$
- *Convolution relation:*
$$\phi_{[p]}(t) = (\phi_{[p-1]} * \phi_{[0]})(t) := \int_{\mathbb{R}} \phi_{[p-1]}(t-s)\phi_{[0]}(s)\,\mathrm{d}s = \int_0^1 \phi_{[p-1]}(t-s)\,\mathrm{d}s. \tag{21}$$

In the remaining of the subsection we derive from the previous properties some results that are needed later on. The next lemma generalizes the symmetry property to derivatives of any order of the cardinal B-splines.

Lemma 3 *Let $\phi_{[p]}$ be the cardinal B-spline as defined in (11)–(12), then*
$$\phi_{[p]}^{(r)}\left(\frac{p+1}{2}+t\right) = (-1)^r \phi_{[p]}^{(r)}\left(\frac{p+1}{2}-t\right).$$

Proof We prove this by induction on the order of derivatives. The base case ($r = 0$) is just the symmetry property (15). As inductive step we increase the order of derivative by one, i.e., $r \to r+1$. Using the recurrence relation for derivatives (19) and the induction hypothesis, we have
$$\begin{aligned}
\phi_{[p]}^{(r+1)}\left(\frac{p+1}{2}+t\right) &= \phi_{[p-1]}^{(r)}\left(\frac{p+1}{2}+t\right) - \phi_{[p-1]}^{(r)}\left(\frac{p+1}{2}+t-1\right) \\
&= (-1)^r \left(\phi_{[p-1]}^{(r)}\left(\frac{p+1}{2}-t-1\right) - \phi_{[p-1]}^{(r)}\left(\frac{p+1}{2}-t\right)\right) \\
&= (-1)^{r+1} \phi_{[p]}^{(r+1)}\left(\frac{p+1}{2}-t\right).
\end{aligned}$$
□

The following lemma provides an expression for inner products of the cardinal B-spline and its integer translates. Similar results for derivatives will be provided in Lemma 5.

Lemma 4 *Let $\phi_{[p]}$ be the cardinal B-spline as defined in (11)–(12), then*
$$\int_{\mathbb{R}} \phi_{[p_1]}(t)\phi_{[p_2]}(t+k)\,\mathrm{d}t = \phi_{[p_1+p_2+1]}(p_1+1+k) = \phi_{[p_1+p_2+1]}(p_2+1-k). \tag{22}$$

Proof Using the convolution relation of cardinal B-splines (21), we obtain

$$\begin{aligned}\phi_{[p_1+p_2+1]}(p_2+1-k) &= \int_0^1 \phi_{[p_1+p_2]}(p_2+1-k-t_1)\,\mathrm{d}t_1\\ &= \int_0^1 \dots \int_0^1 \phi_{[p_2]}(p_2+1-k-(t_1+t_2+\ldots+t_{p_1+1}))\,\mathrm{d}t_1\ldots\mathrm{d}t_{p_1+1}.\end{aligned}$$

From [19, p. 85] we know that for every continuous function f it holds

$$\int_{\mathbb{R}} f(t)\phi_{[p]}(t)\,\mathrm{d}t = \int_0^1 \dots \int_0^1 f(t_1+t_2+\ldots+t_{p+1})\,\mathrm{d}t_1\ldots\mathrm{d}t_{p+1}, \tag{23}$$

and hence

$$\phi_{[p_1+p_2+1]}(p_2+1-k) = \int_{\mathbb{R}} \phi_{[p_2]}(p_2+1-k-t)\phi_{[p_1]}(t)\,\mathrm{d}t.$$

By symmetry of the cardinal B-splines, see (15), we have

$$\phi_{[p_2]}(p_2+1-k-t) = \phi_{[p_2]}(k+t),$$

resulting in

$$\phi_{[p_1+p_2+1]}(p_2+1-k) = \int_{\mathbb{R}} \phi_{[p_2]}(k+t)\phi_{[p_1]}(t)\,\mathrm{d}t.$$

In addition, again by symmetry of the cardinal B-splines, we obtain

$$\phi_{[p_1+p_2+1]}(p_1+1+k) = \phi_{[p_1+p_2+1]}(p_2+1-k),$$

which completes the proof. □

Lemma 5 *Let $\phi_{[p]}$ be the cardinal B-spline as defined in (11)–(12), then*

$$\int_{\mathbb{R}} \phi_{[p_1]}^{(r)}(t)\,\phi_{[p_2]}^{(s)}(t+k)\,\mathrm{d}t = (-1)^r\,\phi_{[p_1+p_2+1]}^{(r+s)}(p_1+1+k) = (-1)^s\,\phi_{[p_1+p_2+1]}^{(r+s)}(p_2+1-k). \tag{24}$$

Proof Because of the (anti-)symmetry of the higher order derivatives of the B-splines given by Lemma 3, we have

$$\begin{aligned}(-1)^r\,\phi_{[p_1+p_2+1]}^{(r+s)}(p_1+1+k) &= (-1)^r\,\phi_{[p_1+p_2+1]}^{(r+s)}\left(\frac{p_1+p_2+2}{2}+\frac{p_1-p_2}{2}+k\right)\\ &= (-1)^r(-1)^{r+s}\,\phi_{[p_1+p_2+1]}^{(r+s)}\left(\frac{p_1+p_2+2}{2}-\frac{p_1-p_2}{2}-k\right)\\ &= (-1)^s\,\phi_{[p_1+p_2+1]}^{(r+s)}(p_2+1-k).\end{aligned}$$

So, we only have to show one of both equalities in (24). This can be proven by induction on the order of derivatives. The base case ($r=s=0$) simply follows from Lemma 4. We consider two inductive steps: in the first inductive step we increase the order of derivative of $\phi_{[p_1]}$ by one, i.e., $r \to r+1$, and in the second inductive step we increase the order of derivative of $\phi_{[p_2]}$ by one, i.e., $s \to s+1$.

1. $(r \to r+1)$. Using (19) and the induction hypothesis, we have

$$\begin{aligned}\int_{\mathbb{R}} \phi_{[p_1]}^{(r+1)}(t)\,\phi_{[p_2]}^{(s)}(t+k)\,\mathrm{d}t &= \int_{\mathbb{R}} \left(\phi_{[p_1-1]}^{(r)}(t) - \phi_{[p_1-1]}^{(r)}(t-1)\right)\,\phi_{[p_2]}^{(s)}(t+k)\,\mathrm{d}t \\ &= \int_{\mathbb{R}} \phi_{[p_1-1]}^{(r)}(t)\phi_{[p_2]}^{(s)}(t+k)\,\mathrm{d}t - \int_{\mathbb{R}} \phi_{[p_1-1]}^{(r)}(t-1)\phi_{[p_2]}^{(s)}(t+k)\,\mathrm{d}t \\ &= \int_{\mathbb{R}} \phi_{[p_1-1]}^{(r)}(t)\phi_{[p_2]}^{(s)}(t+k)\,\mathrm{d}t - \int_{\mathbb{R}} \phi_{[p_1-1]}^{(r)}(t)\phi_{[p_2]}^{(s)}(t+k+1)\,\mathrm{d}t \\ &= (-1)^r \left(\phi_{[p_1+p_2]}^{(r+s)}(p_1+k) - \phi_{[p_1+p_2]}^{(r+s)}(p_1+1+k)\right) \\ &= (-1)^{r+1}\,\phi_{[p_1+p_2+1]}^{(r+s+1)}(p_1+1+k).\end{aligned}$$

2. $(s \to s+1)$. This inductive step can be proven in a completely analogous way as the first inductive step.

□

Finally, we provide some relations about second derivatives of cardinal B-splines.

Lemma 6 *Let $\phi_{[p]}$ be the cardinal B-spline as defined in (11)–(12), and let $\dot\phi_{[p]}$ and $\ddot\phi_{[p]}$ be its first and second derivative, respectively, then*

$$\sum_{k=1}^{p} \ddot\phi_{[2p+1]}(p+1-k) = \dot\phi_{[2p]}(p) = -\frac{1}{2}\ddot\phi_{[2p+1]}(p+1),$$

$$\sum_{k=1}^{p} k^2\,\ddot\phi_{[2p+1]}(p+1-k) = 1.$$

Proof We first note that by (19), (15) and (20), we have

$$-\ddot\phi_{[2p+1]}(p+1) = -2\dot\phi_{[2p]}(p+1) = 2\dot\phi_{[2p]}(p) > 0. \tag{25}$$

Using (18)–(19) and $\phi_{[2p-1]}(-1) = \phi_{[2p-1]}(0) = 0$, we obtain that

$$\begin{aligned}\sum_{k=1}^{p} \ddot\phi_{[2p+1]}(p+1-k) &= \sum_{k=1}^{p} \left(\phi_{[2p-1]}(p+1-k) - 2\phi_{[2p-1]}(p-k) + \phi_{[2p-1]}(p-1-k)\right) \\ &= \phi_{[2p-1]}(p) - \phi_{[2p-1]}(p-1) = \dot\phi_{[2p]}(p).\end{aligned}$$

In a similar way, taking into account that

$$k^2 - 2(k+1)^2 + (k+2)^2 = 2, \quad k \ge 0,$$

we find that

$$\begin{aligned}\sum_{k=1}^{p} k^2\,\ddot\phi_{[2p+1]}(p+1-k) &= \sum_{k=1}^{p} k^2\left(\phi_{[2p-1]}(p+1-k) - 2\phi_{[2p-1]}(p-k) + \phi_{[2p-1]}(p-1-k)\right) \\ &= \phi_{[2p-1]}(p) + 2\sum_{k=2}^{p} \phi_{[2p-1]}(p+1-k) \\ &= \sum_{k=-p+2}^{p} \phi_{[2p-1]}(p+1-k) = \sum_{k=1}^{2p-1} \phi_{[2p-1]}(k) = 1.\end{aligned}$$

The last equalities follow from the symmetry property (15) and the partition of unity property (17) of cardinal B-splines. □

3.2 Fourier transform

In this subsection we will address some relations between inner products of cardinal B-splines, and the Fourier transform of the cardinal B-spline.

We first recall the following result, see [19, Theorem 2.28].

Theorem 6 *Let $\psi \in L_2(\mathbb{R})$ and its Fourier transform $\widehat{\psi}$ satisfy*

$$\psi(t) = O(|t|^{-a}), \quad a > 1, \quad \text{as } |t| \to \infty, \tag{26}$$

and

$$\widehat{\psi}(\theta) = O(|\theta|^{-b}), \quad b > \frac{1}{2}, \quad \text{as } |\theta| \to \infty. \tag{27}$$

Then,

$$\sum_{k\in\mathbb{Z}} \left(\int_{\mathbb{R}} \psi(t-k)\overline{\psi(t)}\,\mathrm{d}t \right) \mathrm{e}^{\mathrm{i}(k\theta)} = \sum_{k\in\mathbb{Z}} |\widehat{\psi}(\theta+2k\pi)|^2, \quad \forall\theta \in [-\pi,\pi]. \tag{28}$$

By using the convolution relation (21) one can easily obtain a simple expression for the Fourier transform of the cardinal B-spline $\phi_{[p]}$, see [19, p. 56]:

$$\widehat{\phi_{[p]}}(\theta) = \left(\frac{1-\mathrm{e}^{-\mathrm{i}\,\theta}}{\mathrm{i}\,\theta} \right)^{p+1}, \tag{29}$$

so that

$$\left| \widehat{\phi_{[p]}}(\theta) \right|^2 = \left(\frac{2-2\cos\theta}{\theta^2} \right)^{p+1}. \tag{30}$$

From (14) and (29) it follows that the cardinal B-spline satisfies the conditions (26)–(27). So, when using the cardinal B-spline of degree p as the function ψ in Theorem 6, we can express the right-hand side in (28) by means of (30). This implies

$$\sum_{k\in\mathbb{Z}} \left| \widehat{\phi_{[p]}}(\theta+2k\pi) \right|^2 \geq \left| \widehat{\phi_{[p]}}(\theta) \right|^2 = \left(\frac{2-2\cos\theta}{\theta^2} \right)^{p+1} \geq \left(\frac{4}{\pi^2} \right)^{p+1}, \quad \theta \in [-\pi,\pi]. \tag{31}$$

A sharper lower bound can be found in [19, p. 89]. It is formulated in terms of the roots of the so-called Euler-Frobenius polynomials of degree $2p$, but these roots are not provided in a closed form expression. On the other hand, to obtain an upper bound, we make use of relation (22) and of the partition of unity property (17). In this way, we obtain

$$\sum_{k\in\mathbb{Z}} \left| \widehat{\phi_{[p]}}(\theta+2k\pi) \right|^2 = \sum_{k\in\mathbb{Z}} \phi_{[2p+1]}(p+1-k)\,\mathrm{e}^{\mathrm{i}(k\theta)} \leq \sum_{k\in\mathbb{Z}} \phi_{[2p+1]}(p+1-k)\,|\mathrm{e}^{\mathrm{i}(k\theta)}| = 1. \tag{32}$$

Note that for the cardinal B-spline of degree p the left-hand side in (28) is a finite sum consisting of $2p+1$ terms.

The next two lemmas provide some properties of the functions associated to certain Toeplitz matrices that we will investigate later on.

Lemma 7 *Let $p \geq 1$, and let $f_p : [-\pi,\pi] \to \mathbb{R}$,*

$$f_p(\theta) := -\ddot{\phi}_{[2p+1]}(p+1) - 2\sum_{k=1}^{p} \ddot{\phi}_{[2p+1]}(p+1-k)\cos(k\theta), \tag{33}$$

and $M_{f_p} := \max_{\theta\in[-\pi,\pi]} f_p(\theta)$. Then the following properties hold.

1. $\forall\theta\in[-\pi,\pi]$,

$$f_p(\theta)=(2-2\cos\theta)\sum_{k\in\mathbb{Z}}\left|\widehat{\phi_{[p-1]}}(\theta+2k\pi)\right|^2, \tag{34}$$

and

$$(2-2\cos\theta)\left(\frac{4}{\pi^2}\right)^p\le f_p(\theta)\le\min\left(2-2\cos\theta,(2-2\cos\theta)^{p+1}\left(\frac{1}{\theta^{2p}}+\frac{1}{6\pi^{2p-2}}\right)\right).$$

2. $\min_{\theta\in[-\pi,\pi]}f_p(\theta)=f_p(0)=0$*, and* $\theta=0$ *is the unique zero of* f_p *over* $[-\pi,\pi]$*, and*

$$M_{f_p}\le\min\left(4,\frac{8}{p+1}+\frac{2\pi^2}{3}\left(\frac{4}{\pi^2}\right)^p,2\dot{\phi}_{[2p]}(p)+2\sum_{k=1}^{p}\left|\ddot{\phi}_{[2p+1]}(p+1-k)\right|\right).$$

In particular, $M_{f_p}\to 0$ *as* $p\to\infty$*.*

Proof Using the recurrence relation for derivatives (18), for every $\theta\in[-\pi,\pi]$ we obtain that

$$\widehat{\dot{\phi}_{[p]}}(\theta)=\left(1-\mathrm{e}^{-\mathrm{i}\,\theta}\right)\widehat{\phi_{[p-1]}}(\theta),$$

and

$$\left|\widehat{\dot{\phi}_{[p]}}(\theta)\right|^2=(2-2\cos\theta)\left|\widehat{\phi_{[p-1]}}(\theta)\right|^2.$$

This implies that

$$\sum_{k\in\mathbb{Z}}\left|\widehat{\dot{\phi}_{[p]}}(\theta+2k\pi)\right|^2=(2-2\cos\theta)\sum_{k\in\mathbb{Z}}\left|\widehat{\phi_{[p-1]}}(\theta+2k\pi)\right|^2. \tag{35}$$

The equality (34) follows from relation (24), Theorem 6 and (35) in the following way:

$$\begin{aligned}f_p(\theta)&=\sum_{k\in\mathbb{Z}}-\ddot{\phi}_{[2p+1]}(p+1-k)\mathrm{e}^{\mathrm{i}(k\theta)}=\sum_{k\in\mathbb{Z}}\left(\int_{\mathbb{R}}\dot{\phi}_{[p]}(t)\dot{\phi}_{[p]}(t-k)\mathrm{d}t\right)\mathrm{e}^{\mathrm{i}(k\theta)}\\&=\sum_{k\in\mathbb{Z}}\left|\widehat{\dot{\phi}_{[p]}}(\theta+2k\pi)\right|^2=(2-2\cos\theta)\sum_{k\in\mathbb{Z}}\left|\widehat{\phi_{[p-1]}}(\theta+2k\pi)\right|^2.\end{aligned}$$

From (34) and from the inequalities (31)–(32), we get

$$(2-2\cos\theta)\left(\frac{4}{\pi^2}\right)^p\le f_p(\theta)\le 2-2\cos\theta,\quad\forall\theta\in[-\pi,\pi]. \tag{36}$$

Furthermore, using (30) in the expression of f_p given by (34), we obtain that

$$f_p(\theta)=(2-2\cos\theta)\sum_{k\in\mathbb{Z}}\left(\frac{2-2\cos(\theta+2k\pi)}{(\theta+2k\pi)^2}\right)^p=(2-2\cos\theta)^{p+1}\sum_{k\in\mathbb{Z}}\frac{1}{(\theta+2k\pi)^{2p}}. \tag{37}$$

Note that for $\theta\in[0,\pi]$

$$\begin{aligned}\sum_{k\in\mathbb{Z}}\frac{1}{(\theta+2k\pi)^{2p}}&=\frac{1}{\theta^{2p}}+\sum_{k=1}^{\infty}\frac{1}{(\theta+2k\pi)^{2p}}+\sum_{k=1}^{\infty}\frac{1}{(-\theta+2k\pi)^{2p}}\\&\le\frac{1}{\theta^{2p}}+\sum_{k=1}^{\infty}\frac{1}{(2k\pi)^{2p}}+\sum_{k=1}^{\infty}\frac{1}{(-\pi+2k\pi)^{2p}}\\&\le\frac{1}{\theta^{2p}}+\frac{1}{\pi^{2p}}\left(\sum_{k=1}^{\infty}\frac{1}{(2k)^2}+\sum_{k=1}^{\infty}\frac{1}{(2k-1)^2}\right)=\frac{1}{\theta^{2p}}+\frac{1}{6\pi^{2p-2}},\end{aligned}$$

and the same bound holds for $\theta \in [-\pi, 0]$ because of the symmetry. By (37), the latter inequality yields

$$f_p(\theta) \leq (2-2\cos\theta)^{p+1}\left(\frac{1}{\theta^{2p}}+\frac{1}{6\pi^{2p-2}}\right), \quad \forall\theta \in [-\pi,\pi]. \tag{38}$$

This proves the first statement in the theorem.

