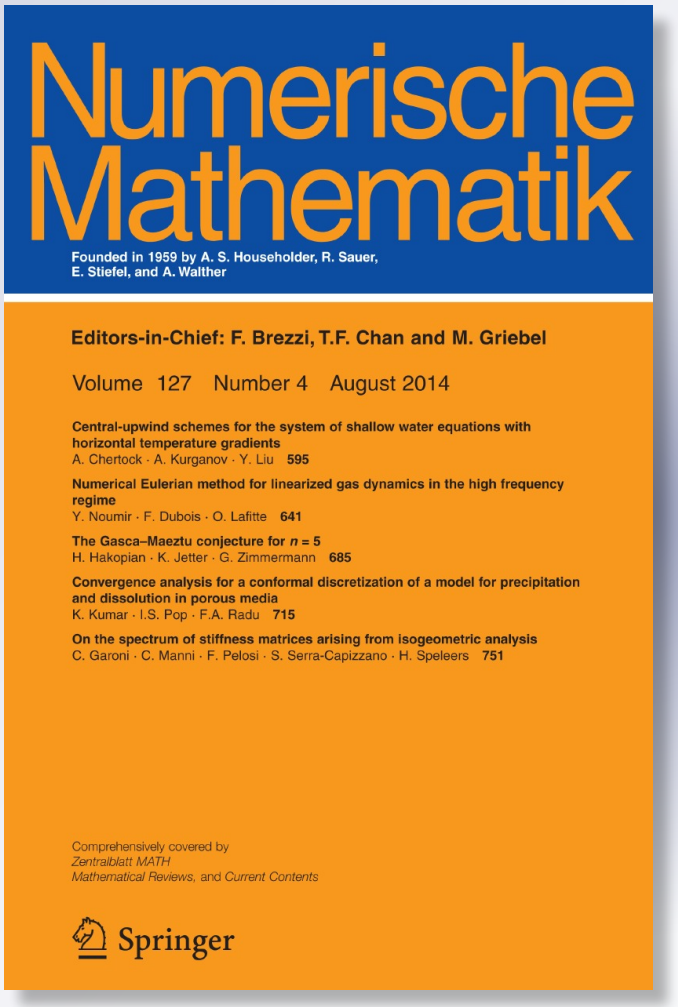
On the spectrum of stiffness matrices arising from isogeometric analysis

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

Numerische Mathematik

ISSN 0029-599X
Volume 127
Number 4

Numer. Math. (2014) 127:751-799
DOI 10.1007/s00211-013-0600-2



Springer

Numer. Math. (2014) 127:751–799
DOI 10.1007/s00211-013-0600-2

Numerische Mathematik

On the spectrum of stiffness matrices arising from isogeometric analysis

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

Received: 2 January 2013 / Revised: 20 September 2013 / Published online: 12 December 2013

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.

Mathematics Subject Classification (2010) 15A12 · 15A18 · 65N30 · 41A15 · 15B05 · 15A69 · 65L10

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

S. Serra-Capizzano
e-mail: stefano.serrac@uninsubria.it

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

F. Pelosi
e-mail: pelosi@mat.uniroma2.it

H. Speleers (✉)
Department of Computer Science, University of Leuven,
Celestijnenlaan 200A, 3001 Heverlee (Leuven), Belgium
e-mail: hendrik.speleers@cs.kuleuven.be

1 Introduction

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

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

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. [10]), 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 [2,26,36])—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 [2] 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 [4]. 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. [2,36]). For instance, one has a simple upper bound for the CG error in energy norm in terms of the spectral condition number. In reality, the upper bound based on the condition number may be not very accurate, especially when superlinear convergence of CG is observed. This superlinear convergence has been detected numerically in the context of discretized elliptic problems in dimension $d \geq 2$, in particular for small stepsizes h. In this setting, the CG convergence is known to be governed by the distribution of the spectrum and has been quantified only recently (see [4,5] and references therein). Similar results are also available for other Krylov methods, when the matrices are not Hermitian positive definite (see [26]).

A discretization of our differential problem for some sequence of stepsizes 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 [10] 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 equispaced 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 [23,31,35] for the one-dimensional setting, and in [29,30] 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 [6]).

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 [7,29–31,35] and Finite Element [6,25] approximations: in fact 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 Sects. 1.1 and 1.2, we present the considered differential problem and the main basics on IgA methods. In Sect. 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 Sect. 4 is devoted to the analysis of matrices arising from the IgA discretization based on B-splines in the 1D case, and Sect. 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 Sect. 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:

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

where $\Omega \subset \mathbb{R}^d$ is a domain with Lipschitz boundary, $\mathrm{f} \in L_2(\Omega)$, $\boldsymbol{\beta} \in \mathbb{R}^d$ 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

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

In the standard Galerkin approach, 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 $\{\varphi_1, \ldots, \varphi_N\}$ for $\mathscr{W}$, then each $v \in \mathscr{W}$ can be written as $v = \sum_{j=1}^{N} v_j \varphi_j$. So, 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). Note that A is symmetric only when $\boldsymbol{\beta} = \mathbf{0}$.

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 IgA based on B-splines

IgA is a paradigm for the analysis of problems governed by partial differential equations [14,20]. Its 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, see [14]. This results in some main 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 [14] 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 like 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 [14,20] and references therein. These appealing features are maintained by the above mentioned generalizations of B-splines, see e.g. [3,21,33].

Finally, the well known properties of the B-spline basis—convex partition of unity, minimal support, local linear independence, optimality of the basis, etc., see e.g. [9]—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 from 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. This topic has not yet been addressed in the literature. Generalizations of this spectral analysis for problems defined on higher-dimensional boxes are straightforward but more involved from the notational point of view. On the other hand, the extension to more complex geometries requires further investigation. Some related results can be found in [12,15].

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. Let us start with introducing some notation and recalling some basic results that will be used throughout this paper. We refer to [8] for more details on basic linear algebra results.

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 $X \in \mathbb{C}^{m\times m}$, $\|X\|_1 \le \mathrm{rank}(X)\|X\| \le m\|X\|$. 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. Whenever $X, Y \in \mathbb{C}^{m\times m}$

are Hermitian, we write $X \geq Y$ if and only if $X - Y$ is nonnegative definite. 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

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

and $X = \mathrm{Re}\, X + \mathrm{i}\,\mathrm{Im}\, X$. The spectrum $\sigma(X)$ of X can be localized in terms of the extremal eigenvalues of $\mathrm{Re}\, X$ and $\mathrm{Im}\, X$, namely

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

Since many of the matrices appearing in Sect. 5 will be formed by a tensor-product of matrices defined in Sect. 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 $\mathbb{C}^{m_1m_2\times m_1m_2}$ given by:

$$X \otimes Y = \begin{bmatrix} x_{11}Y & x_{12}Y & \cdots & x_{1m_1}Y \\ x_{21}Y & x_{22}Y & \cdots & x_{2m_1}Y \\ \vdots & \vdots & & \vdots \\ x_{m_11}Y & x_{m_12}Y & \cdots & x_{m_1m_1}Y \end{bmatrix}.$$

The next lemma, see e.g. [8], collects some 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.