We now prove the second statement. The inequalities in (36) imply that $\min_{\theta\in[-\pi,\pi]} f_p(\theta) = f_p(0) = 0$, that $\theta = 0$ is the only zero of f_p, and that $M_{f_p} \leq 4$. In order to prove that $M_{f_p} \leq \frac{8}{p+1}+\frac{2\pi^2}{3}\left(\frac{4}{\pi^2}\right)^p$, we use the inequalities

$$2-2\cos\theta \leq \theta^2 - \frac{\theta^4}{18} \leq \theta^2, \quad \forall\theta \in [-\pi,\pi].$$

It follows that

$$(2-2\cos\theta)\left(\frac{2-2\cos\theta}{\theta^2}\right)^p \leq \theta^2\left(1-\frac{\theta^2}{18}\right)^p, \quad \forall\theta \in [-\pi,\pi].$$

If $p \geq 2$, the maximum of $\theta^2\left(1-\frac{\theta^2}{18}\right)^p$ over $[-\pi,\pi]$ is located at $\theta^2 = \frac{18}{p+1}$ and its value is given by

$$\frac{18}{p+1}\left(1-\frac{1}{p+1}\right)^p \leq \frac{8}{p+1}.$$

Therefore, if $p \geq 2$, we have

$$\frac{(2-2\cos\theta)^{p+1}}{\theta^{2p}} \leq \frac{8}{p+1}, \qquad \forall\theta \in [-\pi,\pi]. \tag{39}$$

Moreover,

$$\frac{(2-2\cos\theta)^{p+1}}{6\pi^{2p-2}} \leq \frac{4^{p+1}}{6\pi^{2p-2}}, \qquad \forall\theta \in [-\pi,\pi]. \tag{40}$$

Recalling (38), the inequalities (39)–(40) prove that, for $p \geq 2$,

$$M_{f_p} \leq \frac{8}{p+1}+\frac{2\pi^2}{3}\left(\frac{4}{\pi^2}\right)^p. \tag{41}$$

In addition, (41) holds for $p = 1$ too, because $f_1(\theta) = 2-2\cos\theta$ and $M_{f_1} = 4$. To complete the proof of the second statement, we still have to show that

$$M_{f_p} \leq 2\dot{\phi}_{[2p]}(p)+2\sum_{k=1}^{p}|\ddot{\phi}_{[2p+1]}(p+1-k)|, \tag{42}$$

which is easily obtained by using (25) and (33). □

Lemma 8 *Let $p \geq 1$, and let $h_p : [-\pi,\pi] \to \mathbb{R}$,*

$$h_p(\theta) := \phi_{[2p+1]}(p+1)+2\sum_{k=1}^{p}\phi_{[2p+1]}(p+1-k)\cos(k\theta), \tag{43}$$

and $m_{h_p} := \min_{\theta\in[-\pi,\pi]} h_p(\theta)$. Then the following properties hold.

1. $h_p(\theta) = \sum_{k\in\mathbb{Z}} \left|\widehat{\phi_{[p]}}(\theta+2k\pi)\right|^2.$

2. $\max_{\theta\in[-\pi,\pi]} h_p(\theta) = h_p(0) = 1$ *and* $m_{h_p} \geq \left(\frac{4}{\pi^2}\right)^{p+1}.$

Proof From relation (22) and Theorem 6 it follows that

$$\begin{aligned} h_p(\theta) &= \sum_{k\in\mathbb{Z}} \phi_{[2p+1]}(p+1-k)\mathrm{e}^{\mathrm{i}(k\theta)} = \sum_{k\in\mathbb{Z}} \left(\int_{\mathbb{R}} \phi_{[p]}(t)\phi_{[p]}(t-k)\mathrm{d}t\right)\mathrm{e}^{\mathrm{i}(k\theta)} \\ &= \sum_{k\in\mathbb{Z}} \left|\widehat{\phi_{[p]}}(\theta+2k\pi)\right|^2. \end{aligned}$$

The inequalities (31)–(32) imply that

$$\left(\frac{4}{\pi^2}\right)^{p+1} \leq h_p(\theta) \leq 1, \qquad \theta\in[-\pi,\pi]. \tag{44}$$

In addition, by the symmetry property (15) and the partition of unity property (17), we get

$$h(0) = \phi_{[2p+1]}(p+1) + 2\sum_{k=1}^{p} \phi_{[2p+1]}(p+1-k) = \sum_{k=1}^{2p+1} \phi_{[2p+1]}(k) = 1.$$

□

Remark 4 From the expressions of f_p and h_p given in Lemmas 7 and 8, respectively, it follows that for every $\theta\in[-\pi,\pi]$ and $p\geq 2$,

$$f_p(\theta) = (2-2\cos\theta)h_{p-1}(\theta),$$

and for $p\geq 1$,

$$f_p(\theta) = (2-2\cos\theta)\left(\phi_{[2p-1]}(p) + 2\sum_{k=1}^{p-1}\phi_{[2p-1]}(p-k)\cos(k\theta)\right). \tag{45}$$

The latter equality can be easily checked for $p=1$ by a direct computation, with the usual assumption that a sum is empty when the upper index is less than the lower one. Note that (45) is a more elegant and efficient formula to compute f_p.

4 The 1D setting

In this section we focus on the problem (1) in the case where $d=1$ and $\Omega=(0,1)$ is a one-dimensional domain, namely

$$\begin{cases} -u'' + \beta u' + \gamma u = \mathrm{f}, & 0<x<1, \\ u(0)=0, \quad u(1)=0, \end{cases} \tag{46}$$

with $\mathrm{f}\in L_2((0,1))$, $\beta\in\mathbb{R}$, $\gamma\geq 0$. In order to approximate the weak solution u of problem (46) by means of the Galerkin method (4), in the IgA setting we choose the approximation space $\mathscr{W}$ to be a space of smooth spline functions, as we are going to describe now.

Fix $p \geq 1$, $n \geq 2$ and let $\mathscr{V}_n^{[p]}$ be the space of splines of degree p (or order $p+1$) defined over the knot sequence

$$t_1 = \ldots = t_{p+1} = 0 < t_{p+2} < \ldots < t_{p+n} < 1 = t_{p+n+1} = \ldots = t_{2p+n+1}, \tag{47}$$

where

$$t_{p+i+1} := \frac{i}{n}, \quad \forall i = 0, \ldots, n, \tag{48}$$

and the extreme knots have multiplicity $p+1$. More precisely,

$$\mathscr{V}_n^{[p]} := \{s \in C^{p-1}([0,1]) : s|_{[t_{p+i+1}, t_{p+i+2})} \in \mathbb{P}_p, \ \forall i = 0, 1, \ldots, n-1\}.$$

Let $\mathscr{W}_n^{[p]}$ be the subspace of $\mathscr{V}_n^{[p]}$ formed by the spline functions vanishing at the boundary of $[0,1]$, i.e.,

$$\mathscr{W}_n^{[p]} := \{s \in \mathscr{V}_n^{[p]} : s(0) = s(1) = 0\} \subset H_0^1([0,1]). \tag{49}$$

We recall that $\dim \mathscr{V}_n^{[p]} = n+p$ and $\dim \mathscr{W}_n^{[p]} = n+p-2$. In the IgA setting we choose the approximation space $\mathscr{W} = \mathscr{W}_n^{[p]}$ for some $p \geq 1$ and $n \geq 2$.

This space is spanned by the B-spline basis defined as follows (see [14]). Using the convention that a fraction with zero denominator is zero, define the function $N_{i,[k]} : [0,1] \to \mathbb{R}$ for every (k,i) such that $0 \leq k \leq p$, $1 \leq i \leq (n+p)+p-k$:

$$N_{i,[0]}(x) := \begin{cases} 1, & \text{if } x \in [t_i, t_{i+1}), \\ 0, & \text{elsewhere}, \end{cases}$$

and

$$N_{i,[k]}(x) := \frac{x - t_i}{t_{i+k} - t_i} N_{i,[k-1]}(x) + \frac{t_{i+k+1} - x}{t_{i+k+1} - t_{i+1}} N_{i+1,[k-1]}(x), \quad k > 0.$$

Then $\{N_{i,[p]} : i = 1, \ldots, n+p\}$ is a basis of $\mathscr{V}_n^{[p]}$, called the B-spline basis of $\mathscr{V}_n^{[p]}$. Moreover, noting that [14]

$$N_{i,[p]}(0) = N_{i,[p]}(1) = 0, \quad \forall i = 2, \ldots, n+p-1,$$

we deduce that $\{N_{i,[p]} : i = 2, \ldots, n+p-1\}$ is a basis of $\mathscr{W} = \mathscr{W}_n^{[p]}$:

$$\mathscr{W} = \langle N_{i,[p]}, \ i = 2, \ldots, n+p-1 \rangle. \tag{50}$$

If we choose $p = 1$ then we obtain by the above construction the same approximation space $\mathscr{W}$ and the same basis functions considered in classical FEM with linear elements, see [33].

Using the basis (50), the stiffness matrix A in (5) is the object of our interest and, from now onwards, will be denoted by $A_n^{[p]}$ in order to emphasize its dependence on n and p:

$$A_n^{[p]} := A = \left[a(N_{j+1,[p]}, N_{i+1,[p]})\right]_{i,j=1}^{n+p-2}, \tag{51}$$

where in this case $a(u,v) = \int_0^1 u'v' \mathrm{d}x + \beta \int_0^1 u'v \, \mathrm{d}x + \gamma \int_0^1 uv \, \mathrm{d}x$, see (3).

4.1 Construction of the matrices $A_n^{[p]}$

The central basis functions $N_{i,[p]}(x),\ i=p+1,\dots,n,$ defined on the knot sequence (47)–(48) are cardinal B-splines, see Section 3. We have

$$N_{i,[p]}(x)=\phi_{[p]}(nx-i+p+1), \quad i=p+1,\dots,n. \tag{52}$$

Due to the compact support of the B-spline basis, the stiffness matrix $A_n^{[p]}$ has a $(2p+1)$-band structure. We note that

$$N'_{i,[p]}(x)=n\,\dot{\phi}_{[p]}(nx-i+p+1), \quad i=p+1,\dots,n.$$

We now focus on the central part of the stiffness matrix which is only determined by the cardinal B-splines in (52). For each $k=0,1,\dots,p$ and $i=2p,\dots,n-p-1$, the non-zero element in (51) at row i and column $i\pm k$ can be expressed by

$$\begin{aligned}\left(A_n^{[p]}\right)_{i,i\pm k}&=a(N_{i+1\pm k,[p]}(x),N_{i+1,[p]}(x))=a(\phi_{[p]}(nx-i+p\mp k),\phi_{[p]}(nx-i+p))\\&=n^2\int_0^1\dot{\phi}_{[p]}(nx-i+p\mp k)\,\dot{\phi}_{[p]}(nx-i+p)\,\mathrm{d}x\\&\quad+n\beta\int_0^1\dot{\phi}_{[p]}(nx-i+p\mp k)\,\phi_{[p]}(nx-i+p)\,\mathrm{d}x\\&\quad+\gamma\int_0^1\phi_{[p]}(nx-i+p\mp k)\,\phi_{[p]}(nx-i+p)\,\mathrm{d}x\\&=n\int_{\mathbb{R}}\dot{\phi}_{[p]}(t\mp k)\,\dot{\phi}_{[p]}(t)\,\mathrm{d}t+\beta\int_{\mathbb{R}}\dot{\phi}_{[p]}(t\mp k)\,\phi_{[p]}(t)\,\mathrm{d}t+\frac{\gamma}{n}\int_{\mathbb{R}}\phi_{[p]}(t\mp k)\,\phi_{[p]}(t)\,\mathrm{d}t.\end{aligned} \tag{53}$$

Let us consider the following split of the matrix,

$$A_n^{[p]}=nK_n^{[p]}+\beta H_n^{[p]}+\frac{\gamma}{n}M_n^{[p]}, \tag{54}$$

according to the diffusion, advection and reaction terms, respectively. More precisely,

$$nK_n^{[p]}:=\left[\int_0^1 N'_{j+1,[p]}(x)N'_{i+1,[p]}(x)\,\mathrm{d}x\right]_{i,j=1}^{n+p-2}, \tag{55}$$

$$H_n^{[p]}:=\left[\int_0^1 N'_{j+1,[p]}(x)N_{i+1,[p]}(x)\,\mathrm{d}x\right]_{i,j=1}^{n+p-2}, \tag{56}$$

$$\frac{1}{n}M_n^{[p]}:=\left[\int_0^1 N_{j+1,[p]}(x)N_{i+1,[p]}(x)\,\mathrm{d}x\right]_{i,j=1}^{n+p-2}. \tag{57}$$

In view of (53), the parts of these matrices determined by the cardinal B-splines in (52) are

$$\left(K_n^{[p]}\right)_{i,i\pm k}=\int_{\mathbb{R}}\dot{\phi}_{[p]}(t\mp k)\dot{\phi}_{[p]}(t)\,\mathrm{d}t, \tag{58}$$

$$\left(H_n^{[p]}\right)_{i,i\pm k}=\int_{\mathbb{R}}\dot{\phi}_{[p]}(t\mp k)\phi_{[p]}(t)\,\mathrm{d}t, \tag{59}$$

$$\left(M_n^{[p]}\right)_{i,i\pm k}=\int_{\mathbb{R}}\phi_{[p]}(t\mp k)\phi_{[p]}(t)\,\mathrm{d}t, \tag{60}$$

for $k=0,1,\dots,p$ and $i=2p,\dots,n-p-1$. We now derive simple expressions for the elements of the central rows of the matrices $K_n^{[p]}$, $H_n^{[p]}$ and $M_n^{[p]}$ given in (58)–(60), i.e., for the row indices $i=2p,\dots,n-p-1$. Other rules have to be considered for the remaining $2p-1$ initial/final rows. Lemma 5 implies the following result.

Theorem 7 *The matrix $K_n^{[p]}$ is symmetric, the matrix $H_n^{[p]}$ is skew-symmetric and the matrix $M_n^{[p]}$ is symmetric. Moreover, the central non-vanishing elements can be expressed as*

$$\begin{aligned}\left(K_n^{[p]}\right)_{i,i\pm k} &= -\ddot{\phi}_{[2p+1]}(p+1-k),\\ \left(H_n^{[p]}\right)_{i,i+k} &= -\left(H_n^{[p]}\right)_{i,i-k} = \dot{\phi}_{[2p+1]}(p+1-k),\\ \left(M_n^{[p]}\right)_{i,i\pm k} &= \phi_{[2p+1]}(p+1-k),\end{aligned}$$

for $k = 0, 1, \dots, p$ and $i = 2p, \dots, n-p-1$.

From the above theorem, the generic central row of $K_n^{[p]}$ can be expressed as

$$\left[\, 0 \cdots 0 \ -\ddot{\phi}_{[2p+1]}(1) \ \cdots \ -\ddot{\phi}_{[2p+1]}(p) \ -\ddot{\phi}_{[2p+1]}(p+1) \ -\ddot{\phi}_{[2p+1]}(p) \ \cdots \ -\ddot{\phi}_{[2p+1]}(1) \ 0 \cdots 0 \,\right], \tag{61}$$

and in particular, by (25), the diagonal elements can be expressed as

$$\left(K_n^{[p]}\right)_{i,i} = 2\dot{\phi}_{[2p]}(p) > 0.$$

The generic central row of $H_n^{[p]}$ can be expressed as

$$\left[\, 0 \cdots 0 \ -\dot{\phi}_{[2p+1]}(1) \ \cdots \ -\dot{\phi}_{[2p+1]}(p) \ 0 \ \dot{\phi}_{[2p+1]}(p) \ \cdots \ \dot{\phi}_{[2p+1]}(1) \ 0 \cdots 0 \,\right], \tag{62}$$

where we remark that $\left(H_n^{[p]}\right)_{i,i} = \dot{\phi}_{[2p+1]}(p+1) = 0$. The generic central row of $M_n^{[p]}$ can be expressed as

$$\left[\, 0 \cdots 0 \ \phi_{[2p+1]}(1) \ \cdots \ \phi_{[2p+1]}(p) \ \phi_{[2p+1]}(p+1) \ \phi_{[2p+1]}(p) \ \cdots \ \phi_{[2p+1]}(1) \ 0 \cdots 0 \,\right]. \tag{63}$$

As a consequence of Theorem 7, we get the following result.

Corollary 1 *The central non-vanishing elements of the matrix $A_n^{[p]}$ can be expressed as*

$$\left(A_n^{[p]}\right)_{i,i\pm k} = -n\ddot{\phi}_{[2p+1]}(p+1-k) \pm \beta\dot{\phi}_{[2p+1]}(p+1-k) + \frac{\gamma}{n}\phi_{[2p+1]}(p+1-k), \tag{64}$$

for $k = 0, 1, \dots, p$ and $i = 2p, \dots, n-p-1$.

Remark 5 Considering the recurrence relations for derivatives (18)–(19), for the computation of the matrix elements (64) we only need to evaluate cardinal B-splines at integer positions. We sum up some possibilities to evaluate $\phi_{[p]}$ at integer positions.

1. The values of $\phi_{[p]}$ at the integers can be obtained by recurrence relation (12). Recalling that $\phi_{[0]}(k) = \delta_{0k}$, $\phi_{[1]}(k) = \delta_{1k}$, $k \in \mathbb{Z}$, we have

$$\phi_{[p]}(k) = \frac{k}{p}\phi_{[p-1]}(k) + \frac{p+1-k}{p}\phi_{[p-1]}(k-1).$$

2. From (13) it follows that the non-zero values of $\phi_{[p]}$ at the integers are equal to

$$\phi_{[p]}(k) = \frac{1}{p!}\sum_{i=0}^{k-1}\binom{p+1}{i}(-1)^i(k-i)^p, \quad k = 1, \dots, p.$$

4.2 Estimates for the minimal eigenvalues

In this subsection we provide estimates for the minimal eigenvalues of $M_n^{[p]}$ and $K_n^{[p]}$. These estimates will be employed to obtain a lower bound for $|\lambda_{\min}(A_n^{[p]})|$, where $\lambda_{\min}(A_n^{[p]})$ is an eigenvalue of $A_n^{[p]}$ with minimum modulus.

We begin with recalling the following result from [36]. The inequalities in (65) are a special instance for the L_2-norm of the results stated in [36, Theorem 9.27]. We remark that the quantity $\bar{\Delta}$ used in the cited theorem in our context has the value $\frac{1}{n}$, see [36, eq. (6.3)].

Lemma 9 *For every $p \geq 1$, $n \geq 2$, and $\mathbf{x} = (x_1, \ldots, x_{n+p-2}) \in \mathbb{R}^{n+p-2}$,*

$$C_p \frac{\|\mathbf{x}\|^2}{n} \leq \left\| \sum_{i=1}^{n+p-2} x_i N_{i+1,[p]} \right\|^2_{L_2([0,1])} \leq \bar{C}_p \frac{\|\mathbf{x}\|^2}{n}, \tag{65}$$

where the constants $C_p, \bar{C}_p > 0$ do not depend on n and $\mathbf{x}$.

Now we state the Poincaré inequality in the one-dimensional setting. This inequality plays an important role in the proof of Theorem 8.

Lemma 10 (Poincaré's inequality) *For all $v \in H_0^1([0,1])$,*

$$\|v\|_{L_2([0,1])} \leq \frac{1}{\pi} \|v'\|_{L_2([0,1])}. \tag{66}$$

In [16] we find that $\dfrac{1}{\pi} = \sqrt{\dfrac{1}{c_{1,1}}}$ is the best constant such that (66) is satisfied for all $v \in H_0^1([0,1])$. Here, $c_{1,1}$ is the number appearing in (7) for $s = j = 1$, see also Remarks 1–3.

Theorem 8 *Let $C_p > 0$ be the constant in (65), then for all $p \geq 1$ and $n \geq 2$ the following properties hold.*

1. $\lambda_{\min}(M_n^{[p]}) \geq C_p$.
2. $K_n^{[p]} \geq \dfrac{\pi^2}{n^2} M_n^{[p]}$ *and* $\lambda_{\min}(K_n^{[p]}) \geq \dfrac{\pi^2 C_p}{n^2}$.