Now we introduce the fundamental definitions for developing our spectral analysis, see [17, 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

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

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.,

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

We also recall the following results, see [17, 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 $\|X_n\|,\ \|Y_n\| \leq C$ 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 $\|X_n\|,\ \|Y_n\|_1 \leq C$ for all n.*

Set $Z_n := X_n + Y_n$. Then $\{Z_n\} \overset{\lambda}{\sim} f$, and $\{Z_n\}$ 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 $m \geq 1$ in the following way:

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

[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.

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.

Given a sequence $\{x_m\}$, we write $x_m \searrow x$ ($x_m \nearrow x$) to denote that $\{x_m\}$ converges monotonically from above (below) to x.

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 [18], see also [34] 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,*

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

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

Another result due to Parter [22] concerns the asymptotics of the j-th smallest eigenvalue $\lambda_j(T_m(f))$, for j fixed and $m \to \infty$.

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 $(\theta_{\min}-\epsilon, \theta_{\min}+\epsilon)$ for some $\epsilon > 0$ and $f^{(2s)}(\theta_{\min}) > 0$ is the first non-vanishing derivative of f at $\theta_{\min}$. Finally, for every $m \geq 1$, let $\lambda_1(T_m(f)) \leq \cdots \leq \lambda_m(T_m(f))$ be the eigenvalues of $T_m(f)$ arranged in non-decreasing order. Then, for each fixed $j \geq 1$,*

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

i.e., $\lim_{m\to\infty} m^{2s}(\lambda_j(T_m(f)) - m_f) = c_{s,j}\,\frac{f^{(2s)}(\theta_{\min})}{(2s)!}$, *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

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

see [22, p. 191]. Thus, we find that $c_{1,j} = j^2\pi^2$ for all $j \geq 1$, see [16, Remarks 1,2,3] for details.

In view of Sect. 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:

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

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 [28] and again [34] 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,*

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

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

The last result relates tensor-products and Toeplitz matrices. 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 $\phi_{[p]}$ be the cardinal B-spline of degree p over the uniform knot sequence $\{0, 1, \ldots, p+1\}$, which is defined recursively as follows [9]:

$$\phi_{[0]}(t) := \begin{cases} 1, & \text{if } t \in [0, 1), \\ 0, & \text{elsewhere,} \end{cases} \tag{7}$$

and

$$\phi_{[p]}(t) := \frac{t}{p}\phi_{[p-1]}(t) + \frac{p+1-t}{p}\phi_{[p-1]}(t-1), \quad p \geq 1. \tag{8}$$

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 [9,13].

– *Positivity:*

$$\phi_{[p]}(t) > 0, \quad t \in (0, p+1).$$

– *Minimal support:*

$$\phi_{[p]}(t) = 0, \quad t \notin [0, p+1]. \tag{9}$$

– *Symmetry:*

$$\phi_{[p]}\left(\frac{p+1}{2} + t\right) = \phi_{[p]}\left(\frac{p+1}{2} - t\right). \tag{10}$$

– *Partition of unity:*

$$\sum_{k\in\mathbb{Z}} \phi_{[p]}(t-k) = 1,$$

which gives in combination with the local support and smoothness,

$$\sum_{k=1}^{p} \phi_{[p]}(k) = 1, \quad p \geq 1. \tag{11}$$

– *Recurrence relation for derivatives:*

$$\phi_{[p]}^{(r)}(t) = \phi_{[p-1]}^{(r-1)}(t) - \phi_{[p-1]}^{(r-1)}(t-1). \tag{12}$$

– *Convolution relation:*

$$\phi_{[p]}(t) = (\phi_{[p-1]} * \phi_{[0]})(t) := \int_{\mathbb{R}} \phi_{[p-1]}(t-s)\phi_{[0]}(s)\,\mathrm{d}s = \int_0^1 \phi_{[p-1]}(t-s)\,\mathrm{d}s. \tag{13}$$

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-spline.

Lemma 3 *Let $\phi_{[p]}$ be the cardinal B-spline as defined in (7)–(8), then*

$$\phi_{[p]}^{(r)}\left(\frac{p+1}{2}+t\right) = (-1)^r\,\phi_{[p]}^{(r)}\left(\frac{p+1}{2}-t\right).$$

Proof The result follows from repeated differentiations of the symmetry property (10). We can also prove it by induction on the order of derivatives using the recurrence relation (12), as outlined below. The base case ($r = 0$) is just the symmetry property (10). As inductive step we increase the order of derivative by one, i.e., $r \to r+1$. Using the recurrence relation for derivatives (12) and the induction hypothesis, we have

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

□

The following lemma provides an expression for inner products of derivatives of the cardinal B-spline and its integer translates. It generalizes the result given in [13, p. 89].

Lemma 4 *Let $\phi_{[p]}$ be the cardinal B-spline as defined in (7)–(8), then*

$$\begin{aligned}\int_{\mathbb{R}} \phi_{[p_1]}^{(r)}(t)\,\phi_{[p_2]}^{(s)}(t+k)\,\mathrm{d}t &= (-1)^r\,\phi_{[p_1+p_2+1]}^{(r+s)}(p_1+1+k)\\ &= (-1)^s\,\phi_{[p_1+p_2+1]}^{(r+s)}(p_2+1-k). \end{aligned} \tag{14}$$

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 (14).

We first address the case $r=s=0$, namely

$$\int_{\mathbb{R}} \phi_{[p_1]}(t)\phi_{[p_2]}(t+k)\,\mathrm{d}t = \phi_{[p_1+p_2+1]}(p_2+1-k). \tag{15}$$

Using the convolution relation of cardinal B-splines (13), we obtain

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

From [13, p. 85] we also know that for every continuous function f it holds

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

and hence

$$\phi_{[p_1+p_2+1]}(p_2+1-k) = \int_{\mathbb{R}} \phi_{[p_2]}(p_2+1-k-t)\phi_{[p_1]}(t)\,\mathrm{d}t. \tag{16}$$

Moreover, by symmetry of the cardinal B-splines, see (10), we have

$$\phi_{[p_2]}(p_2+1-k-t) = \phi_{[p_2]}(k+t). \tag{17}$$

Combining (16) and (17) results in (15).