Proof Fix $p \geq 1, n \geq 2$. By using the definition of $M_n^{[p]}$, see (57), we have for all $\mathbf{y} \in \mathbb{R}^{n+p-2}$,

$$\begin{aligned}
\mathbf{y}^T \left(\frac{1}{n} M_n^{[p]}\right) \mathbf{y} &= \sum_{i,j=1}^{n+p-2} \left(\frac{1}{n} M_n^{[p]}\right)_{i,j} y_i y_j = \sum_{i,j=1}^{n+p-2} \int_0^1 y_i y_j N_{j+1,[p]}(x) N_{i+1,[p]}(x) \mathrm{d}x \\
&= \int_0^1 \sum_{i=1}^{n+p-2} y_i N_{i+1,[p]}(x) \sum_{j=1}^{n+p-2} y_j N_{j+1,[p]}(x) \mathrm{d}x \\
&= \int_0^1 \left(\sum_{i=1}^{n+p-2} y_i N_{i+1,[p]}(x) \right)^2 \mathrm{d}x = \left\| \sum_{i=1}^{n+p-2} y_i N_{i+1,[p]} \right\|^2_{L_2([0,1])} \geq C_p \frac{\|\mathbf{y}\|^2}{n}.
\end{aligned} \tag{67}$$

The last inequality holds because of (65). Hence, we get $\mathbf{y}^T M_n^{[p]} \mathbf{y} \geq C_p \|\mathbf{y}\|^2$, and from the minimax principle [11, 13] it follows that

$$\lambda_{\min}(M_n^{[p]}) = \min_{\mathbf{y} \neq \mathbf{0}} \frac{\mathbf{y}^T M_n^{[p]} \mathbf{y}}{\|\mathbf{y}\|^2} \geq C_p. \tag{68}$$

This proves the first statement. To prove the second statement, we use the definition of $K_n^{[p]}$, see (55), and obtain for all $\mathbf{y} \in \mathbb{R}^{n+p-2}$,

$$\begin{aligned}
\mathbf{y}^T \left(nK_n^{[p]}\right) \mathbf{y} &= \sum_{i,j=1}^{n+p-2} \left(nK_n^{[p]}\right)_{i,j} y_i y_j = \sum_{i,j=1}^{n+p-2} \int_0^1 y_i y_j N'_{j+1,[p]}(x) N'_{i+1,[p]}(x) \mathrm{d}x \\
&= \int_0^1 \sum_{i=1}^{n+p-2} y_i N'_{i+1,[p]}(x) \sum_{j=1}^{n+p-2} y_j N'_{j+1,[p]}(x) \mathrm{d}x \\
&= \int_0^1 \left(\sum_{i=1}^{n+p-2} y_i N'_{i+1,[p]}(x) \right)^2 \mathrm{d}x = \left\| \sum_{i=1}^{n+p-2} y_i N'_{i+1,[p]} \right\|^2_{L_2([0,1])} = \|v'_{\mathbf{y}}\|^2_{L_2([0,1])}, \qquad (69)
\end{aligned}$$

where $v_{\mathbf{y}} := \sum_{i=1}^{n+p-2} y_i N_{i+1,[p]} \in \mathscr{W}_n^{[p]}$, see (49). Since $\mathscr{W}_n^{[p]} \subset H_0^1([0,1])$, we may apply the Poincaré inequality (66). From (66) and (67) it follows that

$$\mathbf{y}^T \left(nK_n^{[p]}\right) \mathbf{y} = \|v'_{\mathbf{y}}\|^2_{L_2([0,1])} \geq \pi^2 \|v_{\mathbf{y}}\|^2_{L_2([0,1])} = \mathbf{y}^T \left(\frac{\pi^2}{n} M_n^{[p]} \right) \mathbf{y}.$$

Dividing both sides by n we obtain, for all $\mathbf{y} \in \mathbb{R}^{n+p-2}$,

$$\mathbf{y}^T K_n^{[p]} \mathbf{y} \geq \mathbf{y}^T \left(\frac{\pi^2}{n^2} M_n^{[p]} \right) \mathbf{y}.$$

This proves that $K_n^{[p]} \geq \frac{\pi^2}{n^2} M_n^{[p]}$, and applying the minimax principle and (68) yield

$$\lambda_{\min}(K_n^{[p]}) = \min_{\mathbf{y}\neq\mathbf{0}} \frac{\mathbf{y}^T K_n^{[p]} \mathbf{y}}{\|\mathbf{y}\|^2} \geq \min_{\mathbf{y}\neq\mathbf{0}} \frac{\mathbf{y}^T \left(\frac{\pi^2}{n^2} M_n^{[p]}\right) \mathbf{y}}{\|\mathbf{y}\|^2} = \frac{\pi^2}{n^2} \lambda_{\min}(M_n^{[p]}) \geq \frac{\pi^2 C_p}{n^2},$$

which concludes the proof. □

Remark 6 Suppose that, for a given $p \geq 1$, we are able to find a constant $\widetilde{C}_p > 0$ such that[2]

$$\lambda_{\min}(M_n^{[p]}) \geq \widetilde{C}_p.$$

In this case, the statements in Theorem 8 hold with $\widetilde{C}_p$ instead of C_p. Moreover, the left inequality in (65) also holds with $\widetilde{C}_p$ instead of C_p. Indeed, by using a similar argument as in the proof of Theorem 8, we obtain

$$\frac{n \left\| \sum_{i=1}^{n+p-2} x_i N_{i+1,[p]} \right\|^2_{L_2([0,1])}}{\|\mathbf{x}\|^2} = \frac{\mathbf{x}^T M_n^{[p]} \mathbf{x}}{\|\mathbf{x}\|^2} \geq \min_{\mathbf{y}\neq 0} \frac{\mathbf{y}^T M_n^{[p]} \mathbf{y}}{\|\mathbf{y}\|^2} = \lambda_{\min}(M_n^{[p]}) \geq \widetilde{C}_p.$$

[2] Such a constant $\widetilde{C}_p$ could be found by means of the Gershgorin theorems [13]. We refer to Remark 9 in Section 4.6.1 for an example.

n	$\lambda_{\min}(M_n^{[2]})$	$\lambda_{\min}(M_n^{[3]})$	$\lambda_{\min}(M_n^{[4]})$	$n^2\,\lambda_{\min}(K_n^{[2]})$	$n^2\,\lambda_{\min}(K_n^{[3]})$	$n^2\,\lambda_{\min}(K_n^{[4]})$
20	0.1333333	0.0482607	0.0171864	9.8089070	9.7834046	9.7507398
40	0.1333333	0.0486447	0.0173795	9.8543957	9.8486563	9.8419964
80	0.1333333	0.0486538	0.0173821	9.8658001	9.8644478	9.8629796
160	0.1333333	0.0486538	0.0173821	9.8686532	9.8683256	9.8679834
320	0.1333333	0.0486538	0.0173821	9.8693666	9.8692860	9.8692036
640	0.1333333	0.0486538	0.0173821	9.8695450	9.8695250	9.8695048
1280	0.1333333	0.0486538	0.0173821	9.8695896	9.8695846	9.8695796
2560	0.1333333	0.0486538	0.0173821	9.8696007	9.8695994	9.8695982
5120	0.1333333	0.0486538	0.0173821	9.8696035	9.8696032	9.8696029

Table 1 Computation of $\lambda_{\min}(M_n^{[p]})$ and $n^2\,\lambda_{\min}(K_n^{[p]})$ for $p = 2,3,4$ and for increasing values of n.

Table 1 shows the results of some numerical experiments performed on the matrices $M_n^{[p]}$ and $K_n^{[p]}$ for $p = 2,3,4$ and for increasing values of n. From these results it seems that

$$\lambda_{\min}(M_n^{[p]}) \overset{n\to\infty}{\sim} \widetilde{m}_p, \tag{70}$$

with $\widetilde{m}_2 = \frac{2}{15}$, $\widetilde{m}_3 \approx 0.0486538$ and $\widetilde{m}_4 \approx 0.0173821$. Apparently, the sequence $\lambda_{\min}(M_n^{[p]})$ converges to $\widetilde{m}_p$ very quickly as $n \to \infty$. In addition, it seems that[3]

$$\lambda_{\min}(K_n^{[p]}) \overset{n\to\infty}{\sim} \frac{\pi^2}{n^2}. \tag{71}$$

Note that $K_n^{[1]} = \text{Tridiagonal}(-1,2,-1) \in \mathbb{R}^{(n-1)\times(n-1)}$ (see Section 4.5) and for $\lambda_{\min}(K_n^{[1]})$ the asymptotic formula (71) holds, because it is known that

$$\lambda_{\min}(K_n^{[1]}) = 4\left(\sin\frac{\pi}{2n}\right)^2 \overset{n\to\infty}{\sim} \frac{\pi^2}{n^2}.$$

The numerical experiments confirm that, for $p = 2,3,4$, the eigenvalue $\lambda_{\min}(K_n^{[p]})$ converges to 0 as n^{-2}, which means that the lower estimate $\dfrac{\pi^2 C_p}{n^2}$ obtained in Theorem 8 is asymptotically of the same order as $\lambda_{\min}(K_n^{[p]})$ when $n \to \infty$.

In addition, referring to Table 2, we can formulate a deeper conjecture than (71).

Conjecture 1 For every $p \geq 1$, $n \geq 2$ and $j = 1,\dots,n+p-2$, let us denote by $\lambda_j(K_n^{[p]})$ the j-th smallest eigenvalue of $K_n^{[p]}$: $\lambda_1(K_n^{[p]}) \leq \ldots \leq \lambda_{n+p-2}(K_n^{[p]})$. Then, for every $p \geq 1$ and for each fixed $j \geq 1$,

$$\lim_{n\to\infty}\left(n^2\,\lambda_j(K_n^{[p]})\right) = j^2\pi^2. \tag{72}$$

This conjecture can be motivated as follows. The matrix $K_n^{[p]}$ is associated with the (IgA) discretization of the boundary value problem (9), because $nK_n^{[p]}$ coincides with $A_n^{[p]}$ if $\beta = \gamma = 0$. The numbers $j^2\pi^2$, $j = 1,2,\dots$, are precisely the eigenvalues of (9), see Remark 2. The matrices $T_m(2-2\cos\theta)$, $m = 1,2,\dots$, are also associated with the (Finite Difference) discretization of (9) and for these matrices Theorem 4 establishes exactly the limit relation $\lim\limits_{m\to\infty}\left(m^2\,\lambda_j(T_m(2-2\cos\theta))\right) = j^2\pi^2$ for each fixed $j \geq 1$, see Remark 3.

[3] The constant π^2 is precisely $c_{1,1}$, see Remarks 1–3.

n	$\check{\lambda}_{2,n}^{[2]}$	$\check{\lambda}_{2,n}^{[3]}$	$\check{\lambda}_{2,n}^{[4]}$	$\check{\lambda}_{3,n}^{[2]}$	$\check{\lambda}_{3,n}^{[3]}$	$\check{\lambda}_{3,n}^{[4]}$
20	9.6289991	9.5293645	9.4042990	9.3363215	9.1205723	8.8541077
40	9.8089070	9.7860742	9.7596733	9.7335487	9.6826385	9.6241225
80	9.8543957	9.8489935	9.8431323	9.8354171	9.8232883	9.8101443
160	9.8658001	9.8644900	9.8631221	9.8610467	9.8581006	9.8550251
320	9.8686532	9.8683308	9.8680013	9.8674643	9.8667391	9.8659977
640	9.8693666	9.8692867	9.8692058	9.8690693	9.8688895	9.8687076
1280	9.8695449	9.8695250	9.8695050	9.8694706	9.8694259	9.8693808
2560	9.8695895	9.8695846	9.8695796	9.8695710	9.8695598	9.8695486
5120	9.8696007	9.8695994	9.8695982	9.8695960	9.8695933	9.8695905

Table 2 Computation of $\check{\lambda}_{j,n}^{[p]} := (n/j)^2\,\lambda_j(K_n^{[p]})$ for $p = 2,3,4$, for $j = 2,3$ and for increasing values of n.

Theorem 9 *For all $p \geq 1$ and all $n \geq 2$, let $\lambda_{\min}(A_n^{[p]})$ be an eigenvalue of $A_n^{[p]}$ with minimum modulus. Then*

$$\left|\lambda_{\min}(A_n^{[p]})\right| \geq \lambda_{\min}(\mathrm{Re}A_n^{[p]}) \geq \frac{C_p(\pi^2+\gamma)}{n}, \tag{73}$$

with $C_p > 0$ being the same constant appearing in Theorem 8.

Proof By the expression (54) of $A_n^{[p]}$ and recalling that $K_n^{[p]}$, $M_n^{[p]}$ are symmetric, while $H_n^{[p]}$ is skew-symmetric, we infer that the real part of $A_n^{[p]}$ is given by

$$\mathrm{Re}A_n^{[p]} = nK_n^{[p]} + \frac{\gamma}{n}M_n^{[p]}.$$

Therefore, by the minimax principle and by Theorem 8 we obtain

$$\begin{aligned}\lambda_{\min}(\mathrm{Re}A_n^{[p]}) &= \lambda_{\min}\left(nK_n^{[p]} + \frac{\gamma}{n}M_n^{[p]}\right) \geq \lambda_{\min}(nK_n^{[p]}) + \lambda_{\min}\left(\frac{\gamma}{n}M_n^{[p]}\right)\\ &= n\lambda_{\min}(K_n^{[p]}) + \frac{\gamma}{n}\lambda_{\min}(M_n^{[p]}) \geq n\,\frac{\pi^2 C_p}{n^2} + \frac{\gamma}{n}C_p = \frac{C_p(\pi^2+\gamma)}{n}.\end{aligned}$$

From (6) we know that $|\lambda_{\min}(A_n^{[p]})| \geq \lambda_{\min}(\mathrm{Re}A_n^{[p]})$, implying (73). □

The lower bound (73) remains bounded away from 0 for all $\gamma \geq 0$ and, in particular, for the interesting value $\gamma = 0$.

4.3 Conditioning

In this subsection we provide a bound for the condition number

$$\kappa_2(A_n^{[p]}) := \|A_n^{[p]}\|\,\|(A_n^{[p]})^{-1}\|,$$

see Theorem 11. For its proof we need two auxiliary results. The first one (Theorem 10) is the Fan-Hoffman theorem [11, Proposition III.5.1]. The second result (Lemma 11) gives a bound for the infinity norm of the matrices $K_n^{[p]}$, $H_n^{[p]}$ and $M_n^{[p]}$.

Theorem 10 (Fan-Hoffman) *Let $X \in \mathbb{C}^{m\times m}$ and let*

$$\|X\| = s_1(X) \geq s_2(X) \geq \ldots \geq s_m(X), \qquad \lambda_1(\mathrm{Re}X) \geq \lambda_2(\mathrm{Re}X) \geq \ldots \geq \lambda_m(\mathrm{Re}X)$$

be the singular values of X and the eigenvalues of ReX*, respectively. Then*

$$s_j(X) \geq \lambda_j(\mathrm{Re}X), \quad \forall j = 1,\ldots,m.$$

Lemma 11 *For every $p \geq 1$ and every $n \geq 2$,*

$$\left\|\frac{1}{n}M_n^{[p]}\right\|_\infty \leq \frac{1}{n}, \quad \|H_n^{[p]}\|_\infty \leq 2, \quad \|nK_n^{[p]}\|_\infty \leq 4pn.$$

Proof We first note that the derivative and integral of a B-spline $N_{i,[p]}(x)$ are given by (see [14,36]),

$$N'_{i,[p]}(x) = p\left(\frac{N_{i,[p-1]}(x)}{t_{i+p}-t_i} - \frac{N_{i+1,[p-1]}(x)}{t_{i+p+1}-t_{i+1}}\right), \tag{74}$$

and

$$\int_{\mathbb{R}} N_{i,[p]}(x)\,\mathrm{d}x = \frac{t_{i+p+1}-t_i}{p+1}. \tag{75}$$

The sequence of knots (47)–(48) implies that the length of the support of any $N_{i,[p]}$ can be bounded from above by $\frac{p+1}{n}$. Recalling (57), by the positivity property and the partition of unity property of B-splines, we obtain

$$\begin{aligned}\left\|\frac{1}{n}M_n^{[p]}\right\|_\infty &= \max_{i=1\ldots n+p-2} \sum_{j=1}^{n+p-2} \int_0^1 N_{j+1,[p]}(x)N_{i+1,[p]}(x)\,\mathrm{d}x \\ &= \max_{i=1\ldots n+p-2} \int_0^1 \left(\sum_{j=1}^{n+p-2} N_{j+1,[p]}(x)\right) N_{i+1,[p]}(x)\,\mathrm{d}x \\ &\leq \max_{i=1\ldots n+p-2} \int_0^1 N_{i+1,[p]}(x)\,\mathrm{d}x = \max_{i=1\ldots n+p-2} \frac{t_{i+p+2}-t_{i+1}}{p+1} \leq \frac{1}{n}.\end{aligned}$$

Due to the skew-symmetry of $H_n^{[p]}$, see (56), the infinity norm of $H_n^{[p]}$ is equal to the infinity norm of its transpose. By (74) and the positivity property of B-splines, we obtain

$$\begin{aligned}\|H_n^{[p]}\|_\infty &= \max_{i=1\ldots n+p-2} \sum_{j=1}^{n+p-2} \left|\int_0^1 N_{j+1,[p]}(x)N'_{i+1,[p]}(x)\,\mathrm{d}x\right| \\ &= \max_{i=1\ldots n+p-2} p \sum_{j=1}^{n+p-2} \left|\int_0^1 N_{j+1,[p]}(x)\left(\frac{N_{i+1,[p-1]}(x)}{t_{i+p+1}-t_{i+1}} - \frac{N_{i+2,[p-1]}(x)}{t_{i+p+2}-t_{i+2}}\right)\mathrm{d}x\right| \\ &\leq \max_{i=1\ldots n+p-2} p \sum_{j=1}^{n+p-2} \int_0^1 N_{j+1,[p]}(x)\left(\frac{N_{i+1,[p-1]}(x)}{t_{i+p+1}-t_{i+1}} + \frac{N_{i+2,[p-1]}(x)}{t_{i+p+2}-t_{i+2}}\right)\mathrm{d}x.\end{aligned} \tag{76}$$

Using the partition of unity property and (75), we have

$$\sum_{j=1}^{n+p-2} \int_0^1 N_{j+1,[p]}(x)\frac{N_{i+1,[p-1]}(x)}{t_{i+p+1}-t_{i+1}}\,\mathrm{d}x = \int_0^1 \left(\sum_{j=1}^{n+p-2} N_{j+1,[p]}(x)\right)\frac{N_{i+1,[p-1]}(x)}{t_{i+p+1}-t_{i+1}}\,\mathrm{d}x \leq \frac{1}{p},$$

and a similar bound holds for the remaining term in (76). It follows that $\|H_n^{[p]}\|_\infty \leq 2$.