We now prove the general case, i.e.,

$$\int_{\mathbb{R}} \phi^{(r)}_{[p_1]}(t)\,\phi^{(s)}_{[p_2]}(t+k)\,\mathrm{d}t = (-1)^r\,\phi^{(r+s)}_{[p_1+p_2+1]}(p_1+1+k), \tag{18}$$

by induction on the order of derivatives. We consider two inductive steps: in the first inductive step we increase the order of derivative of $\phi_{[p_1]}$ by one, i.e., $r \to r+1$, and in the second inductive step we increase the order of derivative of $\phi_{[p_2]}$ by one, i.e., $s \to s+1$.

1. $(r \to r+1)$. Using (12) and the induction hypothesis, we have

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

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

We will denote by $\dot{\phi}_{[p]}(t)$ and $\ddot{\phi}_{[p]}(t)$ the first and second derivative of $\phi_{[p]}(t)$ with respect to its argument t. The next lemma provides an interesting relation about second derivatives of cardinal B-splines.

Lemma 5 *Let $\phi_{[p]}$ be the cardinal B-spline as defined in (7)–(8), and let $\ddot{\phi}_{[p]}$ be its second derivative, then*

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

Proof By the relation (12), by the fact that $\phi_{[2p-1]}(-1) = \phi_{[2p-1]}(0) = 0$, and by taking into account that

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

we find that

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

The last equalities follow from the symmetry property (10) and the partition of unity property (11) 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 [13, Theorem 2.28].

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

$$\psi(t) = O(|t|^{-a}), \quad a > 1, \quad \text{as } |t| \to \infty, \tag{19}$$

and

$$\widehat{\psi}(\theta) = O(|\theta|^{-b}), \quad b > \frac{1}{2}, \quad \text{as } |\theta| \to \infty. \tag{20}$$

Then,

$$\sum_{k\in\mathbb{Z}} \left(\int_{\mathbb{R}} \psi(t-k)\overline{\psi(t)}\, \mathrm{d}t \right) \mathrm{e}^{\mathrm{i}(k\theta)} = \sum_{k\in\mathbb{Z}} \left| \widehat{\psi}(\theta + 2k\pi) \right|^2, \quad \forall \theta \in [-\pi, \pi]. \tag{21}$$

By using the convolution relation (13) one can easily obtain a simple expression for the Fourier transform of the cardinal B-spline $\phi_{[p]}$, see [13, p. 56]:

$$\widehat{\phi_{[p]}}(\theta) = \left(\frac{1 - \mathrm{e}^{-\mathrm{i}\,\theta}}{\mathrm{i}\,\theta}\right)^{p+1}, \tag{22}$$

so that

$$\left|\widehat{\phi_{[p]}}(\theta)\right|^2 = \left(\frac{2 - 2\cos\theta}{\theta^2}\right)^{p+1}. \tag{23}$$

From (9) and (22) it follows that the cardinal B-spline satisfies the conditions (19)–(20). 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 (21) by means of (23). This implies

$$\begin{aligned}\sum_{k\in\mathbb{Z}} \left|\widehat{\phi_{[p]}}(\theta + 2k\pi)\right|^2 \geq \left|\widehat{\phi_{[p]}}(\theta)\right|^2 &= \left(\frac{2 - 2\cos\theta}{\theta^2}\right)^{p+1} \\ &\geq \left(\frac{4}{\pi^2}\right)^{p+1}, \quad \theta \in [-\pi, \pi].\end{aligned} \tag{24}$$

On the other hand, to obtain an upper bound for (21), we make use of relations (15) and (21) and the partition of unity property (11). In this way, we obtain

$$\begin{aligned}\sum_{k\in\mathbb{Z}} \left|\widehat{\phi_{[p]}}(\theta + 2k\pi)\right|^2 &= \sum_{k\in\mathbb{Z}} \phi_{[2p+1]}(p + 1 - k)\,\mathrm{e}^{\mathrm{i}(k\theta)} \\ &\leq \sum_{k\in\mathbb{Z}} \phi_{[2p+1]}(p + 1 - k)\,|\mathrm{e}^{\mathrm{i}(k\theta)}| = 1.\end{aligned} \tag{25}$$

Note that for the cardinal B-spline of degree p the left-hand side in (21) 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 6 *Let $p \geq 1$, and let $f_p : [-\pi, \pi] \to \mathbb{R}$,*

$$f_p(\theta) := -\ddot{\phi}_{[2p+1]}(p + 1) - 2\sum_{k=1}^{p} \ddot{\phi}_{[2p+1]}(p + 1 - k)\cos(k\theta), \tag{26}$$

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

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

$$f_p(\theta) = (2 - 2\cos\theta)\sum_{k\in\mathbb{Z}} \left|\widehat{\phi_{[p-1]}}(\theta + 2k\pi)\right|^2, \tag{27}$$

and

$$(2-2\cos\theta)\left(\frac{4}{\pi^2}\right)^p \leq f_p(\theta) \leq \min\left(2-2\cos\theta, (2-2\cos\theta)^{p+1}\left(\frac{1}{\theta^{2p}}+\frac{1}{6\,\pi^{2p-2}}\right)\right). \quad (28)$$

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

Proof Using the recurrence relation for derivatives (12), for every $\theta \in [-\pi,\pi]$ we obtain that

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

and

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

This implies that

$$\sum_{k\in\mathbb{Z}}\left|\widehat{\dot{\phi}_{[p]}}(\theta+2k\pi)\right|^2 = (2-2\cos\theta)\sum_{k\in\mathbb{Z}}\left|\widehat{\phi_{[p-1]}}(\theta+2k\pi)\right|^2. \quad (29)$$

The equality (27) follows from relation (14), Theorem 6 and (29) 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 (27) and from the inequalities (24)–(25), we get

$$(2-2\cos\theta)\left(\frac{4}{\pi^2}\right)^p \leq f_p(\theta) \leq 2-2\cos\theta, \quad \forall\theta\in[-\pi,\pi]. \quad (30)$$

Furthermore, using (23) in the expression of f_p given by (27), we obtain that

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

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. This completes the proof of the first statement.

We now prove the second statement. The inequalities in (30) imply that $\min_{\theta\in[-\pi,\pi]} f_p(\theta) = f_p(0) = 0$, and that $\theta = 0$ is the only zero of f_p. From the upper bound in (28) we can also conclude that $M_{f_p} \to 0$ as $p \to \infty$, see [16, proof of Lemma 7] for details. □

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

$$h_p(\theta) := \phi_{[2p+1]}(p+1) + 2\sum_{k=1}^{p} \phi_{[2p+1]}(p+1-k)\cos(k\theta), \tag{31}$$

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* $m_{h_p} \ge \left(\frac{4}{\pi^2}\right)^{p+1}$.