Recalling (55), we obtain

$$\begin{aligned}\|nK_n^{[p]}\|_\infty &= \max_{i=1\ldots n+p-2} \sum_{j=1}^{n+p-2} \left|\int_0^1 N'_{j+1,[p]}(x)N'_{i+1,[p]}(x)\,\mathrm{d}x\right| \\ &= \max_{i=1\ldots n+p-2} p^2 \sum_{j=1}^{n+p-2} \left|\int_0^1 \left(\frac{N_{j+1,[p-1]}(x)}{t_{j+p+1}-t_{j+1}} - \frac{N_{j+2,[p-1]}(x)}{t_{j+p+2}-t_{j+2}}\right)\right. \\ &\qquad \left.\left(\frac{N_{i+1,[p-1]}(x)}{t_{i+p+1}-t_{i+1}} - \frac{N_{i+2,[p-1]}(x)}{t_{i+p+2}-t_{i+2}}\right)\mathrm{d}x\right|.\end{aligned} \tag{77}$$

In addition, we have

$$\sum_{j=1}^{n+p-2}\int_0^1 \frac{N_{j+1,[p-1]}(x)}{t_{j+p+1}-t_{j+1}}\frac{N_{i+1,[p-1]}(x)}{t_{i+p+1}-t_{i+1}}\,\mathrm{d}x=\int_0^1\left(\sum_{j=1}^{n+p-2}\frac{N_{j+1,[p-1]}(x)}{t_{j+p+1}-t_{j+1}}\right)\frac{N_{i+1,[p-1]}(x)}{t_{i+p+1}-t_{i+1}}\,\mathrm{d}x$$
$$\le n\int_0^1\frac{N_{i+1,[p-1]}(x)}{t_{i+p+1}-t_{i+1}}\,\mathrm{d}x=\frac{n}{p},$$

and in a similar way we can also bound the remaining terms in (77). This results in

$$\|nK_n^{[p]}\|_\infty\le\max_{i=1\ldots n+p-2}p^2 4\frac{n}{p}=4pn.$$

□

Remark 7 A consequence of Lemma 11 is that we can take $\bar{C}_p=1$ in (65), independently of p. Indeed, Lemma 11 implies that $\lambda_{\max}(M_n^{[p]})\le\|M_n^{[p]}\|_\infty\le 1$ for all $p\ge 1$ and $n\ge 2$. Thus, by the minimax principle, along the lines of the proof of Theorem 8, we have

$$\frac{n\left\|\sum_{i=1}^{n+p-2}x_iN_{i+1,[p]}\right\|^2_{L_2([0,1])}}{\|\mathbf{x}\|^2}=\frac{\mathbf{x}^TM_n^{[p]}\mathbf{x}}{\|\mathbf{x}\|^2}\le\max_{\mathbf{y}\neq\mathbf{0}}\frac{\mathbf{y}^TM_n^{[p]}\mathbf{y}}{\|\mathbf{y}\|^2}=\lambda_{\max}(M_n^{[p]})\le 1.$$

Theorem 11 *For every $p\ge 1$ there exists a constant $\alpha_p>0$ such that*

$$\kappa_2(A_n^{[p]})\le\alpha_pn^2,\quad\forall n\ge 2.\tag{78}$$

Proof Fix $p\ge 1$ and $n\ge 2$. By Theorem 7 it follows that $K_n^{[p]}$, $H_n^{[p]}$ and $M_n^{[p]}$ are normal matrices, and by applying Lemma 11 we obtain for $\|A_n^{[p]}\|$ the following bound:

$$\begin{aligned}\|A_n^{[p]}\|&=\left\|nK_n^{[p]}+\beta H_n^{[p]}+\frac{\gamma}{n}M_n^{[p]}\right\|\le\|nK_n^{[p]}\|+|\beta|\|H_n^{[p]}\|+\gamma\left\|\frac{1}{n}M_n^{[p]}\right\|\\&=\rho(nK_n^{[p]})+|\beta|\rho(H_n^{[p]})+\gamma\rho\left(\frac{1}{n}M_n^{[p]}\right)\\&\le\|nK_n^{[p]}\|_\infty+|\beta|\|H_n^{[p]}\|_\infty+\gamma\left\|\frac{1}{n}M_n^{[p]}\right\|_\infty\le 4pn+2|\beta|+\frac{\gamma}{n}.\end{aligned}\tag{79}$$

We now give a bound for $\|(A_n^{[p]})^{-1}\|$. Using Theorems 9 and 10, we obtain

$$s_{n+p-2}(A_n^{[p]})\ge\lambda_{\min}(\mathrm{Re}A_n^{[p]})\ge\frac{C_p(\pi^2+\gamma)}{n},$$

where $s_{n+p-2}(A_n^{[p]})$ is the minimum singular value of $A_n^{[p]}$. Hence,

$$\|(A_n^{[p]})^{-1}\|=\frac{1}{s_{n+p-2}(A_n^{[p]})}\le\frac{n}{C_p(\pi^2+\gamma)}.\tag{80}$$

Combining (79) with (80), we get $\kappa_2(A_n^{[p]})\le\dfrac{4pn^2+2n|\beta|+\gamma}{C_p(\pi^2+\gamma)}$, which implies (78) with

$$\alpha_p:=\frac{1}{C_p(\pi^2+\gamma)}\left[4p+|\beta|+\frac{\gamma}{4}\right].$$

□

4.4 Spectral distribution

We will now study, for a fixed $p \geq 1$, the spectral distribution of the sequence

$$\frac{1}{n}A_n^{[p]} = K_n^{[p]} + \frac{\beta}{n}H_n^{[p]} + \frac{\gamma}{n^2}M_n^{[p]}, \tag{81}$$

formed by the scaled stiffness matrices. Recall from (51) that $A_n^{[p]}$ is of size $(n+p-2) \times (n+p-2)$. The central rows of $A_n^{[p]}$ (given in Corollary 1) are those with index ranging from $i = 2p$ to $i = n-p-1$. Thus, the condition on n to ensure that $A_n^{[p]}$ has at least one central row is $n-p-1 \geq 2p$, i.e., $n \geq 3p+1$.

For every $n \geq 3p+1$, we decompose the matrix $K_n^{[p]}$ into

$$K_n^{[p]} = B_n^{[p]} + R_n^{[p]}, \tag{82}$$

where $B_n^{[p]}$ is the symmetric $(2p+1)$-band matrix whose generic central row is given by (61), while $R_n^{[p]} := K_n^{[p]} - B_n^{[p]}$ is a low-rank correction term. Indeed, $R_n^{[p]}$ has at most $2(2p-1)$ non-zero rows and so

$$\mathrm{rank}(R_n^{[p]}) \leq 2(2p-1). \tag{83}$$

Similarly, we decompose the matrix $M_n^{[p]}$ into

$$M_n^{[p]} = C_n^{[p]} + S_n^{[p]}, \tag{84}$$

where $C_n^{[p]}$ is the symmetric $(2p+1)$-band matrix whose generic central row is given by (63), while $S_n^{[p]} := M_n^{[p]} - C_n^{[p]}$ is a low-rank correction term analogous to $R_n^{[p]}$ and

$$\mathrm{rank}(S_n^{[p]}) \leq 2(2p-1). \tag{85}$$

Now we analyze the spectral properties of $B_n^{[p]}$ and $C_n^{[p]}$. Besides being interesting in its own right, some of the given properties are needed for the proof of Theorem 12, which yields the spectral distribution of the sequence $\{\frac{1}{n}A_n^{[p]}\}$.

Lemma 12 *Let f_p and M_{f_p} be defined as in Lemma 7. For all $n \geq 3p+1$, $B_n^{[p]} = T_{n+p-2}(f_p)$. Moreover,*

1. *$\sigma(B_n^{[p]}) \subset (0, M_{f_p}), \quad \forall n \geq 3p+1$;*
2. *$\lambda_{\min}(B_n^{[p]}) \searrow 0$ and $\lambda_{\max}(B_n^{[p]}) \nearrow M_{f_p}$ as $n \to \infty$;*
3. *$\{B_n^{[p]}\} \overset{\lambda}{\sim} f_p$;*
4. *for each fixed $j \geq 1$,*

$$\lambda_j(B_n^{[p]}) \overset{n\to\infty}{\sim} \frac{j^2\pi^2}{n^2},$$

where $\lambda_1(B_n^{[p]}) \leq \ldots \leq \lambda_{n+p-2}(B_n^{[p]})$ are the eigenvalues of $B_n^{[p]}$ arranged in increasing order.

Proof From the definitions of $B_n^{[p]}$ and f_p it follows that $B_n^{[p]} = T_{n+p-2}(f_p)$ for all $n \geq 3p+1$. Hence, the first three statements are a consequence of Theorem 3 and Lemma 7.

We now prove the last statement. From Lemma 7 we know that $\theta = 0$ is the unique zero of f_p over $[-\pi, \pi]$. Furthermore, from (36) it is easy to derive that $f_p'(0) = 0$ and, by using Lemma 6,

$$f_p''(0) = 2\sum_{k=1}^{p} k^2 \ddot{\phi}_{[2p+1]}(p+1-k) = 2.$$

This means that the function f_p satisfies all the hypotheses of Theorem 4 with $s = 1$, $\theta_{\min} = 0$ and $f_p^{(2s)}(\theta_{\min}) = 2$. Then, for each fixed $j \geq 1$,

$$\lambda_j(B_n^{[p]}) = \lambda_j(T_{n+p-2}(f_p)) \overset{n\to\infty}{\sim} \frac{c_{1,j}}{(n+p-2)^2} \overset{n\to\infty}{\sim} \frac{j^2\pi^2}{n^2},$$

where the last asymptotic equivalence holds because $c_{1,j} = j^2\pi^2$, see Remarks 2–3. □

Remark 8 In Section 4.2, looking at the numerical results summarized in the Tables 1–2, we conjectured that (72) holds for all $p \geq 1$ and $j \geq 1$. In Lemma 12 we have seen that (72) holds with $\lambda_j(B_n^{[p]})$ instead of $\lambda_j(K_n^{[p]})$. Furthermore, using the Cauchy interlacing theorem [11] and the fact that $B_n^{[p]}$ is a principal submatrix of $K_{n+4p-2}^{[p]}$, it can be shown that

$$\lambda_j(K_{n+4p-2}^{[p]}) \leq \lambda_j(B_n^{[p]}) \leq \lambda_{j+4p-2}(K_{n+4p-2}^{[p]}), \quad \forall p \geq 1, \forall n \geq 2, \forall j = 1, \ldots, n+p-2.$$

Hence, if there exists a constant $\widetilde{k}_{p,j}$ such that

$$\lambda_j(K_n^{[p]}) \overset{n\to\infty}{\sim} \frac{\widetilde{k}_{p,j}}{n^2},$$

then $\widetilde{k}_{p,j} \leq j^2\pi^2$, and if $j > 4p-2$ then $(j-4p+2)^2\pi^2 \leq \widetilde{k}_{p,j} \leq j^2\pi^2$.

Lemma 13 *Let h_p and m_{h_p} be defined as in Lemma 8. For all $n \geq 3p+1$, $C_n^{[p]} = T_{n+p-2}(h_p)$. Moreover,*

1. $\sigma(C_n^{[p]}) \subset (m_{h_p}, 1)$, $\quad \forall n \geq 3p+1$;
2. $\lambda_{\min}(C_n^{[p]}) \searrow m_{h_p}$ *and* $\lambda_{\max}(C_n^{[p]}) \nearrow 1$ *as* $n \to \infty$;
3. $\{C_n^{[p]}\} \overset{\lambda}{\sim} h_p$.

Proof From the definitions of $C_n^{[p]}$ and h_p it follows that $C_n^{[p]} = T_{n+p-2}(h_p)$ for all $n \geq 3p+1$. Theorem 3 and Lemma 8 conclude the proof. □

The function f_p in (33) (as well as the function h_p in (43)) can be easily computed for particular p using the evaluation methods for cardinal B-splines described in Remark 5.

Theorem 12 *The sequence of matrices $\{\frac{1}{n}A_n^{[p]}\}$ is distributed like the function f_p defined in (33) in the sense of the eigenvalues, i.e.,*

$$\lim_{n\to\infty} \frac{1}{n+p-2} \sum_{j=1}^{n+p-2} F\left(\lambda_j\left(\frac{1}{n}A_n^{[p]}\right)\right) = \frac{1}{2\pi}\int_{-\pi}^{\pi} F(f_p(\theta))\,\mathrm{d}\theta, \quad \forall F \in C_c(\mathbb{C}, \mathbb{C}).$$

Furthermore, $\{\frac{1}{n}A_n^{[p]}\}$ is strongly clustered at the range $[0, M_{f_p}]$ of f_p.

Proof Recalling (81)–(82), we have

$$\frac{1}{n}A_n^{[p]} = B_n^{[p]} + R_n^{[p]} + \frac{\beta}{n}H_n^{[p]} + \frac{\gamma}{n^2}M_n^{[p]}. \tag{86}$$

We now prove that the hypotheses of Theorem 2 are satisfied with $Z_n = \frac{1}{n}A_n^{[p]}$, $X_n = B_n^{[p]}$ and Y_n the remaining term in the right-hand side of (86). We have seen in Lemma 12 that $\{B_n^{[p]}\} \overset{\lambda}{\sim} f_p$. Noting that $B_n^{[p]}$ is symmetric, by Lemma 12 we obtain that for all $n \geq 3p+1$,

$$\|B_n^{[p]}\| = \rho(B_n^{[p]}) \leq M_{f_p},$$

where M_{f_p} is a constant independent of n. Since $\operatorname{rank}(R_n^{[p]}) \leq 2(2p-1)$ (see (83)) and since $K_n^{[p]}$, $H_n^{[p]}$ and $M_n^{[p]}$ are normal matrices, we get

$$\begin{aligned}
\left\| R_n^{[p]} + \frac{\beta}{n}H_n^{[p]} + \frac{\gamma}{n^2}M_n^{[p]} \right\|_1 &\leq \|R_n^{[p]}\|_1 + \frac{|\beta|}{n}\|H_n^{[p]}\|_1 + \frac{\gamma}{n^2}\|M_n^{[p]}\|_1 \\
&\leq \operatorname{rank}(R_n^{[p]})\|R_n^{[p]}\| + |\beta|\frac{(n+p-2)}{n}\|H_n^{[p]}\| + \gamma\frac{(n+p-2)}{n^2}\|M_n^{[p]}\| \\
&\leq 2(2p-1)\|K_n^{[p]} - B_n^{[p]}\| + |\beta|\frac{(n+p-2)}{n}\|H_n^{[p]}\| + \gamma\frac{(n+p-2)}{n^2}\|M_n^{[p]}\| \\
&\leq 2(2p-1)\|B_n^{[p]}\| + 2(2p-1)\|K_n^{[p]}\| + |\beta|\frac{(n+p-2)}{n}\|H_n^{[p]}\| + \gamma\frac{(n+p-2)}{n^2}\|M_n^{[p]}\| \\
&\leq 2(2p-1)M_{f_p} + 2(2p-1)\|K_n^{[p]}\|_\infty + |\beta|\frac{(n+p-2)}{n}\|H_n^{[p]}\|_\infty + \gamma\frac{(n+p-2)}{n^2}\|M_n^{[p]}\|_\infty.
\end{aligned}$$

From Lemma 11 it follows that the right-hand side of the last inequality can be bounded from above by a constant independent of n, $\forall n \geq 3p+1$, implying that all the hypotheses of Theorem 2 are satisfied. □

In the next two subsections we discuss in more detail the spectral properties of the matrices $A_n^{[p]}$ for the cases $p=1$ and $p=2$.

4.5 The linear case $p = 1$

In the case $p=1$, for every $n \geq 4$, the matrix $A_n^{[1]}$ is of size $(n-1)\times(n-1)$ and is given by

$$A_n^{[1]} = nK_n^{[1]} + \beta H_n^{[1]} + \frac{\gamma}{n}M_n^{[1]}, \tag{87}$$

where

$$K_n^{[1]} = \begin{bmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 2 \end{bmatrix}, \quad H_n^{[1]} = \frac{1}{2}\begin{bmatrix} 0 & 1 & & & \\ -1 & 0 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 0 & 1 \\ & & & -1 & 0 \end{bmatrix}, \quad M_n^{[1]} = \frac{1}{6}\begin{bmatrix} 4 & 1 & & & \\ 1 & 4 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & 4 & 1 \\ & & & 1 & 4 \end{bmatrix}.$$

The matrix $A_n^{[1]}$ is nothing else than the stiffness matrix arising from classical FEM with linear elements.

Note that the scaled matrix

$$\frac{1}{n}A_n^{[1]} = K_n^{[1]} + \frac{\beta}{n}H_n^{[1]} + \frac{\gamma}{n^2}M_n^{[1]} \tag{88}$$

is a real Toeplitz tridiagonal matrix, namely

$$\frac{1}{n}A_n^{[1]} = \mathrm{Tridiagonal}\left(-1-\frac{\beta}{2n}+\frac{\gamma}{6n^2},\, 2+\frac{2\gamma}{3n^2},\, -1+\frac{\beta}{2n}+\frac{\gamma}{6n^2}\right).$$

Moreover, for n large enough, the elements $-1-\frac{\beta}{2n}+\frac{\gamma}{6n^2}$ and $-1+\frac{\beta}{2n}+\frac{\gamma}{6n^2}$ are both negative. This means that we can compute all the eigenvalues of $\frac{1}{n}A_n^{[1]}$ (for n large enough) by means of the following result.

Theorem 13 *Let*

$$X := \begin{bmatrix} b & c & & & \\ a & b & c & & \\ & \ddots & \ddots & \ddots & \\ & & a & b & c \\ & & & a & b \end{bmatrix} = \mathrm{Tridiagonal}(a,b,c) \in \mathbb{R}^{m\times m}$$

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

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

Proof By direct computation we have $X = \mathrm{diag}_{0\le j\le m-1}(r^j)\cdot Y \cdot \mathrm{diag}_{0\le j\le m-1}(r^{-j})$ with $r = \sqrt{\frac{a}{c}}$ and $Y = \mathrm{Tridiagonal}(\sqrt{ac}, b, \sqrt{ac})$. Thus, X is similar to Y, whose eigenvalues

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

are known since Y belongs to the τ-algebra, see [12]. □

Applying Theorem 13 to our case we obtain the following corollary.

Corollary 2 *Let $n \ge 4$ be such that $-1-\frac{\beta}{2n}+\frac{\gamma}{6n^2}$ and $-1+\frac{\beta}{2n}+\frac{\gamma}{6n^2}$ are both negative. Then, $\frac{1}{n}A_n^{[1]}$ has $n-1$ real distinct eigenvalues*

$$\lambda_j\left(\frac{1}{n}A_n^{[1]}\right) = 2 + \frac{2\gamma}{3n^2} + 2\sqrt{1-\left(\frac{\gamma}{3}+\frac{\beta^2}{4}\right)\frac{1}{n^2}+\frac{\gamma^2}{36n^4}}\,\cos\frac{j\pi}{n}, \quad j = 1,\dots,n-1. \tag{89}$$

By using the expression (89) for the eigenvalues, one can show (by a direct computation) that the sequence $\{\frac{1}{n}A_n^{[1]}\}$ is distributed like the function $f_1(\theta) = 2-2\cos\theta$ in the sense of the eigenvalues, which is in agreement with Theorem 12. Note that also the sequence $\{K_n^{[1]}\}$ is distributed like f_1 in the sense of the eigenvalues, because $K_n^{[1]}$ is the $(n-1)$-th Toeplitz matrix associated with the function f_1.

For all $n \ge 4$ such that $-1-\frac{\beta}{2n}+\frac{\gamma}{6n^2}$ and $-1+\frac{\beta}{2n}+\frac{\gamma}{6n^2}$ are both negative, by using (89) and some asymptotic expansion, one can prove that

$$\lambda_{\min}\left(\frac{1}{n}A_n^{[1]}\right) \ge 4\left(\sin\frac{\pi}{2n}\right)^2 + \frac{2\gamma}{3n^2}.$$

Moreover, by Gershgorin's first theorem [13], we have $\lambda_{\min}\left(\frac{1}{n}A_n^{[1]}\right) \geq \frac{\gamma}{n^2}$. Hence,

$$\sigma\left(\frac{1}{n}A_n^{[1]}\right) \subset \left[\max\left(4\left(\sin\frac{\pi}{2n}\right)^2 + \frac{2\gamma}{3n^2}, \frac{\gamma}{n^2}\right), 4 + \frac{\gamma}{3n^2}\right).$$

This gives a sharper lower bound for $\lambda_{\min}(\frac{1}{n}A_n^{[p]})$ than the one provided in Theorem 9.

From (89) it also follows that

$$n^2\lambda_{\min}\left(\frac{1}{n}A_n^{[1]}\right) = n^2\lambda_{n-1}\left(\frac{1}{n}A_n^{[1]}\right) \overset{n\to\infty}{\longrightarrow} \pi^2 + \gamma + \frac{\beta^2}{4},$$

$$n^2\left(4 - \lambda_{\max}\left(\frac{1}{n}A_n^{[1]}\right)\right) = n^2\left(4 - \lambda_1\left(\frac{1}{n}A_n^{[1]}\right)\right) \overset{n\to\infty}{\longrightarrow} \pi^2 - \frac{\gamma}{3} + \frac{\beta^2}{4}.$$

In particular, $\{\frac{1}{n}A_n^{[1]}\}$ is strongly clustered at $[0,4]$ according to Definition 2. Note that $[0,4]$ is precisely the range of the function $f_1(\theta) = 2 - 2\cos\theta$ (cfr. Theorem 12).

We conclude this subsection by collecting in the next lemma some results which can be derived by the Gershgorin theorems and will be used in later sections.

Lemma 14 *For all $n \geq 4$,*

- *$H_n^{[1]}$ is skew-symmetric, irreducible, and $\sigma(H_n^{[1]}) \subset \{0\} \times (-1,1)$;*
- *$M_n^{[1]}$ is symmetric, irreducible, and $\sigma(M_n^{[1]}) \subset \left(\frac{1}{3},1\right)$.*

4.6 The quadratic case $p = 2$

The spectral analysis of $\frac{1}{n}A_n^{[1]}$ has not been difficult because Theorem 13 provided us with the explicit expression (89) for the eigenvalues of $\frac{1}{n}A_n^{[1]}$. For $p \geq 2$ such an expression for the eigenvalues of $\frac{1}{n}A_n^{[p]}$ is not available and so our spectral analysis must rely on other considerations.

In the case $p = 2$, for every $n \geq 5$ the matrix $\frac{1}{n}A_n^{[2]}$ is of size $n \times n$ and is given by

$$\frac{1}{n}A_n^{[2]} = K_n^{[2]} + \frac{\beta}{n}H_n^{[2]} + \frac{\gamma}{n^2}M_n^{[2]},$$

where

$$K_n^{[2]} = \frac{1}{6}\begin{bmatrix} 8 & -1 & -1 & & & & \\ -1 & 6 & -2 & -1 & & & \\ -1 & -2 & 6 & -2 & -1 & & \\ & \ddots & \ddots & \ddots & \ddots & \ddots & \\ & & -1 & -2 & 6 & -2 & -1 \\ & & & -1 & -2 & 6 & -1 \\ & & & & -1 & -1 & 8 \end{bmatrix}, \quad H_n^{[2]} = \frac{1}{24}\begin{bmatrix} 0 & 9 & 1 & & & & \\ -9 & 0 & 10 & 1 & & & \\ -1 & -10 & 0 & 10 & 1 & & \\ & \ddots & \ddots & \ddots & \ddots & \ddots & \\ & & -1 & -10 & 0 & 10 & 1 \\ & & & -1 & -10 & 0 & 9 \\ & & & & -1 & -9 & 0 \end{bmatrix},$$

$$M_n^{[2]} = \frac{1}{120}\begin{bmatrix} 40 & 25 & 1 & & & & \\ 25 & 66 & 26 & 1 & & & \\ 1 & 26 & 66 & 26 & 1 & & \\ & \ddots & \ddots & \ddots & \ddots & \ddots & \\ & & 1 & 26 & 66 & 26 & 1 \\ & & & 1 & 26 & 66 & 25 \\ & & & & 1 & 25 & 40 \end{bmatrix}.$$

Theorem 12 reads in the case $p=2$ as $\{\frac{1}{n}A_n^{[2]}\}\overset{\lambda}{\sim} f_2$, with

$$f_2(\theta)=1-\frac{2}{3}\cos\theta-\frac{1}{3}\cos(2\theta)=(2-2\cos\theta)\left(\frac{2}{3}+\frac{1}{3}\cos\theta\right).$$

Moreover, $\{\frac{1}{n}A_n^{[2]}\}$ is strongly clustered at $\left[0,\frac{3}{2}\right]$, which is the range of f_2. In the next subsections we provide more specific results about the spectral properties of $\frac{1}{n}A_n^{[2]}$.

4.6.1 Localization of the eigenvalues

We are now looking for a good localization of $\sigma\left(\frac{1}{n}A_n^{[2]}\right)$. In order to prove Theorem 14, we need some auxiliary lemmas. Using the Gershgorin theorems, we can derive the following bounds for the spectra of the matrices $K_n^{[2]}$, $H_n^{[2]}$ and $M_n^{[2]}$.

Lemma 15 *For all $n\geq 5$,*

- *$K_n^{[2]}$ is symmetric, irreducible, and $\sigma(K_n^{[2]})\subset(0,2)$;*
- *$H_n^{[2]}$ is skew-symmetric, irreducible, and $\sigma(H_n^{[2]})\subset\{0\}\times\left(-\frac{11}{12},\frac{11}{12}\right)$;*
- *$M_n^{[2]}$ is symmetric, irreducible, and $\sigma(M_n^{[2]})\subset\left(\frac{1}{10},1\right)$;*
- *if $\frac{25\gamma}{120n^2}<\frac{1}{6}$, then $K_n^{[2]}+\frac{\gamma}{n^2}M_n^{[2]}$ is symmetric, irreducible, and*

$$\sigma\left(K_n^{[2]}+\frac{\gamma}{n^2}M_n^{[2]}\right)\subset\left(\frac{\gamma}{n^2},2+\frac{\gamma}{10n^2}\right).$$

Remark 9 Lemma 15 implies that $\lambda_{\min}(M_n^{[2]})>\frac{1}{10}$ for all $n\geq 5$. From Remark 6 and Theorem 8 it follows that

$$\lambda_{\min}(K_n^{[2]})>\frac{\pi^2}{10n^2},\quad \forall n\geq 5. \tag{90}$$

Moreover, by Remark 6, we have $\left\|\sum_{i=1}^{n}x_iN_{i+1,[2]}\right\|^2_{L_2([0,1])}\geq\frac{\|\mathbf{x}\|^2}{10n}$ for all $n\geq 5$ and all $\mathbf{x}\in\mathbb{R}^n$.

The next lemma concerns the low-rank matrix $R_n^{[2]}$ introduced in (82).

Lemma 16 *For every $n\geq 5$, we have* $R_n^{[2]}=\frac{1}{6}\begin{bmatrix}2&1&&&\\1&0&&&\\&&&&\\&&&0&1\\&&&1&2\end{bmatrix}\in\mathbb{R}^{n\times n}$*, and its characteristic polynomial is $\frac{1}{1296}\lambda^{n-4}(36\lambda^2-12\lambda-1)^2$. Hence, the eigenvalues of $R_n^{[2]}$ are $\frac{1+\sqrt{2}}{6}$ (with multiplicity 2), $\frac{1-\sqrt{2}}{6}$ (with multiplicity 2) and 0 (with multiplicity $n-4$).*

Theorem 14 *For every $n\geq 5$ such that $\frac{25\gamma}{120n^2}<\frac{1}{6}$,*

$$\begin{aligned}\sigma\left(\frac{1}{n}A_n^{[2]}\right)\subset\Bigg(&\max\left(\frac{\gamma}{n^2},\frac{\pi^2+\gamma}{10n^2}\right),\min\left(\frac{3}{2}+\frac{1+\sqrt{2}}{6}+\frac{\gamma}{n^2},2+\frac{\gamma}{10n^2}\right)\Bigg)\\ &\times\left[-\frac{11|\beta|}{12n},\frac{11|\beta|}{12n}\right]\subset\mathbb{C}.\end{aligned} \tag{91}$$

Proof Fix $n \geq 5$ such that the condition $\frac{25\gamma}{120n^2} < \frac{1}{6}$ is met. By computing the real and imaginary part of $\frac{1}{n}A_n^{[2]}$, we obtain

$$\operatorname{Re}\frac{1}{n}A_n^{[2]} = K_n^{[2]} + \frac{\gamma}{n^2}M_n^{[2]} = B_n^{[2]} + R_n^{[2]} + \frac{\gamma}{n^2}M_n^{[2]}, \qquad \operatorname{Im}\frac{1}{n}A_n^{[2]} = \frac{\beta}{\mathrm{i}n}H_n^{[2]}.$$

We aim at localizing the spectra $\sigma\left(\operatorname{Re}\frac{1}{n}A_n^{[2]}\right)$ and $\sigma\left(\operatorname{Im}\frac{1}{n}A_n^{[2]}\right)$.

We begin with $\sigma\left(\operatorname{Re}\frac{1}{n}A_n^{[2]}\right)$. Since n satisfies the condition $\frac{25\gamma}{120n^2} < \frac{1}{6}$, by Lemma 15 we have

$$\sigma\left(\operatorname{Re}\frac{1}{n}A_n^{[2]}\right) \subset \left(\frac{\gamma}{n^2}, 2 + \frac{\gamma}{10n^2}\right). \tag{92}$$

We can improve (92) as follows. By combining the minimax principle with Lemmas 12, 15 and 16, and taking into account that $M_{f_2} = \frac{3}{2}$, we obtain

$$\begin{aligned}\lambda_{\max}\left(\operatorname{Re}\frac{1}{n}A_n^{[2]}\right) &= \lambda_{\max}\left(B_n^{[2]} + R_n^{[2]} + \frac{\gamma}{n^2}M_n^{[2]}\right) \leq \lambda_{\max}(B_n^{[2]}) + \lambda_{\max}(R_n^{[2]}) + \frac{\gamma}{n^2}\lambda_{\max}(M_n^{[2]}) \\ &< \frac{3}{2} + \frac{1+\sqrt{2}}{6} + \frac{\gamma}{n^2}.\end{aligned}$$

Similarly, by using (90) and Lemma 15,

$$\lambda_{\min}\left(\operatorname{Re}\frac{1}{n}A_n^{[2]}\right) = \lambda_{\min}\left(K_n^{[2]} + \frac{\gamma}{n^2}M_n^{[2]}\right) \geq \lambda_{\min}(K_n^{[2]}) + \frac{\gamma}{n^2}\lambda_{\min}(M_n^{[2]}) > \frac{\pi^2+\gamma}{10n^2}.$$

Thus, we can replace (92) with

$$\sigma\left(\operatorname{Re}\frac{1}{n}A_n^{[2]}\right) \subset \left(\max\left(\frac{\gamma}{n^2}, \frac{\pi^2+\gamma}{10n^2}\right), \min\left(\frac{3}{2} + \frac{1+\sqrt{2}}{6} + \frac{\gamma}{n^2}, 2 + \frac{\gamma}{10n^2}\right)\right). \tag{93}$$

Now we localize the spectrum $\sigma\left(\operatorname{Im}A_n^{[2]}\right)$. Since $\operatorname{Im}\frac{1}{n}A_n^{[2]}$ is Hermitian, from Lemma 15 we obtain[4]

$$\sigma\left(\operatorname{Im}\frac{1}{n}A_n^{[2]}\right) \subset \left[-\frac{11|\beta|}{12n}, \frac{11|\beta|}{12n}\right]. \tag{94}$$

Combining (93)–(94) with (6), we obtain (91). □

4.6.2 Clustering

We are now dealing with the clustering properties of the sequence $\{\frac{1}{n}A_n^{[2]}\}$. We have already mentioned that $\{\frac{1}{n}A_n^{[2]}\}$ is strongly clustered at $\left[0, \frac{3}{2}\right]$, but we have no bounds on the number of outliers, i.e., those eigenvalues of $\frac{1}{n}A_n^{[2]}$ lying outside the ε-expansion $[0, \frac{3}{2}]_\varepsilon = \left[-\varepsilon, \frac{3}{2}+\varepsilon\right] \times [-\varepsilon, \varepsilon]$. Theorem 17 provides an estimate for the number of outliers, and its proof requires the following two results from numerical linear algebra. The first result is the classical interlacing principle, see e.g. [11].

[4] If $\beta \neq 0$ then $\operatorname{Im}\frac{1}{n}A_n^{[2]}$ is irreducible and $\sigma\left(\operatorname{Im}\frac{1}{n}A_n^{[2]}\right) \subset \left(-\frac{11|\beta|}{12n}, \frac{11|\beta|}{12n}\right)$. In (94) we have included the endpoints $\pm\frac{11|\beta|}{12n}$ to cover the case $\beta = 0$.

Theorem 15 *Let $K := B + R$, where $B \in \mathbb{C}^{m\times m}$ is Hermitian and*

$$R := \sum_{j=1}^{k^+} r_j \mathbf{u}_j \mathbf{u}_j^* + \sum_{j=1}^{k^-} t_j \mathbf{v}_j \mathbf{v}_j^*,$$

with $r_j > 0$ for each $j = 1,\dots,k^+$, $t_j < 0$ for each $j = 1,\dots,k^-$ and $\mathbf{u}_1,\dots,\mathbf{u}_{k^+},\mathbf{v}_1,\dots,\mathbf{v}_{k^-} \in \mathbb{C}^m\backslash\{\mathbf{0}\}$. Let

$$\lambda_1(B) \ge \dots \ge \lambda_m(B) \quad \textit{and} \quad \lambda_1(K) \ge \dots \ge \lambda_m(K)$$

be the eigenvalues of B and K arranged in decreasing order. Then

$$\lambda_{j-k^+}(B) \ge \lambda_j(K) \ge \lambda_{j+k^-}(B),$$

for every $j = k^+ + 1,\dots,m-k^-$.

The second result is the Ky-Fan theorem [11, Proposition III.5.3].

Theorem 16 (Ky-Fan) *Let $A \in \mathbb{C}^{m\times m}$ and let $\lambda_j(A)$ and $\lambda_j(\mathrm{Re}A)$, $j = 1,\dots,m$, be the eigenvalues of A and $\mathrm{Re}A$, respectively, arranged in decreasing order:*

$$\mathrm{Re}(\lambda_1(A)) \ge \dots \ge \mathrm{Re}(\lambda_m(A)) \quad \textit{and} \quad \lambda_1(\mathrm{Re}A) \ge \dots \ge \lambda_m(\mathrm{Re}A).$$

Then

$$\sum_{j=1}^{k} \mathrm{Re}(\lambda_j(A)) \le \sum_{j=1}^{k} \lambda_j(\mathrm{Re}A),$$

for every $k = 1,\dots,m$. For $k = m$, the equality holds.

Theorem 17 *For all $\varepsilon \in (0,1)$ and $n \ge \max\left(5, \frac{\sqrt{2}\gamma}{\varepsilon}\right)$, it holds*

$$q_n^+(\varepsilon) \le \left\lceil \frac{1+\sqrt{2}}{3\varepsilon} \right\rceil, \tag{95}$$

where $q_n^+(\varepsilon)$ is the number of eigenvalues of $\frac{1}{n}A_n^{[2]}$ whose real parts are $\ge \frac{3}{2} + \varepsilon$.

Proof For every $n \ge 5$, we consider again the decomposition $K_n^{[2]} = B_n^{[2]} + R_n^{[2]}$ introduced in (82). The matrix $R_n^{[2]}$ is symmetric and we know the eigenvalues of $R_n^{[2]}$ from Lemma 16. By the spectral (Schur) decomposition of $R_n^{[2]}$ we see that

$$R_n^{[2]} = \frac{1+\sqrt{2}}{6}\mathbf{u}_1\mathbf{u}_1^* + \frac{1+\sqrt{2}}{6}\mathbf{u}_2\mathbf{u}_2^* + \frac{1-\sqrt{2}}{6}\mathbf{v}_1\mathbf{v}_1^* + \frac{1-\sqrt{2}}{6}\mathbf{v}_2\mathbf{v}_2^*,$$

where $\mathbf{u}_1,\mathbf{u}_2,\mathbf{v}_1,\mathbf{v}_2 \in \mathbb{C}^n$ are orthonormal vectors. Hence, by Theorem 15,

$$\lambda_{j-2}(B_n^{[2]}) \ge \lambda_j(K_n^{[2]}) \ge \lambda_{j+2}(B_n^{[2]}),$$

for every $j = 3,\dots,n-2$, where the eigenvalues of $B_n^{[2]}$ and $K_n^{[2]}$ are arranged in decreasing order. In particular, from Lemma 12 and $M_{f_2} = \frac{3}{2}$, it follows that $\sigma(B_n^{[2]}) \subset \left(0, \frac{3}{2}\right)$, and

$$\frac{3}{2} > \lambda_1(B_n^{[2]}) \ge \lambda_3(K_n^{[2]}) \ge \dots \ge \lambda_n(K_n^{[2]}) > 0, \tag{96}$$

where the last inequality is a consequence of Lemma 15. Moreover, by the minimax principle,

$$\lambda_{\max}(K_n^{[2]}) = \lambda_{\max}(B_n^{[2]} + R_n^{[2]}) \le \lambda_{\max}(B_n^{[2]}) + \lambda_{\max}(R_n^{[2]}) < \frac{3}{2} + \frac{1+\sqrt{2}}{6}. \tag{97}$$

Assume that the eigenvalues of $\frac{1}{n}A_n^{[2]}$ and $\operatorname{Re}\frac{1}{n}A_n^{[2]}$ are arranged in decreasing order:

$$\operatorname{Re}\left(\lambda_1\left(\frac{1}{n}A_n^{[2]}\right)\right) \ge \ldots \ge \operatorname{Re}\left(\lambda_n\left(\frac{1}{n}A_n^{[2]}\right)\right),$$

and

$$\lambda_1\left(\operatorname{Re}\frac{1}{n}A_n^{[2]}\right) \ge \ldots \ge \lambda_n\left(\operatorname{Re}\frac{1}{n}A_n^{[2]}\right).$$

Recalling from Lemma 15 that $\sigma(M_n^{[2]}) \subset \left(\frac{1}{10}, 1\right)$ and applying again the minimax principle, for every $j = 1, \ldots, n$ we have

$$\begin{aligned}\lambda_j(K_n^{[2]}) &= \min_{\substack{V\subseteq\mathbb{C}^n\\ \dim V=n+1-j}} \max_{\substack{\mathbf{x}\in V\\ \|\mathbf{x}\|=1}} (\mathbf{x}^* K_n^{[2]}\mathbf{x}) = \min_{\substack{V\subseteq\mathbb{C}^n\\ \dim V=n+1-j}} \max_{\substack{\mathbf{x}\in V\\ \|\mathbf{x}\|=1}} \left(\mathbf{x}^*\left(\operatorname{Re}\frac{1}{n}A_n^{[2]} - \frac{\gamma}{n^2}M_n^{[2]}\right)\mathbf{x}\right)\\ &> \min_{\substack{V\subseteq\mathbb{C}^n\\ \dim V=n+1-j}} \max_{\substack{\mathbf{x}\in V\\ \|\mathbf{x}\|=1}} \left(\mathbf{x}^*\left(\operatorname{Re}\frac{1}{n}A_n^{[2]}\right)\mathbf{x} - \frac{\gamma}{n^2}\right) = \lambda_j\left(\operatorname{Re}\frac{1}{n}A_n^{[2]}\right) - \frac{\gamma}{n^2},\end{aligned}$$

and so

$$\lambda_j\left(\operatorname{Re}\frac{1}{n}A_n^{[2]}\right) < \lambda_j(K_n^{[2]}) + \frac{\gamma}{n^2}, \quad \forall j = 1, \ldots, n. \tag{98}$$

Now fix $\varepsilon > 0$ and let $q_n^+(\varepsilon)$ be the number of eigenvalues of $\frac{1}{n}A_n^{[2]}$ whose real parts are greater than or equal to $\frac{3}{2} + \varepsilon$. Following the argument used in [24, proof of Theorem 3.5] and keeping in mind (96)–(98), we apply Theorem 16 to obtain

$$\begin{aligned}\left(\frac{3}{2}+\varepsilon\right)q_n^+(\varepsilon) &\le \sum_{j=1}^{q_n^+(\varepsilon)} \operatorname{Re}\left(\lambda_j\left(\frac{1}{n}A_n^{[2]}\right)\right) \le \sum_{j=1}^{q_n^+(\varepsilon)} \lambda_j\left(\operatorname{Re}\frac{1}{n}A_n^{[2]}\right) \le \sum_{j=1}^{q_n^+(\varepsilon)} \left(\lambda_j(K_n^{[2]}) + \frac{\gamma}{n^2}\right)\\ &= \sum_{j=1}^{q_n^+(\varepsilon)} \lambda_j(K_n^{[2]}) + \frac{\gamma q_n^+(\varepsilon)}{n^2} = \lambda_1(K_n^{[2]}) + \lambda_2(K_n^{[2]}) + \sum_{j=3}^{q_n^+(\varepsilon)} \lambda_j(K_n^{[2]}) + \frac{\gamma q_n^+(\varepsilon)}{n^2}\\ &< 2\left(\frac{3}{2} + \frac{1+\sqrt{2}}{6}\right) + (q_n^+(\varepsilon)-2)\frac{3}{2} + \frac{\gamma q_n^+(\varepsilon)}{n^2} = \left(\frac{3}{2}+\frac{\gamma}{n^2}\right)q_n^+(\varepsilon) + \frac{1+\sqrt{2}}{3},\end{aligned}$$

and so, for every $\varepsilon > 0$ and $n \ge 5$ such that $\frac{\gamma}{n^2} < \varepsilon$ we have

$$q_n^+(\varepsilon) < \frac{1+\sqrt{2}}{3\left(\varepsilon - \frac{\gamma}{n^2}\right)}. \tag{99}$$