Proof From relation (15) 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 (24)–(25) imply that

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

In addition, by the symmetry property (10) and the partition of unity property (11), 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 2 From the expressions of f_p and h_p given in Lemmas 6 and 7, 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 $p \geq 1$,

$$f_p(\theta) = (2 - 2\cos\theta) \left(\phi_{[2p-1]}(p) + 2 \sum_{k=1}^{p-1} \phi_{[2p-1]}(p-k) \cos(k\theta) \right). \quad (32)$$

The latter equality can be easily checked for $p = 1$ by a direct computation, with the usual assumption that a sum is zero when the upper index is less than the lower one. Note that (32) 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 $\Omega = (0, 1)$, namely

$$\begin{cases} -u'' + \beta u' + \gamma u = \mathrm{f}, & 0 < x < 1, \\ u(0) = 0, \quad u(1) = 0, \end{cases} \quad (33)$$

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 (33) 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 defined over the knot sequence

$$t_1 = \cdots = t_{p+1} = 0 < t_{p+2} < \cdots < t_{p+n} < 1 = t_{p+n+1} = \cdots = t_{2p+n+1}, \quad (34)$$

where

$$t_{p+i+1} := \frac{i}{n}, \quad \forall i = 0, \ldots, n, \quad (35)$$

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.,

$$\mathscr{W}_n^{[p]} := \{ s \in \mathscr{V}_n^{[p]} : s(0) = s(1) = 0 \} \subset H_0^1([0, 1]). \quad (36)$$

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 [9]). 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, by recalling from [9]

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

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

$$\mathscr{W} = \langle N_{i,[p]},\ i = 2, \ldots, n + p - 1 \rangle. \tag{37}$$

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 [24].

Using the basis (37), 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:

$$A_n^{[p]} := A = \left[a(N_{j+1,[p]}, N_{i+1,[p]})\right]_{i,j=1}^{n+p-2}, \tag{38}$$

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 $N_{i,[p]}(x),\ i = p+1, \ldots, n$, defined on the knot sequence (34)–(35), are cardinal B-splines, see Sect. 3. We have

$$N_{i,[p]}(x) = \phi_{[p]}(nx - i + p + 1), \quad i = p + 1, \ldots, n. \tag{39}$$

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

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

We now focus on the central part of the stiffness matrix which is only determined by the cardinal B-splines in (39). For each $k = 0, 1, \ldots, p$ and $i = 2p, \ldots, n-p-1$, the non-zero element in (38) at row i and column $i \pm k$ can be expressed by

$$\begin{aligned}
\left(A_n^{[p]}\right)_{i,i\pm k} &= a(N_{i+1\pm k,[p]}(x), N_{i+1,[p]}(x)) \\
&= a(\phi_{[p]}(nx - i + p \mp k), \phi_{[p]}(nx - i + p)) \\
&= n^2 \int_0^1 \dot{\phi}_{[p]}(nx - i + p \mp k)\, \dot{\phi}_{[p]}(nx - i + p)\, \mathrm{d}x \\
&\quad + n\beta \int_0^1 \dot{\phi}_{[p]}(nx - i + p \mp k)\, \phi_{[p]}(nx - i + p)\, \mathrm{d}x \\
&\quad + \gamma \int_0^1 \phi_{[p]}(nx - i + p \mp k)\, \phi_{[p]}(nx - i + p)\, \mathrm{d}x \\
&= n \int_{\mathbb{R}} \dot{\phi}_{[p]}(t \mp k)\, \dot{\phi}_{[p]}(t)\, \mathrm{d}t + \beta \int_{\mathbb{R}} \dot{\phi}_{[p]}(t \mp k)\, \phi_{[p]}(t)\, \mathrm{d}t \\
&\quad + \frac{\gamma}{n} \int_{\mathbb{R}} \phi_{[p]}(t \mp k)\, \phi_{[p]}(t)\, \mathrm{d}t. \qquad (40)
\end{aligned}$$

Let us consider the following split of the matrix,

$$A_n^{[p]} = nK_n^{[p]} + \beta H_n^{[p]} + \frac{\gamma}{n} M_n^{[p]}, \qquad (41)$$

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

$$nK_n^{[p]} := \left[\int_0^1 N'_{j+1,[p]}(x) N'_{i+1,[p]}(x)\, \mathrm{d}x \right]_{i,j=1}^{n+p-2}, \qquad (42)$$

$$H_n^{[p]} := \left[\int_0^1 N'_{j+1,[p]}(x) N_{i+1,[p]}(x)\, \mathrm{d}x \right]_{i,j=1}^{n+p-2}, \qquad (43)$$

$$\frac{1}{n} M_n^{[p]} := \left[\int_0^1 N_{j+1,[p]}(x) N_{i+1,[p]}(x)\, \mathrm{d}x \right]_{i,j=1}^{n+p-2}. \qquad (44)$$

In view of (40), the parts of these matrices determined by the cardinal B-splines in (39) are

$$\left(K_n^{[p]}\right)_{i,i\pm k} = \int_{\mathbb{R}} \dot{\phi}_{[p]}(t \mp k)\dot{\phi}_{[p]}(t)\,\mathrm{d}t, \tag{45}$$

$$\left(H_n^{[p]}\right)_{i,i\pm k} = \int_{\mathbb{R}} \dot{\phi}_{[p]}(t \mp k)\phi_{[p]}(t)\,\mathrm{d}t, \tag{46}$$

$$\left(M_n^{[p]}\right)_{i,i\pm k} = \int_{\mathbb{R}} \phi_{[p]}(t \mp k)\phi_{[p]}(t)\,\mathrm{d}t, \tag{47}$$

for $k = 0, 1, \ldots, p$ and $i = 2p, \ldots, n-p-1$. 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 (45)–(47), i.e., for the row indices $i = 2p, \ldots, n-p-1$. Other rules have to be considered for the remaining $2p-1$ initial/final rows. Lemma 4 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 $k = 0, 1, \ldots, p$ and $i = 2p, \ldots, n-p-1$.