If $0 < \varepsilon < 1$ and $n > \max\left(5, \sqrt{\frac{\gamma}{\varepsilon}}\right)$, then

$$\frac{1+\sqrt{2}}{3\left(\varepsilon - \frac{\gamma}{n^2}\right)} \le \frac{1+\sqrt{2}}{3\varepsilon} + 1 \quad \Leftrightarrow \quad n \ge \sqrt{\frac{(1+\sqrt{2}+3\varepsilon)\gamma}{3\varepsilon^2}}.$$

From the inequality

$$\sqrt{\frac{(1+\sqrt{2}+3\varepsilon)\gamma}{3\varepsilon^2}} \leq \frac{\sqrt{2\gamma}}{\varepsilon},$$

and from (99) it follows that (95) holds $\forall\varepsilon \in (0,1)$ and $\forall n \geq \max\left(5, \frac{\sqrt{2\gamma}}{\varepsilon}\right)$. □

Let $q_n(\varepsilon)$ be the number of eigenvalues of $\frac{1}{n}A_n^{[2]}$ lying outside the ε-expansion $\left[0,\frac{3}{2}\right]_\varepsilon$. By combining (91) and (95), we are able to find an upper bound for $q_n(\varepsilon)$. Indeed, $\forall\varepsilon \in (0,1)$ and $\forall n \geq \max\left(5, \frac{11|\beta|}{12\varepsilon}, \frac{\sqrt{2\gamma}}{\varepsilon}\right) = O\left(\frac{1}{\varepsilon}\right)$,

$$q_n(\varepsilon) \leq \left\lceil \frac{1+\sqrt{2}}{3\varepsilon} \right\rceil.$$

Note that, by Theorem 14, $\forall\varepsilon \in (0,1)$, $\forall n \geq \max\left(5, \frac{11|\beta|}{12\varepsilon}, \sqrt{\frac{5\gamma}{4\varepsilon}}\right)$, there are no eigenvalues of $\frac{1}{n}A_n^{[2]}$ lying outside $\left[0,\frac{3}{2}+\frac{1+\sqrt{2}}{6}\right]_\varepsilon$. Thus, $\forall\varepsilon \in (0,1)$, $\forall n \geq \max\left(5, \frac{11|\beta|}{12\varepsilon}, \frac{\sqrt{2\gamma}}{\varepsilon}\right)$, $q_n(\varepsilon)$ is just the number of eigenvalues of $\frac{1}{n}A_n^{[2]}$ lying in

$$\left[0,\frac{3}{2}+\frac{1+\sqrt{2}}{6}\right]_\varepsilon \setminus \left[0,\frac{3}{2}\right]_\varepsilon = \left(\frac{3}{2}+\varepsilon, \frac{3}{2}+\frac{1+\sqrt{2}}{6}+\varepsilon\right] \times [-\varepsilon,\varepsilon].$$

5 The 2D setting

We now consider our model problem (1) on the two-dimensional domain $\Omega = (0,1)^2$. More precisely,

$$\begin{cases} -\Delta u(x,y) + \beta\cdot\nabla u(x,y) + \gamma u(x,y) = \mathrm{f}(x,y), & \forall(x,y)\in\Omega, \\ u(x,y) = 0, & \forall(x,y)\in\partial\Omega, \end{cases} \tag{100}$$

with $\mathrm{f} \in L_2((0,1)^2)$, $\beta = [\beta_1\ \beta_2]^T \in \mathbb{R}^2$, $\gamma \geq 0$. In order to approximate the weak solution of problem (100) by means of the Galerkin method (4), the approximation space $\mathscr{W}$ is chosen as the space of smooth tensor-product splines that we now describe.

We consider two univariate B-spline bases as defined in Section 4 (for the x and y directions):

- the B-spline basis $\{N_{i,[p_1]}(x),\ i = 1,\dots,n_1+p_1\}$ over the knot sequence

  $$s_1 = \ldots = s_{p_1+1} = 0 < s_{p_1+2} < \ldots < s_{p_1+n_1} < 1 = s_{p_1+n_1+1} = \ldots = s_{2p_1+n_1+1},$$

  where

  $$s_{p_1+i+1} := \frac{i}{n_1}, \quad \forall i = 0,\dots,n_1;$$

- the B-spline basis $\{N_{i,[p_2]}(y),\ i = 1,\dots,n_2+p_2\}$ over the knot sequence

  $$t_1 = \ldots = t_{p_2+1} = 0 < t_{p_2+2} < \ldots < t_{p_2+n_2} < 1 = t_{p_2+n_2+1} = \ldots = t_{2p_2+n_2+1},$$

  where

  $$t_{p_2+i+1} := \frac{i}{n_2}, \quad \forall i = 0,\dots,n_2.$$

The bivariate tensor-product B-spline basis $\{N_{i,j,[p_1,p_2]},\ i=1,\dots,n_1+p_1,\ j=1,\dots,n_2+p_2\}$ is given by

$$N_{i,j,[p_1,p_2]}(x,y) := \left(N_{i,[p_1]} \otimes N_{j,[p_2]}\right)(x,y) = N_{i,[p_1]}(x)N_{j,[p_2]}(y).$$

We choose the space $\mathscr{W}_{n_1,n_2}^{[p_1,p_2]}$ as approximation space $\mathscr{W}$ in the Galerkin problem (4), where

$$\mathscr{W}_{n_1,n_2}^{[p_1,p_2]} := \langle N_{i,j,[p_1,p_2]} :\ i=2,\dots,n_1+p_1-1,\ j=2,\dots,n_2+p_2-1\rangle, \tag{101}$$

and we consider the elements of the basis (101) ordered as follows:

$$\varphi_{(n_1+p_1-2)(j-1)+i} = N_{i+1,j+1,[p_1,p_2]}, \quad i=1,\dots,n_1+p_1-2,\ j=1,\dots,n_2+p_2-2. \tag{102}$$

Once we have fixed the tensor-product B-spline basis ordered as in (102), the Galerkin problem (4) leads to a linear system (5). The stiffness matrix A in (5) is the object of our interest and, from now onwards, will be denoted by $A_{n_1,n_2}^{[p_1,p_2]}$ in order to emphasize its dependence on n_1,n_2 and p_1,p_2:

$$A_{n_1,n_2}^{[p_1,p_2]} := A = [a(\varphi_j,\varphi_i)]_{i,j=1}^{(n_1+p_1-2)(n_2+p_2-2)}, \tag{103}$$

where in this case $a(u,v) = \int_0^1\int_0^1 \nabla u\cdot\nabla v\,\mathrm{d}x\mathrm{d}y + \beta\cdot\int_0^1\int_0^1 \nabla u\ v\,\mathrm{d}x\mathrm{d}y + \gamma\int_0^1\int_0^1 uv\,\mathrm{d}x\mathrm{d}y$, see (3).

5.1 Construction of the matrices $A_{n_1,n_2}^{[p_1,p_2]}$

Using the integration rules described in Section 4.1, we obtain that

$$A_{n_1,n_2}^{[p_1,p_2]} = \frac{n_1}{n_2}\widehat{K}_{n_1,n_2}^{[p_1,p_2]} + \frac{n_2}{n_1}\widetilde{K}_{n_1,n_2}^{[p_1,p_2]} + \frac{\beta_1}{n_2}\widehat{H}_{n_1,n_2}^{[p_1,p_2]} + \frac{\beta_2}{n_1}\widetilde{H}_{n_1,n_2}^{[p_1,p_2]} + \frac{\gamma}{n_1n_2}M_{n_1,n_2}^{[p_1,p_2]}, \tag{104}$$

where

$$\widehat{K}_{n_1,n_2}^{[p_1,p_2]} := M_{n_2}^{[p_2]} \otimes K_{n_1}^{[p_1]}, \qquad \widetilde{K}_{n_1,n_2}^{[p_1,p_2]} := K_{n_2}^{[p_2]} \otimes M_{n_1}^{[p_1]},$$

$$\widehat{H}_{n_1,n_2}^{[p_1,p_2]} := M_{n_2}^{[p_2]} \otimes H_{n_1}^{[p_1]}, \qquad \widetilde{H}_{n_1,n_2}^{[p_1,p_2]} := H_{n_2}^{[p_2]} \otimes M_{n_1}^{[p_1]},$$

$$M_{n_1,n_2}^{[p_1,p_2]} := M_{n_2}^{[p_2]} \otimes M_{n_1}^{[p_1]}.$$

In particular, for the case $n_1=n_2=n$ and $p_1=p_2=p$,

$$A_{n,n}^{[p,p]} = K_{n,n}^{[p,p]} + \frac{\beta_1}{n}\widehat{H}_{n,n}^{[p,p]} + \frac{\beta_2}{n}\widetilde{H}_{n,n}^{[p,p]} + \frac{\gamma}{n^2}M_{n,n}^{[p,p]}, \tag{105}$$

with $K_{n,n}^{[p,p]} := \widehat{K}_{n,n}^{[p,p]} + \widetilde{K}_{n,n}^{[p,p]}$.

5.2 Spectral distribution

We will now study, for fixed $p_1, p_2 \geq 1$, the spectral distribution of the sequence of matrices (104) under the additional mild assumption that the ratio $\frac{n_2}{n_1} =: \nu$ is constant as $n_1 \to \infty$.[5] With this assumption we have

$$A_{n_1,n_2}^{[p_1,p_2]} = \frac{1}{\nu}\widehat{K}_{n_1,n_2}^{[p_1,p_2]} + \nu\widetilde{K}_{n_1,n_2}^{[p_1,p_2]} + \frac{\beta_1}{\nu n_1}\widehat{H}_{n_1,n_2}^{[p_1,p_2]} + \frac{\beta_2}{n_1}\widetilde{H}_{n_1,n_2}^{[p_1,p_2]} + \frac{\gamma}{\nu(n_1)^2}M_{n_1,n_2}^{[p_1,p_2]}. \tag{106}$$

For every $n_1 \geq 3p_1+1$ such that $n_2 = \nu n_1 \geq 3p_2+1$, we decompose the matrices $\widehat{K}_{n_1,n_2}^{[p_1,p_2]}$ and $\widetilde{K}_{n_1,n_2}^{[p_1,p_2]}$ into

$$\widehat{K}_{n_1,n_2}^{[p_1,p_2]} = \widehat{B}_{n_1,n_2}^{[p_1,p_2]} + \widehat{R}_{n_1,n_2}^{[p_1,p_2]}, \qquad \widetilde{K}_{n_1,n_2}^{[p_1,p_2]} = \widetilde{B}_{n_1,n_2}^{[p_1,p_2]} + \widetilde{R}_{n_1,n_2}^{[p_1,p_2]}, \tag{107}$$

where

$$\widehat{B}_{n_1,n_2}^{[p_1,p_2]} := C_{n_2}^{[p_2]} \otimes B_{n_1}^{[p_1]}, \qquad \widetilde{B}_{n_1,n_2}^{[p_1,p_2]} := B_{n_2}^{[p_2]} \otimes C_{n_1}^{[p_1]},$$

and

$$\begin{aligned}\widehat{R}_{n_1,n_2}^{[p_1,p_2]} &:= \widehat{K}_{n_1,n_2}^{[p_1,p_2]} - \widehat{B}_{n_1,n_2}^{[p_1,p_2]} = C_{n_2}^{[p_2]} \otimes R_{n_1}^{[p_1]} + S_{n_2}^{[p_2]} \otimes B_{n_1}^{[p_1]} + S_{n_2}^{[p_2]} \otimes R_{n_1}^{[p_1]},\\ \widetilde{R}_{n_1,n_2}^{[p_1,p_2]} &:= \widetilde{K}_{n_1,n_2}^{[p_1,p_2]} - \widetilde{B}_{n_1,n_2}^{[p_1,p_2]} = B_{n_2}^{[p_2]} \otimes S_{n_1}^{[p_1]} + R_{n_2}^{[p_2]} \otimes C_{n_1}^{[p_1]} + R_{n_2}^{[p_2]} \otimes S_{n_1}^{[p_1]}.\end{aligned}$$

We recall that the matrices $B_n^{[p]}$, $R_n^{[p]}$, $C_n^{[p]}$, $S_n^{[p]}$ were introduced in Section 4.4, see (82)–(85). Finally, we define

$$B_{n_1,n_2}^{[p_1,p_2]} := \frac{1}{\nu}\widehat{B}_{n_1,n_2}^{[p_1,p_2]} + \nu\widetilde{B}_{n_1,n_2}^{[p_1,p_2]}, \tag{108}$$

$$R_{n_1,n_2}^{[p_1,p_2]} := \frac{1}{\nu}\widehat{R}_{n_1,n_2}^{[p_1,p_2]} + \nu\widetilde{R}_{n_1,n_2}^{[p_1,p_2]}. \tag{109}$$

From Lemmas 12 and 13 we know that $B_n^{[p]} = T_{n+p-2}(f_p)$ and $C_n^{[p]} = T_{n+p-2}(h_p)$ for $p \geq 1$ and $n \geq 3p+1$. By Lemma 2 we then obtain

$$\begin{aligned}\widehat{B}_{n_1,n_2}^{[p_1,p_2]} &= T_{n_2+p_2-2}(h_{p_2}) \otimes T_{n_1+p_1-2}(f_{p_1}) = T_{n_2+p_2-2,n_1+p_1-2}(h_{p_2} \otimes f_{p_1}),\\ \widetilde{B}_{n_1,n_2}^{[p_1,p_2]} &= T_{n_2+p_2-2}(f_{p_2}) \otimes T_{n_1+p_1-2}(h_{p_1}) = T_{n_2+p_2-2,n_1+p_1-2}(f_{p_2} \otimes h_{p_1}),\end{aligned}$$

and

$$B_{n_1,n_2}^{[p_1,p_2]} = T_{n_2+p_2-2,n_1+p_1-2}\left(\frac{1}{\nu}h_{p_2} \otimes f_{p_1} + \nu f_{p_2} \otimes h_{p_1}\right). \tag{110}$$

Hence, by Theorem 5,

$$\{\widehat{B}_{n_1,n_2}^{[p_1,p_2]}\} \overset{\lambda}{\sim} h_{p_2} \otimes f_{p_1}, \qquad \{\widetilde{B}_{n_1,n_2}^{[p_1,p_2]}\} \overset{\lambda}{\sim} f_{p_2} \otimes h_{p_1},$$

[5] In this way, $A_{n_1,n_2}^{[p_1,p_2]}$ is really a sequence of matrices, since only n_1 is a free parameter. The relation $n_2 = \nu n_1$ must be kept in mind while reading this section. We point out that this request could be replaced by even milder conditions, but at the price of heavier notations.
If n_1 and n_2 are not proportional, i.e., n_2/n_1 does not converge to a positive constant, then:

- either a distribution does not exist (when n_2/n_1 does not have a limit),
- or it exists, but this distribution completely ignores the differential operator in x (if n_1/n_2 converges to zero) or the operator in y (if n_2/n_1 converges to zero).

and

$$\{B^{[p_1,p_2]}_{n_1,n_2}\} \overset{\lambda}{\sim} \frac{1}{\nu} h_{p_2} \otimes f_{p_1} + \nu f_{p_2} \otimes h_{p_1}. \tag{111}$$

By Lemma 1 and the inequalities (83), (85), the two matrices $\widehat{R}^{[p_1,p_2]}_{n_1,n_2}$ and $\widetilde{R}^{[p_1,p_2]}_{n_1,n_2}$ satisfy

$$\begin{aligned}\mathrm{rank}(\widehat{R}^{[p_1,p_2]}_{n_1,n_2}) &\leq \mathrm{rank}(C^{[p_2]}_{n_2} \otimes R^{[p_1]}_{n_1}) + \mathrm{rank}(S^{[p_2]}_{n_2} \otimes B^{[p_1]}_{n_1}) + \mathrm{rank}(S^{[p_2]}_{n_2} \otimes R^{[p_1]}_{n_1}) \\ &= \mathrm{rank}(C^{[p_2]}_{n_2})\mathrm{rank}(R^{[p_1]}_{n_1}) + \mathrm{rank}(S^{[p_2]}_{n_2})\mathrm{rank}(B^{[p_1]}_{n_1}) + \mathrm{rank}(S^{[p_2]}_{n_2})\mathrm{rank}(R^{[p_1]}_{n_1}) \\ &\leq (\nu n_1 + p_2 - 2)2(2p_1 - 1) + 2(2p_2 - 1)(n_1 + p_1 - 2) + 2(2p_2 - 1)2(2p_1 - 1) \\ &= o((n_1 + p_1 - 2)(\nu n_1 + p_2 - 2)), \quad \text{as } n_1 \to \infty,\end{aligned}$$

and in a similar way we also obtain

$$\mathrm{rank}(\widetilde{R}^{[p_1,p_2]}_{n_1,n_2}) = o((n_1 + p_1 - 2)(\nu n_1 + p_2 - 2)), \quad \text{as } n_1 \to \infty.$$

Thus,

$$\mathrm{rank}(R^{[p_1,p_2]}_{n_1,n_2}) \leq \mathrm{rank}(\widehat{R}^{[p_1,p_2]}_{n_1,n_2}) + \mathrm{rank}(\widetilde{R}^{[p_1,p_2]}_{n_1,n_2}) = o((n_1 + p_1 - 2)(\nu n_1 + p_2 - 2)), \tag{112}$$

as $n_1 \to \infty$. Note that $(n_1 + p_1 - 2)(\nu n_1 + p_2 - 2)$ is the dimension of the matrix $A^{[p_1,p_2]}_{n_1,n_2}$. Moreover, using Lemmas 1, 12, 13 and the fact that the matrices $K^{[p]}_n$, $H^{[p]}_n$, $M^{[p]}_n$, $B^{[p]}_n$, $C^{[p]}_n$ are normal, we obtain

$$\begin{aligned}\|R^{[p_1,p_2]}_{n_1,n_2}\| &\leq \frac{1}{\nu}\|\widehat{R}^{[p_1,p_2]}_{n_1,n_2}\| + \nu\|\widetilde{R}^{[p_1,p_2]}_{n_1,n_2}\| = \frac{1}{\nu}\|\widehat{K}^{[p_1,p_2]}_{n_1,n_2} - \widehat{B}^{[p_1,p_2]}_{n_1,n_2}\| + \nu\|\widetilde{K}^{[p_1,p_2]}_{n_1,n_2} - \widetilde{B}^{[p_1,p_2]}_{n_1,n_2}\| \\ &\leq \frac{1}{\nu}\|\widehat{K}^{[p_1,p_2]}_{n_1,n_2}\| + \frac{1}{\nu}\|\widehat{B}^{[p_1,p_2]}_{n_1,n_2}\| + \nu\|\widetilde{K}^{[p_1,p_2]}_{n_1,n_2}\| + \nu\|\widetilde{B}^{[p_1,p_2]}_{n_1,n_2}\| \\ &= \frac{1}{\nu}\|M^{[p_2]}_{n_2}\|\|K^{[p_1]}_{n_1}\| + \frac{1}{\nu}\|C^{[p_2]}_{n_2}\|\|B^{[p_1]}_{n_1}\| + \nu\|K^{[p_2]}_{n_2}\|\|M^{[p_1]}_{n_1}\| + \nu\|B^{[p_2]}_{n_2}\|\|C^{[p_1]}_{n_1}\| \\ &\leq \frac{1}{\nu}\|M^{[p_2]}_{n_2}\|_\infty\|K^{[p_1]}_{n_1}\|_\infty + \frac{1}{\nu}M_{f_{p_1}} + \nu\|K^{[p_2]}_{n_2}\|_\infty\|M^{[p_1]}_{n_1}\|_\infty + \nu M_{f_{p_2}}.\end{aligned}$$

From Lemma 11 it follows

$$\|R^{[p_1,p_2]}_{n_1,n_2}\| \leq Q_{p_1,p_2}, \tag{113}$$

where Q_{p_1,p_2} is a constant independent of n_1.