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

$$\left[0 \cdots 0\ -\ddot{\phi}_{[2p+1]}(1) \cdots -\ddot{\phi}_{[2p+1]}(p)\ -\ddot{\phi}_{[2p+1]}(p+1)\ -\ddot{\phi}_{[2p+1]}(p) \cdots -\ddot{\phi}_{[2p+1]}(1)\, 0 \cdots 0\right], \tag{48}$$

and in particular, by using (12) and (10), the diagonal elements can be expressed as

$$\left(K_n^{[p]}\right)_{i,i} = -\ddot{\phi}_{[2p+1]}(p+1) = -2\dot{\phi}_{[2p]}(p+1) = 2\dot{\phi}_{[2p]}(p).$$

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

$$\left[0 \cdots 0\ -\dot{\phi}_{[2p+1]}(1) \cdots -\dot{\phi}_{[2p+1]}(p)\ 0\ \dot{\phi}_{[2p+1]}(p) \cdots \dot{\phi}_{[2p+1]}(1)\ 0 \cdots 0\right]. \tag{49}$$

Since $\phi_{[2p+1]}$ is symmetric with respect to $p+1$, see (10), we have $\dot{\phi}_{[2p+1]}(p+1) = 0$, and hence

$$\left(H_n^{[p]}\right)_{i,i} = \dot{\phi}_{[2p+1]}(p+1) = 0.$$

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

$$\left[0 \cdots 0\ \phi_{[2p+1]}(1) \cdots \phi_{[2p+1]}(p)\ \phi_{[2p+1]}(p+1)\ \phi_{[2p+1]}(p) \cdots \phi_{[2p+1]}(1)\ 0 \cdots 0\right]. \tag{50}$$

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*

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

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

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 [27]. For every $p \geq 1$, $n \geq 2$, and $\mathbf{x} = (x_1, \ldots, x_{n+p-2}) \in \mathbb{R}^{n+p-2}$,

$$C_p \frac{\|\mathbf{x}\|^2}{n} \leq \left\| \sum_{i=1}^{n+p-2} x_i N_{i+1,[p]} \right\|_{L_2([0,1])}^2 \leq \bar{C}_p \frac{\|\mathbf{x}\|^2}{n}, \tag{51}$$

where the constants $C_p, \bar{C}_p > 0$ do not depend on n and $\mathbf{x}$. The inequalities in (51) are a special instance for the L_2-norm of the results stated in [27, 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 [27, eq. (6.3)].

We also recall the Poincaré inequality in the one-dimensional setting:

$$\|v\|_{L_2([0,1])} \leq \frac{1}{\pi}\|v'\|_{L_2([0,1])}, \quad \forall\, v \in H_0^1([0,1]). \tag{52}$$

In [11] we find that $\frac{1}{\pi}$ is the best constant for which (52) is satisfied.

The inequalities (51)–(52) play an important role in the proof of Theorem 8.

Theorem 8 *Let $C_p > 0$ be the constant in (51), then for all $p \geq 1$ and $n \geq 2$ the following properties hold.*

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

Proof Fix $p \geq 1$, $n \geq 2$. By using the definition of $M_n^{[p]}$, see (44), we have for all $\mathbf{y} \in \mathbb{R}^{n+p-2}$,

$$\begin{aligned}
\mathbf{y}^T \left(\frac{1}{n} M_n^{[p]}\right) \mathbf{y} &= \sum_{i,j=1}^{n+p-2} \left(\frac{1}{n} M_n^{[p]}\right)_{i,j} y_i y_j \\
&= \sum_{i,j=1}^{n+p-2} \int_0^1 y_i y_j N_{j+1,[p]}(x) N_{i+1,[p]}(x) \mathrm{d}x \\
&= \int_0^1 \sum_{i=1}^{n+p-2} y_i N_{i+1,[p]}(x) \sum_{j=1}^{n+p-2} y_j N_{j+1,[p]}(x) \mathrm{d}x \\
&= \int_0^1 \left(\sum_{i=1}^{n+p-2} y_i N_{i+1,[p]}(x) \right)^2 \mathrm{d}x \\
&= \left\| \sum_{i=1}^{n+p-2} y_i N_{i+1,[p]} \right\|^2_{L_2([0,1])} \geq C_p \frac{\|\mathbf{y}\|^2}{n}. \qquad (53)
\end{aligned}$$

The last inequality holds because of (51). Hence, we get $\mathbf{y}^T M_n^{[p]} \mathbf{y} \geq C_p \|\mathbf{y}\|^2$, and from the minimax principle [8] it follows that

$$\lambda_{\min}(M_n^{[p]}) = \min_{\mathbf{y} \neq \mathbf{0}} \frac{\mathbf{y}^T M_n^{[p]} \mathbf{y}}{\|\mathbf{y}\|^2} \geq C_p. \qquad (54)$$

This proves the first statement. To prove the second statement, we use the definition of $K_n^{[p]}$, see (42), and obtain for all $\mathbf{y} \in \mathbb{R}^{n+p-2}$,

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

where $v_{\mathbf{y}} := \sum_{i=1}^{n+p-2} y_i N_{i+1,[p]} \in \mathscr{W}_n^{[p]}$, see (36)–(37). Since $\mathscr{W}_n^{[p]} \subset H_0^1([0,1])$, we may apply the Poincaré inequality (52). From (52) and (53) it follows that

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

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

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

This proves that $K_n^{[p]} \geq \frac{\pi^2}{n^2} M_n^{[p]}$. Moreover, the minimax principle and (54) 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 3 For every $p \geq 1$, $n \geq 2$ and $j = 1, \ldots, n+p-2$, let $\lambda_j(K_n^{[p]})$ be the j-th smallest eigenvalue of $K_n^{[p]}$, i.e., $\lambda_1(K_n^{[p]}) \leq \cdots \leq \lambda_{n+p-2}(K_n^{[p]})$. Then, we conjecture that for every $p \geq 1$ and for each fixed $j \geq 1$,

$$\lim_{n \to \infty} \left(n^2 \lambda_j(K_n^{[p]}) \right) = j^2 \pi^2. \tag{55}$$

This conjecture can be motivated as follows. The matrix $K_n^{[p]}$ is associated with the (IgA) discretization of the boundary value problem (33) with $\beta = \gamma = 0$. The numbers $j^2\pi^2$, $j = 1, 2, \ldots$, are precisely the eigenvalues of this boundary value problem, see Remark 1. We have verified this conjecture numerically for $p = 2, 3, 4$, for $j = 1, 2, 3$ and for increasing values of n, see [16, p. 23].

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,*

$$\left| \lambda_{\min}(A_n^{[p]}) \right| \geq \lambda_{\min}(\mathrm{Re}\, A_n^{[p]}) \geq \frac{C_p(\pi^2 + \gamma)}{n}, \tag{56}$$

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

Proof By the expression (41) 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

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

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

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

From (6) we know that $|\lambda_{\min}(A_n^{[p]})| \geq \lambda_{\min}(\mathrm{Re}\, A_n^{[p]})$, implying (56). □

The lower bound (56) 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 [8, Proposition III.5.1]. The second result (Lemma 8) 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*

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

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

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

Lemma 8 *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 [9,27]),

$$N'_{i,[p]}(x) = p\left(\frac{N_{i,[p-1]}(x)}{t_{i+p} - t_i} - \frac{N_{i+1,[p-1]}(x)}{t_{i+p+1} - t_{i+1}}\right), \tag{57}$$

and

$$\int_{\mathbb{R}} N_{i,[p]}(x)\, \mathrm{d}x = \frac{t_{i+p+1} - t_i}{p+1}. \tag{58}$$

The sequence of knots (34)–(35) implies that the length of the support of any $N_{i,[p]}$ can be bounded from above by $\frac{p+1}{n}$. Recalling (44), by the positivity property and the partition of unity property of B-splines, we obtain

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

Due to the skew-symmetry of $H_n^{[p]}$, see (43), the infinity norm of $H_n^{[p]}$ is equal to the infinity norm of its transpose. By (57) and the positivity property of B-splines, we obtain

$$\begin{aligned}
\|H_n^{[p]}\|_\infty &= \max_{i=1,\dots,n+p-2} \sum_{j=1}^{n+p-2} \left| \int_0^1 N_{j+1,[p]}(x) N'_{i+1,[p]}(x)\,\mathrm{d}x \right| \\
&= \max_{i=1,\dots,n+p-2} p \sum_{j=1}^{n+p-2} \left| \int_0^1 N_{j+1,[p]}(x) \left(\frac{N_{i+1,[p-1]}(x)}{t_{i+p+1} - t_{i+1}} - \frac{N_{i+2,[p-1]}(x)}{t_{i+p+2} - t_{i+2}} \right) \mathrm{d}x \right| \\
&\le \max_{i=1,\dots,n+p-2} p \sum_{j=1}^{n+p-2} \int_0^1 N_{j+1,[p]}(x) \left(\frac{N_{i+1,[p-1]}(x)}{t_{i+p+1} - t_{i+1}} + \frac{N_{i+2,[p-1]}(x)}{t_{i+p+2} - t_{i+2}} \right) \mathrm{d}x.
\end{aligned} \tag{59}$$

By using the partition of unity property and (58), we have

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

and a similar bound holds for the remaining term in (59). It follows that $\|H_n^{[p]}\|_\infty \le 2$.

Recalling (42), we obtain

$$\begin{aligned}\|nK_n^{[p]}\|_\infty &= \max_{i=1,\dots,n+p-2} \sum_{j=1}^{n+p-2} \left| \int_0^1 N'_{j+1,[p]}(x) N'_{i+1,[p]}(x)\,\mathrm{d}x \right| \\ &= \max_{i=1,\dots,n+p-2} p^2 \sum_{j=1}^{n+p-2} \left| \int_0^1 \left(\frac{N_{j+1,[p-1]}(x)}{t_{j+p+1}-t_{j+1}} - \frac{N_{j+2,[p-1]}(x)}{t_{j+p+2}-t_{j+2}} \right) \right. \\ &\quad \left. \left(\frac{N_{i+1,[p-1]}(x)}{t_{i+p+1}-t_{i+1}} - \frac{N_{i+2,[p-1]}(x)}{t_{i+p+2}-t_{i+2}} \right) \mathrm{d}x \right|. \end{aligned} \tag{60}$$

In addition, we have

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

and in a similar way we can also bound the remaining terms in (60). This results in

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

□

Remark 4 A consequence of Lemma 8 is that we can take $\bar{C}_p = 1$ in (51), independently of p. Indeed, Lemma 8 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 (see (53)), we have

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

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

$$\kappa_2(A_n^{[p]}) \le \alpha_p n^2, \quad \forall n \ge 2. \tag{61}$$

Proof Fix $p \geq 1$ and $n \geq 2$. By Theorem 7 it follows that $K_n^{[p]}$, $H_n^{[p]}$ and $M_n^{[p]}$ are normal matrices, and by applying Lemma 8 we obtain for $\|A_n^{[p]}\|$ the following bound:

$$\begin{aligned}\|A_n^{[p]}\| &= \left\| nK_n^{[p]} + \beta H_n^{[p]} + \frac{\gamma}{n} M_n^{[p]} \right\| \leq \|nK_n^{[p]}\| + |\beta| \|H_n^{[p]}\| + \gamma \left\| \frac{1}{n} M_n^{[p]} \right\| \\ &\leq \|nK_n^{[p]}\|_\infty + |\beta| \|H_n^{[p]}\|_\infty + \gamma \left\| \frac{1}{n} M_n^{[p]} \right\|_\infty \leq 4pn + 2|\beta| + \frac{\gamma}{n}. \end{aligned} \tag{62}$$

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

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

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

$$\|(A_n^{[p]})^{-1}\| = \frac{1}{s_{n+p-2}(A_n^{[p]})} \leq \frac{n}{C_p(\pi^2 + \gamma)}. \tag{63}$$

Combining (62) with (63), we get $\kappa_2(A_n^{[p]}) \leq \frac{4pn^2 + 2n|\beta| + \gamma}{C_p(\pi^2 + \gamma)}$, which implies (61) 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

$$\frac{1}{n} A_n^{[p]} = K_n^{[p]} + \frac{\beta}{n} H_n^{[p]} + \frac{\gamma}{n^2} M_n^{[p]}, \tag{64}$$

formed by the scaled stiffness matrices. Recall from (38) 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

$$K_n^{[p]} = B_n^{[p]} + R_n^{[p]}, \tag{65}$$

where $B_n^{[p]}$ is the symmetric $(2p+1)$-band matrix whose generic central row is given by (48), 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

$$\mathrm{rank}(R_n^{[p]}) \leq 2(2p - 1). \tag{66}$$

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

$$M_n^{[p]} = C_n^{[p]} + S_n^{[p]}, \tag{67}$$

where $C_n^{[p]}$ is the symmetric $(2p+1)$-band matrix whose generic central row is given by (50), while $S_n^{[p]} := M_n^{[p]} - C_n^{[p]}$ is a low-rank correction term analogous to $R_n^{[p]}$ and

$$\operatorname{rank}(S_n^{[p]}) \leq 2(2p-1). \tag{68}$$

Now we analyze the spectral properties of $B_n^{[p]}$ and $C_n^{[p]}$. These properties will be used in the proof of Theorem 12, which yields the spectral distribution of the sequence $\{\frac{1}{n}A_n^{[p]}\}$.

Lemma 9 *Let f_p and M_{f_p} be defined as in Lemma 6. 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 $\lambda_1(B_n^{[p]}) \leq \cdots \leq \lambda_{n+p-2}(B_n^{[p]})$ are the eigenvalues of $B_n^{[p]}$ in non-decreasing 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 6.