Theorem 18 *The sequence of matrices $\{A^{[p_1,p_2]}_{n_1,n_2}\}$ is distributed like the function $g_{p_1,p_2} : [-\pi,\pi]^2 \to \mathbb{R}$,*

$$g_{p_1,p_2} := \frac{1}{\nu} h_{p_2} \otimes f_{p_1} + \nu f_{p_2} \otimes h_{p_1}, \tag{114}$$

in the sense of the eigenvalues.

Proof We have to show that, $\forall F \in C_c(\mathbb{C},\mathbb{C})$,

$$\begin{aligned}\lim_{n_1\to\infty} \frac{1}{(n_1 + p_1 - 2)(\nu n_1 + p_2 - 2)} &\sum_{j=1}^{(n_1+p_1-2)(\nu n_1+p_2-2)} F(\lambda_j(A^{[p_1,p_2]}_{n_1,n_2})) \\ &= \frac{1}{(2\pi)^2}\int_{-\pi}^{\pi}\int_{-\pi}^{\pi} F(g_{p_1,p_2}(\theta_1,\theta_2))\mathrm{d}\theta_1\mathrm{d}\theta_2.\end{aligned}$$

By (106)–(109), we have

$$A_{n_1,n_2}^{[p_1,p_2]} = B_{n_1,n_2}^{[p_1,p_2]} + R_{n_1,n_2}^{[p_1,p_2]} + \frac{\beta_1}{\nu n_1}\widehat{H}_{n_1,n_2}^{[p_1,p_2]} + \frac{\beta_2}{n_1}\widetilde{H}_{n_1,n_2}^{[p_1,p_2]} + \frac{\gamma}{\nu(n_1)^2}M_{n_1,n_2}^{[p_1,p_2]}. \tag{115}$$

Recalling that $n_2 = \nu n_1$ is determined as a function of n_1, we now prove that the hypotheses of Theorem 1 are satisfied with $Z_{n_1} = A_{n_1,n_2}^{[p_1,p_2]}$, $X_{n_1} = B_{n_1,n_2}^{[p_1,p_2]}$ and Y_{n_1} the remaining term in the right-hand side of (115). We have seen in (111) that $\{B_{n_1,n_2}^{[p_1,p_2]}\} \overset{\lambda}{\sim} g_{p_1,p_2}$. Noting that $B_{n_1,n_2}^{[p_1,p_2]}$ is symmetric, by Theorem 5 we obtain

$$\|B_{n_1,n_2}^{[p_1,p_2]}\| = \rho(B_{n_1,n_2}^{[p_1,p_2]}) < M_{g_{p_1,p_2}},$$

where $M_{g_{p_1,p_2}} := \max_{(\theta_1,\theta_2)\in[-\pi,\pi]^2} g_{p_1,p_2}(\theta_1,\theta_2)$ is a constant independent of n_1.

By Lemma 1, we get

$$\begin{aligned}
&\left\|\frac{\beta_1}{\nu n_1}\widehat{H}_{n_1,n_2}^{[p_1,p_2]} + \frac{\beta_2}{n_1}\widetilde{H}_{n_1,n_2}^{[p_1,p_2]} + \frac{\gamma}{\nu(n_1)^2}M_{n_1,n_2}^{[p_1,p_2]}\right\| \\
&\quad\le \frac{|\beta_1|}{\nu n_1}\|\widehat{H}_{n_1,n_2}^{[p_1,p_2]}\| + \frac{|\beta_2|}{n_1}\|\widetilde{H}_{n_1,n_2}^{[p_1,p_2]}\| + \frac{\gamma}{\nu(n_1)^2}\|M_{n_1,n_2}^{[p_1,p_2]}\| \\
&\quad\le \frac{|\beta_1|}{\nu n_1}\|M_{n_2}^{[p_2]}\|\|H_{n_1}^{[p_1]}\| + \frac{|\beta_2|}{n_1}\|H_{n_2}^{[p_2]}\|\|M_{n_1}^{[p_1]}\| + \frac{\gamma}{\nu(n_1)^2}\|M_{n_2}^{[p_2]}\|\|M_{n_1}^{[p_1]}\| \\
&\quad\le \frac{|\beta_1|}{\nu n_1}\|M_{n_2}^{[p_2]}\|_\infty\|H_{n_1}^{[p_1]}\|_\infty + \frac{|\beta_2|}{n_1}\|H_{n_2}^{[p_2]}\|_\infty\|M_{n_1}^{[p_1]}\|_\infty + \frac{\gamma}{\nu(n_1)^2}\|M_{n_2}^{[p_2]}\|_\infty\|M_{n_1}^{[p_1]}\|_\infty,
\end{aligned}$$

and from Lemma 11 it follows that

$$\left\|\frac{\beta_1}{\nu n_1}\widehat{H}_{n_1,n_2}^{[p_1,p_2]} + \frac{\beta_2}{n_1}\widetilde{H}_{n_1,n_2}^{[p_1,p_2]} + \frac{\gamma}{\nu(n_1)^2}M_{n_1,n_2}^{[p_1,p_2]}\right\| = O\left(\frac{1}{n_1}\right). \tag{116}$$

Combining (113) and (116), we obtain

$$\left\|R_{n_1,n_2}^{[p_1,p_2]} + \frac{\beta_1}{\nu n_1}\widehat{H}_{n_1,n_2}^{[p_1,p_2]} + \frac{\beta_2}{n_1}\widetilde{H}_{n_1,n_2}^{[p_1,p_2]} + \frac{\gamma}{\nu(n_1)^2}M_{n_1,n_2}^{[p_1,p_2]}\right\| \le \bar{Q}_{p_1,p_2},$$

where $\bar{Q}_{p_1,p_2}$ is a constant independent of n_1.

On the other hand, by using (112)–(113) and (116), we get

$$\begin{aligned}
&\left\|R_{n_1,n_2}^{[p_1,p_2]} + \frac{\beta_1}{\nu n_1}\widehat{H}_{n_1,n_2}^{[p_1,p_2]} + \frac{\beta_2}{n_1}\widetilde{H}_{n_1,n_2}^{[p_1,p_2]} + \frac{\gamma}{\nu(n_1)^2}M_{n_1,n_2}^{[p_1,p_2]}\right\|_1 \\
&\quad\le \|R_{n_1,n_2}^{[p_1,p_2]}\|_1 + \left\|\frac{\beta_1}{\nu n_1}\widehat{H}_{n_1,n_2}^{[p_1,p_2]} + \frac{\beta_2}{n_1}\widetilde{H}_{n_1,n_2}^{[p_1,p_2]} + \frac{\gamma}{\nu(n_1)^2}M_{n_1,n_2}^{[p_1,p_2]}\right\|_1 \\
&\quad\le \mathrm{rank}(R_{n_1,n_2}^{[p_1,p_2]})\|R_{n_1,n_2}^{[p_1,p_2]}\| \\
&\qquad\quad + (n_1+p_1-2)(\nu n_1+p_2-2)\left\|\frac{\beta_1}{\nu n_1}\widehat{H}_{n_1,n_2}^{[p_1,p_2]} + \frac{\beta_2}{n_1}\widetilde{H}_{n_1,n_2}^{[p_1,p_2]} + \frac{\gamma}{\nu(n_1)^2}M_{n_1,n_2}^{[p_1,p_2]}\right\| \\
&\quad\le \mathrm{rank}(R_{n_1,n_2}^{[p_1,p_2]})Q_{p_1,p_2} + (n_1+p_1-2)(\nu n_1+p_2-2)\,O\left(\frac{1}{n_1}\right) \\
&\quad= o((n_1+p_1-2)(\nu n_1+p_2-2)), \quad \text{as } n_1\to\infty.
\end{aligned}$$

Hence, all the hypotheses of Theorem 1 are satisfied, and the symbol (114) follows. □

In the next two subsections we discuss in more detail the spectral properties of the matrices $A_{n_1,n_2}^{[p_1,p_2]}$ with $n_1 = n_2 = n$ in the cases $p_1 = p_2 = 1$ and $p_1 = p_2 = 2$.

5.3 The bilinear case $p_1 = p_2 = 1$

In the case $p_1 = p_2 = 1$, for every $n_1 = n_2 = n \geq 4$, the matrix $A_{n,n}^{[1,1]}$ is $(n-1)^2 \times (n-1)^2$ and is given by

$$A_{n,n}^{[1,1]} = K_{n,n}^{[1,1]} + \frac{\beta_1}{n}\widehat{H}_{n,n}^{[1,1]} + \frac{\beta_2}{n}\widetilde{H}_{n,n}^{[1,1]} + \frac{\gamma}{n^2}M_{n,n}^{[1,1]}, \tag{117}$$

where the matrices $K_{n,n}^{[1,1]}, \widehat{H}_{n,n}^{[1,1]}, \widetilde{H}_{n,n}^{[1,1]}, M_{n,n}^{[1,1]}$ are described in Section 5.1. Theorem 18 reads in the case $p_1 = p_2 = 1$ as $\{A_{n,n}^{[1,1]}\} \overset{\lambda}{\sim} g_{1,1}$, with

$$\begin{aligned} g_{1,1}(\theta_1,\theta_2) &= (f_1 \otimes h_1)(\theta_1,\theta_2) + (h_1 \otimes f_1)(\theta_1,\theta_2) \\ &= \frac{8}{3} - \frac{2}{3}\cos(\theta_1) - \frac{2}{3}\cos(\theta_2) - \frac{4}{3}\cos(\theta_1)\cos(\theta_2). \end{aligned}$$

5.3.1 Localization of the eigenvalues and clustering

Theorem 19 *For every $n \geq 4$ such that $\frac{\gamma}{9n^2} < \frac{1}{3}$*

$$\begin{aligned} \sigma(A_{n,n}^{[1,1]}) \subset & \left(\max\left(\frac{\gamma}{n^2}, \frac{8}{3}\left(\sin\frac{\pi}{2n}\right)^2 + \frac{\gamma}{9n^2}\right), \min\left(4 + \frac{\gamma}{n^2}, \frac{16}{3} - \frac{\gamma}{9n^2}\right)\right) \\ & \times \left[-\frac{|\beta_1| + |\beta_2|}{n}, \frac{|\beta_1| + |\beta_2|}{n}\right] \subset \mathbb{C}. \end{aligned} \tag{118}$$

Proof Fix $n \geq 4$. By computing the real and imaginary part of $A_{n,n}^{[1,1]}$, we obtain

$$\mathrm{Re}A_{n,n}^{[1,1]} = K_{n,n}^{[1,1]} + \frac{\gamma}{n^2}M_{n,n}^{[1,1]}, \qquad \mathrm{Im}A_{n,n}^{[1,1]} = \frac{\beta_1}{\mathrm{i}n}\widehat{H}_{n,n}^{[1,1]} + \frac{\beta_2}{\mathrm{i}n}\widetilde{H}_{n,n}^{[1,1]}.$$

The target is the localization of $\sigma(\mathrm{Re}A_{n,n}^{[1,1]})$ and $\sigma(\mathrm{Im}A_{n,n}^{[1,1]})$.

We begin with $\sigma(\mathrm{Re}A_{n,n}^{[1,1]})$. Since n satisfies the condition $\frac{\gamma}{9n^2} < \frac{1}{3}$, $\mathrm{Re}A_{n,n}^{[1,1]}$ is Hermitian, irreducible and, by Gershgorin's theorems,

$$\sigma(\mathrm{Re}A_{n,n}^{[1,1]}) \subset \left(\frac{\gamma}{n^2}, \frac{16}{3} - \frac{\gamma}{9n^2}\right).$$

We can improve this range as follows. The matrix $K_{n,n}^{[1,1]}$ is equal to the matrix $B_{n,n}^{[1,1]}$ defined in (108), taking into account that in this case $\nu = \frac{n}{n} = 1$. Therefore, by (110) we obtain

$$K_{n,n}^{[1,1]} = B_{n,n}^{[1,1]} = T_{n-1,n-1}(h_1 \otimes f_1 + f_1 \otimes h_1) = T_{n-1,n-1}(g_{1,1}).$$

The range of $g_{1,1}$ is $[0,4]$ and so, by Theorem 5, $\sigma(K_{n,n}^{[1,1]}) \subset (0,4)$. Moreover, since $M_{n,n}^{[1,1]} = M_n^{[1]} \otimes M_n^{[1]}$, from Lemmas 1 and 14 it follows that $M_{n,n}^{[1,1]}$ is symmetric and that $\sigma(M_{n,n}^{[1,1]}) \subset (\frac{1}{9}, 1)$. By the minimax principle we then have

$$\lambda_{\max}(\mathrm{Re}A_{n,n}^{[1,1]}) = \lambda_{\max}\left(K_{n,n}^{[1,1]} + \frac{\gamma}{n^2}M_{n,n}^{[1,1]}\right) \leq \lambda_{\max}(K_{n,n}^{[1,1]}) + \frac{\gamma}{n^2}\lambda_{\max}(M_{n,n}^{[1,1]}) < 4 + \frac{\gamma}{n^2}.$$

In addition, by the minimax principle, by Lemmas 1 and 14, and by the fact that $\lambda_{\min}(K_n^{[1]}) = 4\left(\sin\frac{\pi}{2n}\right)^2$, we obtain

$$\begin{aligned}
\lambda_{\min}\left(\mathrm{Re}A_{n,n}^{[1,1]}\right) &= \lambda_{\min}\left(K_{n,n}^{[1,1]} + \frac{\gamma}{n^2}M_{n,n}^{[1,1]}\right) \\
&= \lambda_{\min}\left(K_n^{[1]} \otimes M_n^{[1]} + M_n^{[1]} \otimes K_n^{[1]} + \frac{\gamma}{n^2}M_n^{[1]} \otimes M_n^{[1]}\right) \\
&\geq \lambda_{\min}(K_n^{[1]})\lambda_{\min}(M_n^{[1]}) + \lambda_{\min}(M_n^{[1]})\lambda_{\min}(K_n^{[1]}) + \frac{\gamma}{n^2}\,\lambda_{\min}(M_n^{[1]})^2 \\
&> 2\cdot 4\left(\sin\frac{\pi}{2n}\right)^2\frac{1}{3} + \frac{\gamma}{9n^2} = \frac{8}{3}\left(\sin\frac{\pi}{2n}\right)^2 + \frac{\gamma}{9n^2}.
\end{aligned}$$

Therefore, we obtain for $\sigma(\mathrm{Re}A_{n,n}^{[1,1]})$ the localization

$$\sigma(\mathrm{Re}A_{n,n}^{[1,1]}) \subset \left(\max\left(\frac{\gamma}{n^2}, \frac{8}{3}\left(\sin\frac{\pi}{2n}\right)^2 + \frac{\gamma}{9n^2}\right), \min\left(4+\frac{\gamma}{n^2}, \frac{16}{3} - \frac{\gamma}{9n^2}\right)\right). \tag{119}$$

We now localize the spectrum $\sigma(\mathrm{Im}A_{n,n}^{[1,1]})$. The matrices $\widehat{H}_{n,n}^{[1,1]}$ and $\widetilde{H}_{n,n}^{[1,1]}$ are skew-symmetric.[6] As a consequence, the matrices $\mathrm{i}\widehat{H}_{n,n}^{[1,1]}$ and $\mathrm{i}\widetilde{H}_{n,n}^{[1,1]}$ are Hermitian, proving that all the eigenvalues of $\widehat{H}_{n,n}^{[1,1]}$ and $\widetilde{H}_{n,n}^{[1,1]}$ are purely imaginary. Moreover, $\widehat{H}_{n,n}^{[1,1]} = M_n^{[1]} \otimes H_n^{[1]}$ and $\widetilde{H}_{n,n}^{[1,1]} = H_n^{[1]} \otimes M_n^{[1]}$. Thus, by Lemmas 1 and 14, $\sigma(\widehat{H}_{n,n}^{[1,1]}) = \sigma(\widetilde{H}_{n,n}^{[1,1]}) \subset \{0\} \times (-1,1)$. Hence, by the minimax principle,

$$\begin{aligned}
\lambda_{\min}(\mathrm{Im}A_{n,n}^{[1,1]}) &= \lambda_{\min}\left(\frac{\beta_1}{n}\frac{1}{\mathrm{i}}\widehat{H}_{n,n}^{[1,1]} + \frac{\beta_2}{n}\frac{1}{\mathrm{i}}\widetilde{H}_{n,n}^{[1,1]}\right) \\
&\geq \lambda_{\min}\left(\frac{\beta_1}{n}\frac{1}{\mathrm{i}}\widehat{H}_{n,n}^{[1,1]}\right) + \lambda_{\min}\left(\frac{\beta_2}{n}\frac{1}{\mathrm{i}}\widetilde{H}_{n,n}^{[1,1]}\right) \geq -\frac{|\beta_1|}{n} - \frac{|\beta_2|}{n},
\end{aligned}$$

and similarly it can be proved that

$$\lambda_{\max}(\mathrm{Im}A_{n,n}^{[1,1]}) \leq \frac{|\beta_1|}{n} + \frac{|\beta_2|}{n}.$$

Therefore, we obtain for $\sigma(\mathrm{Im}A_{n,n}^{[1,1]})$ the localization

$$\sigma(\mathrm{Im}A_{n,n}^{[1,1]}) \subseteq \left[-\frac{|\beta_1|+|\beta_2|}{n}, \frac{|\beta_1|+|\beta_2|}{n}\right]. \tag{120}$$

Combining (6) with (119)–(120), we obtain (118). □

Theorem 19 shows that $\{A_{n,n}^{[1,1]}\}$ is strongly clustered at $[0,4]$, the range of the function $g_{1,1}$. This is confirmed by the following corollary.

Corollary 3 *$\forall\varepsilon \in (0,1)$ and $\forall n \geq \max\left(4, \sqrt{\frac{\gamma}{\varepsilon}}, \frac{|\beta_1|+|\beta_2|}{\varepsilon}\right)$, we have*

$$q_n(\varepsilon) = 0,$$

where $q_n(\varepsilon)$ is the number of eigenvalues of $A_{n,n}^{[1,1]}$ lying outside $[0,4]_\varepsilon$.

[6] This follows from their definition and from Lemma 1, taking into account that $H_n^{[1]}$ is skew-symmetric, while $M_n^{[1]}$ is symmetric, see Theorem 7.