We now prove the last statement. From Lemma 6 we know that $\theta = 0$ is the unique zero of f_p over $[-\pi, \pi]$. Furthermore, from (30) it is easy to derive that $f_p'(0) = 0$ and, by using Lemma 5,

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

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 Remark 1. □

In Lemma 9 we have seen that (55) holds with $\lambda_j(B_n^{[p]})$ instead of $\lambda_j(K_n^{[p]})$. Since $B_n^{[p]}$ is equal to $K_n^{[p]}$ up to a low-rank correction term, see (65)–(66), this may further support the conjecture formulated in Remark 3.

Lemma 10 *Let h_p and m_{h_p} be defined as in Lemma 7. 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 7 conclude the proof. □

Theorem 12 *The sequence of matrices $\{\frac{1}{n}A_n^{[p]}\}$ is distributed like the function f_p defined in (26) 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 (64)–(65), we have

$$\frac{1}{n}A_n^{[p]} = B_n^{[p]} + R_n^{[p]} + \frac{\beta}{n}H_n^{[p]} + \frac{\gamma}{n^2}M_n^{[p]}. \tag{69}$$

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 (69). We have seen in Lemma 9 that $\{B_n^{[p]}\} \overset{\lambda}{\sim} f_p$. Noting that $B_n^{[p]}$ is symmetric, by Lemma 9 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 (66)) and since $K_n^{[p]}$, $H_n^{[p]}$ and $M_n^{[p]}$ are normal matrices, we get

$$\begin{aligned}
&\left\| R_n^{[p]} + \frac{\beta}{n}H_n^{[p]} + \frac{\gamma}{n^2}M_n^{[p]} \right\|_1 \\
&\quad \leq \|R_n^{[p]}\|_1 + \frac{|\beta|}{n}\|H_n^{[p]}\|_1 + \frac{\gamma}{n^2}\|M_n^{[p]}\|_1 \\
&\quad \leq \operatorname{rank}(R_n^{[p]})\|R_n^{[p]}\| + |\beta|\frac{(n+p-2)}{n}\|H_n^{[p]}\| + \gamma\frac{(n+p-2)}{n^2}\|M_n^{[p]}\|
\end{aligned}$$

$$
\begin{aligned}
&\le 2(2p-1)\|K_n^{[p]} - B_n^{[p]}\| + |\beta|\frac{(n+p-2)}{n}\|H_n^{[p]}\| + \gamma\frac{(n+p-2)}{n^2}\|M_n^{[p]}\| \\
&\le 2(2p-1)\|B_n^{[p]}\| + 2(2p-1)\|K_n^{[p]}\| + |\beta|\frac{(n+p-2)}{n}\|H_n^{[p]}\| \\
&\quad + \gamma\frac{(n+p-2)}{n^2}\|M_n^{[p]}\| \\
&\le 2(2p-1)M_{f_p} + 2(2p-1)\|K_n^{[p]}\|_\infty + |\beta|\frac{(n+p-2)}{n}\|H_n^{[p]}\|_\infty \\
&\quad + \gamma\frac{(n+p-2)}{n^2}\|M_n^{[p]}\|_\infty.
\end{aligned}
$$

From Lemma 8 it follows that the right-hand side of the last inequality can be bounded from above by a constant independent of n, $\forall n \ge 3p+1$. Hence, all the hypotheses of Theorem 2 are satisfied. □

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

4.5 The linear case $p = 1$

In the case $p = 1$, for every $n \ge 4$, the matrix $\frac{1}{n}A_n^{[1]}$ is of size $(n-1)\times(n-1)$ and is given by (64) where

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

$$
M_n^{[1]} = \frac{1}{6}\begin{bmatrix} 4 & 1 & & & \\ 1 & 4 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & 4 & 1 \\ & & & 1 & 4 \end{bmatrix}.
$$

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

We can give an explicit expression for the eigenvalues of $\frac{1}{n}A_n^{[1]}$. To this end, we recall the following simple result, see e.g. [32, p. 154]. Let $X := \text{Tridiagonal}(a, b, c) \in \mathbb{R}^{m\times m}$ be a real Toeplitz tridiagonal matrix such that $ac > 0$. Then, X has m real distinct eigenvalues,

$$
\lambda_j(X) = b + 2\sqrt{ac}\,\cos\frac{j\pi}{m+1}, \quad j = 1, \ldots, m. \tag{70}
$$

Note that $\frac{1}{n}A_n^{[1]}$ is a real Toeplitz tridiagonal matrix, namely

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

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. In this case we can apply (70), and we obtain the following theorem.

Theorem 13 *Let $n \geq 4$ 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,*

$$\lambda_j\left(\frac{1}{n}A_n^{[1]}\right) = 2+\frac{2\gamma}{3n^2}+2\sqrt{1-\left(\frac{\gamma}{3}+\frac{\beta^2}{4}\right)\frac{1}{n^2}+\frac{\gamma^2}{36n^4}}\cos\frac{j\pi}{n}, \tag{71}$$

for $j = 1,\dots,n-1$.

By using the expression (71) 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.

For all $n \geq 4$ such that $-1-\frac{\beta}{2n}+\frac{\gamma}{6n^2}$ and $-1+\frac{\beta}{2n}+\frac{\gamma}{6n^2}$ are both negative, by using (71) and some asymptotic expansion, one can also prove that

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

Moreover, by the first Gershgorin theorem [8], 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 (71) it also follows that

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

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 $f_1(\theta) = 2-2\cos\theta$ (cf. 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 11 *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 (71) 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 (64) where

$$K_n^{[2]} = \frac{1}{6}\begin{bmatrix} 8 & -1 & -1 & & & & \\ -1 & 6 & -2 & -1 & & & \\ -1 & -2 & 6 & -2 & -1 & & \\ & \ddots & \ddots & \ddots & \ddots & \ddots & \\ & & -1 & -2 & 6 & -2 & -1 \\ & & & -1 & -2 & 6 & -1 \\ & & & & -1 & -1 & 8 \end{bmatrix},$$

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

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

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 12 *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 $n^2 > \frac{5}{4}\gamma$, then $K_n^{[2]} + \frac{\gamma}{n^2}M_n^{[2]}$ is symmetric, irreducible, and*

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

Lemma 12 implies that $\lambda_{\min}(M_n^{[2]}) > \frac{1}{10}$ for all $n \geq 5$. From Theorem 8 it follows that

$$\lambda_{\min}(K_n^{[2]}) > \frac{\pi^2}{10n^2}, \quad \forall n \geq 5. \tag{72}$$

The next lemma concerns the low-rank matrix $R_n^{[2]}$ introduced in (65).