Proof Fix $\varepsilon \in (0,1)$ and $n \geq \max\left(4, \sqrt{\frac{\gamma}{\varepsilon}}, \frac{|\beta_1|+|\beta_2|}{\varepsilon}\right)$. Since n satisfies the conditions $\frac{\gamma}{9n^2} < \frac{1}{3}$, $\frac{\gamma}{n^2} \leq \varepsilon$ and $\frac{|\beta_1|+|\beta_2|}{n} \leq \varepsilon$, by Theorem 19 we have

$$\sigma(A_{n,n}^{[1,1]}) \subset \left(\frac{\gamma}{n^2}, 4+\frac{\gamma}{n^2}\right) \times \left[-\frac{|\beta_1|+|\beta_2|}{n}, \frac{|\beta_1|+|\beta_2|}{n}\right] \subset [-\varepsilon, 4+\varepsilon] \times [-\varepsilon, \varepsilon] = [0,4]_\varepsilon.$$

Hence, $q_n(\varepsilon) = 0$. □

5.4 The biquadratic case $p_1 = p_2 = 2$

In the case $p_1 = p_2 = 2$, for every $n_1 = n_2 = n \geq 5$, the matrix $A_{n,n}^{[2,2]}$ is $n^2 \times n^2$ and

$$A_{n,n}^{[2,2]} = K_{n,n}^{[2,2]} + \frac{\beta_1}{n}\widehat{H}_{n,n}^{[2,2]} + \frac{\beta_2}{n}\widetilde{H}_{n,n}^{[2,2]} + \frac{\gamma}{n^2}M_{n,n}^{[2,2]},$$

where the matrices $K_{n,n}^{[2,2]}, \widehat{H}_{n,n}^{[2,2]}, \widetilde{H}_{n,n}^{[2,2]}, M_{n,n}^{[2,2]}$ are described in Section 5.1. Theorem 18 reads in the case $p_1 = p_2 = 2$ as $\{A_{n,n}^{[2,2]}\} \overset{\lambda}{\sim} g_{2,2}$, with

$$\begin{aligned} g_{2,2}(\theta_1,\theta_2) &= (f_2 \otimes h_2)(\theta_1,\theta_2) + (h_2 \otimes f_2)(\theta_1,\theta_2) \\ &= \frac{1}{90}[99 + 6\cos(\theta_1) + 6\cos(\theta_2) - 15\cos(2\theta_1) - 15\cos(2\theta_2) - 52\cos(\theta_1)\cos(\theta_2) \\ &\qquad - 14\cos(\theta_1)\cos(2\theta_2) - 14\cos(\theta_2)\cos(2\theta_1) - \cos(2\theta_1)\cos(2\theta_2)]. \end{aligned}$$

5.4.1 Localization of the eigenvalues

Theorem 20 *For every $n \geq 5$ such that $\frac{25\gamma}{120n^2} < \frac{1}{6}$*

$$\begin{aligned} \sigma(A_{n,n}^{[2,2]}) \subset \left(\max\left(\frac{\pi^2+10\gamma}{100n^2}, \frac{2\pi^2+\gamma}{100n^2}\right), \frac{49}{24}+\frac{\gamma}{n^2}\right) \\ \times \left[-\frac{11}{12}\frac{|\beta_1|+|\beta_2|}{n}, \frac{11}{12}\frac{|\beta_1|+|\beta_2|}{n}\right] \subset \mathbb{C}. \end{aligned} \tag{121}$$

Proof Fix $n \geq 5$ such that the condition $\frac{25\gamma}{120n^2} < \frac{1}{6}$ is met. By computing the real and imaginary part of $A_{n,n}^{[2,2]}$ we obtain

$$\mathrm{Re}A_{n,n}^{[2,2]} = K_{n,n}^{[2,2]} + \frac{\gamma}{n^2}M_{n,n}^{[2,2]}, \quad \text{and} \quad \mathrm{Im}A_{n,n}^{[2,2]} = \frac{\beta_1}{\mathrm{i}n}\widehat{H}_{n,n}^{[2,2]} + \frac{\beta_2}{\mathrm{i}n}\widetilde{H}_{n,n}^{[2,2]}.$$

The target is now the localization of $\sigma(\mathrm{Re}A_{n,n}^{[2,2]})$ and $\sigma(\mathrm{Im}A_{n,n}^{[2,2]})$.

First we localize the spectrum of $\mathrm{Re}A_{n,n}^{[2,2]}$. Note that

$$\begin{aligned} \mathrm{Re}A_{n,n}^{[2,2]} &= K_{n,n}^{[2,2]} + \frac{\gamma}{n^2}M_{n,n}^{[2,2]} = M_n^{[2]} \otimes K_n^{[2]} + K_n^{[2]} \otimes M_n^{[2]} + \frac{\gamma}{n^2}M_n^{[2]} \otimes M_n^{[2]} \\ &= M_n^{[2]} \otimes K_n^{[2]} + \left(K_n^{[2]} + \frac{\gamma}{n^2}M_n^{[2]}\right) \otimes M_n^{[2]}. \end{aligned}$$

Therefore, by the minimax principle, by Lemmas 1 and 15, and by (90),

$$\begin{aligned}\lambda_{\min}(\mathrm{Re}A_{n,n}^{[2,2]}) &\geq \lambda_{\min}(M_n^{[2]}\otimes K_n^{[2]})+\lambda_{\min}(K_n^{[2]}\otimes M_n^{[2]})+\frac{\gamma}{n^2}\lambda_{\min}(M_n^{[2]}\otimes M_n^{[2]})\\ &= \lambda_{\min}(M_n^{[2]})\lambda_{\min}(K_n^{[2]})+\lambda_{\min}(K_n^{[2]})\lambda_{\min}(M_n^{[2]})+\frac{\gamma}{n^2}\lambda_{\min}(M_n^{[2]})\lambda_{\min}(M_n^{[2]})\\ &> 2\cdot\frac{\pi^2}{10n^2}\,\frac{1}{10}+\frac{\gamma}{100n^2}=\frac{2\pi^2+\gamma}{100n^2}. \end{aligned} \tag{122}$$

Moreover, recalling that $n\geq 5$ satisfies the condition $\frac{25\gamma}{120n^2}<\frac{1}{6}$, we can use the bound provided in Lemma 15 for the spectrum of the matrix $\left(K_n^{[2]}+\frac{\gamma}{n^2}M_n^{[2]}\right)$. Hence, by the minimax principle, by Lemmas 1 and 15, and by (90),

$$\begin{aligned}\lambda_{\min}(\mathrm{Re}A_{n,n}^{[2,2]}) &\geq \lambda_{\min}\left(M_n^{[2]}\otimes K_n^{[2]}\right)+\lambda_{\min}\left(\left(K_n^{[2]}+\frac{\gamma}{n^2}M_n^{[2]}\right)\otimes M_n^{[2]}\right)\\ &= \lambda_{\min}\left(M_n^{[2]}\right)\lambda_{\min}\left(K_n^{[2]}\right)+\lambda_{\min}\left(K_n^{[2]}+\frac{\gamma}{n^2}M_n^{[2]}\right)\lambda_{\min}\left(M_n^{[2]}\right)\\ &> \frac{1}{10}\,\frac{\pi^2}{10n^2}+\frac{\gamma}{n^2}\,\frac{1}{10}=\frac{\pi^2+10\gamma}{100n^2}. \end{aligned} \tag{123}$$

Furthermore, since $K_{n,n}^{[2,2]}=B_{n,n}^{[2,2]}+R_{n,n}^{[2,2]}$, we can decompose $\mathrm{Re}A_{n,n}^{[2,2]}$ as

$$\mathrm{Re}A_{n,n}^{[2,2]}=B_{n,n}^{[2,2]}+R_{n,n}^{[2,2]}+\frac{\gamma}{n^2}M_{n,n}^{[2,2]}.$$

We recall from (110) that $B_{n,n}^{[2,2]}=T_{n,n}(g_{2,2})$. The range of $g_{2,2}$ is $\left[0,\frac{3}{2}\right]$, and so by Theorem 5 we obtain $\sigma(B_{n,n}^{[2,2]})\subset\left(0,\frac{3}{2}\right)$. Concerning the symmetric matrix $R_{n,n}^{[2,2]}$, we find by Gershgorin's first theorem that $\sigma(R_{n,n}^{[2,2]})\subset\left[-\frac{269}{360},\frac{13}{24}\right]$. Using Lemmas 1 and 15, we also find that $\sigma(M_{n,n}^{[2,2]})\subset(\frac{1}{100},1)$. Then, we apply again the minimax principle to obtain an upper bound for $\lambda_{\max}(\mathrm{Re}A_{n,n}^{[2,2]})$:

$$\lambda_{\max}(\mathrm{Re}A_{n,n}^{[2,2]})\leq\lambda_{\max}(B_{n,n}^{[2,2]})+\lambda_{\max}(R_{n,n}^{[2,2]})+\frac{\gamma}{n^2}\lambda_{\max}(M_{n,n}^{[2,2]})<\frac{3}{2}+\frac{13}{24}+\frac{\gamma}{n^2}=\frac{49}{24}+\frac{\gamma}{n^2}. \tag{124}$$

Combining (122)–(124) we obtain

$$\sigma(\mathrm{Re}A_{n,n}^{[2,2]})\subset\left(\max\left(\frac{\pi^2+10\gamma}{100n^2},\frac{2\pi^2+\gamma}{100n^2}\right),\frac{49}{24}+\frac{\gamma}{n^2}\right). \tag{125}$$

Now we localize the spectrum of $\mathrm{Im}A_{n,n}^{[2,2]}$. The matrices $\widehat{H}_{n,n}^{[2,2]}$, $\widetilde{H}_{n,n}^{[2,2]}$ are skew-symmetric and $\widehat{H}_{n,n}^{[2,2]}=M_n^{[2]}\otimes H_n^{[2]}$, $\widetilde{H}_{n,n}^{[2,2]}=H_n^{[2]}\otimes M_n^{[2]}$. By Lemmas 1 and 15, we have $\sigma(\widehat{H}_{n,n}^{[2,2]})=\sigma(\widetilde{H}_{n,n}^{[2,2]})\subset\{0\}\times(-\frac{11}{12},\frac{11}{12})$. Hence, by the minimax principle,

$$\begin{aligned}\lambda_{\min}(\mathrm{Im}A_{n,n}^{[2,2]}) &= \lambda_{\min}\left(\frac{\beta_1}{n}\,\frac{1}{\mathrm{i}}\widehat{H}_{n,n}^{[2,2]}+\frac{\beta_2}{n}\,\frac{1}{\mathrm{i}}\widetilde{H}_{n,n}^{[2,2]}\right)\\ &\geq \lambda_{\min}\left(\frac{\beta_1}{n}\,\frac{1}{\mathrm{i}}\widehat{H}_{n,n}^{[2,2]}\right)+\lambda_{\min}\left(\frac{\beta_2}{n}\,\frac{1}{\mathrm{i}}\widetilde{H}_{n,n}^{[2,2]}\right)\geq-\frac{|\beta_1|}{n}\,\frac{11}{12}-\frac{|\beta_2|}{n}\,\frac{11}{12}, \end{aligned}$$

and similarly it can be proved that

$$\lambda_{\max}(\mathrm{Im}A_{n,n}^{[2,2]}) \leq \frac{|\beta_1|}{n}\frac{11}{12} + \frac{|\beta_2|}{n}\frac{11}{12}.$$

Thus,

$$\sigma(\mathrm{Im}A_{n,n}^{[2,2]}) \subseteq \left[-\frac{11}{12}\frac{|\beta_1|+|\beta_2|}{n}, \frac{11}{12}\frac{|\beta_1|+|\beta_2|}{n}\right]. \tag{126}$$

Using (6) in combination with (125) and (126), we obtain (121). □

6 Conclusions

We have studied the spectral properties of stiffness matrices that arise when isogeometric analysis is employed for the numerical solution of classical second order elliptic problems. Motivated by the applicative interest in the fast solution of the related linear systems, we have provided a spectral characterization of the involved matrices. In particular, we have given an asymptotic analysis of

1. the eigenvalue of minimal modulus and the eigenvalue of maximal modulus,
2. the conditioning,
3. the localization of the spectrum,
4. the global behavior of the spectrum.

Concerning all these items, as in the case of Finite Differences and Finite Elements, the crucial information comes from a symbol that describes the spectrum. The current analysis is not yet complete since we have to take into account more involved geometries, variable coefficients operators, etc. We expect that the global symbol of the associated matrix sequence, describing the spectrum in such a general context, will be formed, in analogy with the Finite Difference and Finite Element context, by using the information from the main operator (the principal symbol in the Hörmander Theory [26]), the used approximation techniques, and the involved domain.

Of course, a second challenging step will be the use of such spectral information for designing optimal preconditioners in the Krylov methods, optimal multigrid methods, and efficient combinations of these techniques.

References

1. Aricó, A., Donatelli, M., Serra-Capizzano, S.: *V-cycle optimal convergence for certain (multilevel) structured linear systems*. SIAM J. Matrix Anal. Appl. **26**, 186–214 (2004)
2. Axelsson, O., Barker, V.: *Finite Element Solution of Boundary Value Problems, Theory and Computation*. Academic Press Inc., New York (1984)
3. Axelsson, O., Lindskog, G.: *On the rate of convergence of the preconditioned conjugate gradient method*. Numer. Math. **48**, 499–523 (1986)
4. Bazilevs, Y., Calo, V.M., Cottrell, J.A., Evans, J.A., Hughes, T.J.R., Lipton, S., Scott, M.A., Sederberg, T.W.: *Isogeometric analysis using T-splines*. Comput. Methods Appl. Mech. Engrg. **199**, 229–263 (2010)
5. Beckermann, B., Kuijlaars, A.B.J.: *Superlinear convergence of Conjugate Gradients*. SIAM J. Numer. Anal. **39**, 300–329 (2001)
6. Beckermann, B., Kuijlaars, A.B.J.: *On the sharpness of an asymptotic error estimate for Conjugate Gradients*. BIT Numer. Anal. **41**, 856–867 (2001)
7. Beckermann, B., Kuijlaars, A.B.J.: *Superlinear CG convergence for special right-hand sides*. Electr. Trans. Numer. Anal. **14**, 1–19 (2002)

8. Beckermann, B., Serra-Capizzano, S.: *On the asymptotic spectrum of Finite Elements matrices*. SIAM J. Numer. Anal. **45**, 746–769 (2007)
9. Beirão da Veiga, L., Buffa, A., Cho, D., Sangalli, G.: *Analysis-suitable T-splines are Dual-Compatible*. Comput. Methods Appl. Mech. Engrg. **249–252**, 42–51 (2012)
10. Bertaccini, D., Golub, G., Serra-Capizzano, S., Tablino Possio, C.: *Preconditioned HSS method for the solution of non-Hermitian positive definite linear systems and applications to the discrete convection-diffusion equation*. Numer. Math. **99**, 441–484 (2005)
11. Bhatia, R.: *Matrix analysis*. Springer-Verlag, New York (1997)
12. Bini, D., Capovani, M.: *Spectral and computational properties of band symmetric Toeplitz matrices*. Linear Algebra Appl. **52–53**, 99–126 (1983)
13. Bini, D., Capovani, M., Menchi, O.: *Metodi numerici per l'algebra lineare*. Zanichelli (1988)
14. de Boor, C.: *A practical guide to splines*. Springer-Verlag, New York (2001)
15. Böttcher, A., Silbermann, B.: *Introduction to Large Truncated Toeplitz Matrices*. Springer, New York (1999)
16. Böttcher, A., Widom, H.: *From Toeplitz eigenvalues through Green's kernels to higher-order Wirtinger-Sobolev inequalities*. Operator Theory: Advances and Applications **171**, 73–87 (2007)
17. Brezis, H.: *Functional analysis, Sobolev spaces and partial differential equations*. Springer (2011)
18. Buffa, A., Harbrecht, H., Kunoth, A., Sangalli, G.: *BPX-preconditioning for isogeometric analysis*. Comput. Methods Appl. Mech. Engrg. **265**, 63–70 (2013)
19. Chui, C.K.: *An introduction to wavelets*. Academic Press (1992)
20. Cottrell, J.A., Hughes, T.J.R., Bazilevs, Y.: *Isogeometric analysis: toward integration of CAD and FEA*. John Wiley & Sons (2009)
21. Dörfel, M., Jüttler, B., Simeon, B.: *Adaptive isogeometric analysis by local h-refinement with T-splines*. Comput. Methods Appl. Mech. Engrg. **199**, 264–275 (2010)
22. Fiorentino, G., Serra-Capizzano, S.: *Multigrid methods for symmetric positive definite block Toeplitz matrices with nonnegative generating functions*. SIAM J. Sci. Comput. **17**, 1068–1081 (1996)
23. Gahalaut, K.P.S., Kraus, J.K., Tomar, S.K.: *Multigrid methods for isogeometric discretization*. Comput. Methods Appl. Mech. Engrg. **253**, 413–425 (2013)
24. Golinskii, L., Serra-Capizzano, S.: *The asymptotic properties of the spectrum of nonsymmetrically perturbed Jacobi matrix sequences*. J. Approx. Theory **144**, 84–102 (2007)
25. Grenander, U., Szegö, G.: *Toeplitz forms and their applications*, second edition. Chelsea, New York (1984)
26. Hörmander, L.: *Pseudo-differential operators and non-elliptic boundary problems*. Annals of Math. **2**, 129–209 (1966)
27. 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**, 4135–4195 (2005)
28. Johnson, C.: *Numerical Solutions of Partial Differential Equations by the Finite Elements Methods*. Cambridge Univ. Press, Cambridge (1988)
29. Manni, C., Pelosi, F., Sampoli, M.L.: *Generalized B-splines as a tool in isogeometric analysis*. Comput. Methods Appl. Mech. Engrg. **200**, 867–881 (2011)
30. Parter, S.V.: *On the extreme eigenvalues of truncated Toeplitz matrices*. Bull. Amer. Math. Soc. **67**, 191–197 (1961)
31. Parter, S.V.: *On the extreme eigenvalues of Toeplitz matrices*. Trans. Amer. Math. Soc. **100**, 263–276 (1961)
32. Parter, S.V.: *On the eigenvalues of certain generalizations of Toeplitz matrices*. Arch. Rat. Math. Mech. **3**, 244–257 (1962)
33. Quarteroni, A.: *Numerical models for differential problems*. Springer-Verlag Italia (2009)
34. Russo, A., Tablino Possio, C.: *Preconditioned Hermitian and skew-Hermitian splitting method for finite element approximations of convection-diffusion equations*. SIAM J. Matrix Anal. Appl. **31**, 997–1018 (2009)
35. Saad, Y.: *Iterative Methods for Sparse Linear Systems*. PWS Publishing, Boston, MA (1996)
36. Schumaker, L.L.: *Spline functions: basic theory*, third edition. Cambridge Mathematical Library (2007)
37. Serra-Capizzano, S.: *Preconditioning strategies for asymptotically ill-conditioned block Toeplitz systems*. BIT Numer. Anal. **34**, 579–594 (1994)
38. Serra-Capizzano, S.: *Convergence analysis of Two-Grid methods for elliptic Toeplitz and PDEs matrix-sequences*. Numer. Math. **92**, 433–465 (2002)
39. Serra-Capizzano, S.: *Generalized Locally Toeplitz sequences: spectral analysis and applications to discretized Partial Differential Equations*. Linear Algebra Appl. **366**, 371–402 (2003)

40. Serra-Capizzano, S.: *GLT sequences as a Generalized Fourier Analysis and applications*. Linear Algebra Appl. **419**, 180–233 (2006)
41. Serra-Capizzano, S., Tablino Possio, C.: *Spectral and structural analysis of high precision Finite Difference matrices for Elliptic Operators*. Linear Algebra Appl. **293**, 85–131 (1999)
42. Speleers, H., Manni, C., Pelosi, F., Sampoli, M.L.: *Isogeometric analysis with Powell-Sabin splines for advection-diffusion-reaction problems*. Comput. Methods Appl. Mech. Engrg. **221–222**, 132–148 (2012)
43. Tilli, P.: *A note on the spectral distribution of Toeplitz matrices*. Linear Multilinear Algebra **45**, 147–159 (1998)
44. Tilli, P.: *Locally Toeplitz sequences: spectral properties and applications*. Linear Algebra Appl. **278**, 91–120 (1998)
45. van der Sluis, A., van der Vorst, H.A.: *The rate of convergence of conjugate gradients*. Numer. Math. **48**, 543–560 (1986)