Lemma 13 *For every $n \geq 5$, the characteristic polynomial of $R_n^{[2]}$ 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 $n^2 > \frac{5}{4}\gamma$,*

$$\sigma\left(\frac{1}{n}A_n^{[2]}\right) \subset \left(\max\left(\frac{\gamma}{n^2}, \frac{\pi^2+\gamma}{10n^2}\right), \min\left(\frac{3}{2} + \frac{1+\sqrt{2}}{6} + \frac{\gamma}{n^2}, 2 + \frac{\gamma}{10n^2}\right)\right) \times \left[-\frac{11|\beta|}{12n}, \frac{11|\beta|}{12n}\right] \subset \mathbb{C}. \tag{73}$$

Proof Fix $n \geq 5$ such that the condition $n^2 > \frac{5}{4}\gamma$ is met. By computing the real and imaginary part of $\frac{1}{n}A_n^{[2]}$, we obtain

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

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

We begin with $\sigma\left(\mathrm{Re}\,\frac{1}{n}A_n^{[2]}\right)$. Since n satisfies the condition $n^2 > \frac{5}{4}\gamma$, by Lemma 12 we obtain

$$\sigma\left(\mathrm{Re}\,\frac{1}{n}A_n^{[2]}\right) \subset \left(\frac{\gamma}{n^2}, 2+\frac{\gamma}{10n^2}\right). \tag{74}$$

We can improve (74) as follows. By combining the minimax principle with Lemmas 9, 12 and 13, and taking into account that $M_{f_2} = \frac{3}{2}$, we obtain

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

Similarly, by using the minimax principle, the bound (72) and Lemma 12,

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

Thus, we can replace (74) with

$$\sigma\left(\mathrm{Re}\,\frac{1}{n}A_n^{[2]}\right) \subset \left(\max\left(\frac{\gamma}{n^2}, \frac{\pi^2+\gamma}{10n^2}\right), \min\left(\frac{3}{2}+\frac{1+\sqrt{2}}{6}+\frac{\gamma}{n^2}, 2+\frac{\gamma}{10n^2}\right)\right). \tag{75}$$

Now we localize the spectrum $\sigma\left(\mathrm{Im}\,\frac{1}{n}A_n^{[2]}\right)$. Since $\mathrm{Im}\,\frac{1}{n}A_n^{[2]}$ is Hermitian, from Lemma 12 we obtain[2]

$$\sigma\left(\mathrm{Im}\,\frac{1}{n}A_n^{[2]}\right) \subset \left[-\frac{11|\beta|}{12n}, \frac{11|\beta|}{12n}\right]. \tag{76}$$

Combining (75)–(76) with (6), we obtain (73). □

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. [8].

[2] If $\beta \neq 0$ then $\mathrm{Im}\,\frac{1}{n}A_n^{[2]}$ is irreducible and $\sigma\left(\mathrm{Im}\,\frac{1}{n}A_n^{[2]}\right) \subset \left(-\frac{11|\beta|}{12n}, \frac{11|\beta|}{12n}\right)$. In (76) 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

$$\lambda_1(B) \geq \cdots \geq \lambda_m(B) \quad \text{and} \quad \lambda_1(K) \geq \cdots \geq \lambda_m(K)$$

be the eigenvalues of B and K arranged in non-increasing order. Then

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

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

The second result is the Ky–Fan theorem [8, 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 non-increasing order:*

$$\mathrm{Re}(\lambda_1(A)) \geq \cdots \geq \mathrm{Re}(\lambda_m(A)) \quad \text{and} \quad \lambda_1(\mathrm{Re}\, A) \geq \cdots \geq \lambda_m(\mathrm{Re}\, A).$$

Then

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

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

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

$$q_n^+(\varepsilon) \leq \left\lceil \frac{1+\sqrt{2}}{3\varepsilon} \right\rceil, \tag{77}$$

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

Proof For every $n \geq 5$, we consider the decomposition $K_n^{[2]} = B_n^{[2]} + R_n^{[2]}$ introduced in (65). The matrix $R_n^{[2]}$ is symmetric and we know the eigenvalues of $R_n^{[2]}$ from Lemma 13. 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,

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

for every $j = 3, \ldots, n-2$, where the eigenvalues of $B_n^{[2]}$ and $K_n^{[2]}$ are arranged in non-increasing order. In particular, from Lemma 9 and $M_{f_2} = \frac{3}{2}$, it follows that $\sigma(B_n^{[2]}) \subset \left(0, \frac{3}{2}\right)$, and

$$\frac{3}{2} > \lambda_1(B_n^{[2]}) \geq \lambda_3(K_n^{[2]}) \geq \cdots \geq \lambda_n(K_n^{[2]}) > 0, \tag{78}$$

where the last inequality is a consequence of Lemma 12. Moreover, by the minimax principle,

$$\lambda_{\max}(K_n^{[2]}) = \lambda_{\max}(B_n^{[2]} + R_n^{[2]}) \leq \lambda_{\max}(B_n^{[2]}) + \lambda_{\max}(R_n^{[2]}) < \frac{3}{2} + \frac{1+\sqrt{2}}{6}. \tag{79}$$

Assume that the eigenvalues of $\frac{1}{n}A_n^{[2]}$ and $\operatorname{Re}\frac{1}{n}A_n^{[2]}$ are arranged in non-increasing order. Recalling from Lemma 12 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

$$\begin{aligned}\lambda_j\left(\operatorname{Re}\frac{1}{n}A_n^{[2]}\right) &= \min_{\substack{V\subseteq\mathbb{C}^n \\ \dim V = n+1-j}} \max_{\substack{\mathbf{x}\in V \\ \|\mathbf{x}\|=1}} \left(\mathbf{x}^*\left(K_n^{[2]} + \frac{\gamma}{n^2}M_n^{[2]}\right)\mathbf{x}\right) \\ &< \min_{\substack{V\subseteq\mathbb{C}^n \\ \dim V = n+1-j}} \max_{\substack{\mathbf{x}\in V \\ \|\mathbf{x}\|=1}} \left(\mathbf{x}^* K_n^{[2]}\mathbf{x} + \frac{\gamma}{n^2}\right) = \lambda_j(K_n^{[2]}) + \frac{\gamma}{n^2}. \end{aligned} \tag{80}$$

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 [17, proof of Theorem 3.5] and keeping in mind (78)–(80), we apply Theorem 16 to obtain

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

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

$$q_n^+(\varepsilon) < \frac{1+\sqrt{2}}{3\left(\varepsilon - \frac{\gamma}{n^2}\right)}. \tag{81}$$

Note that if $0 < \varepsilon < 1$ and $n \geq \frac{\sqrt{2\gamma}}{\varepsilon}$, then

$$\frac{1+\sqrt{2}}{3\left(\varepsilon - \frac{\gamma}{n^2}\right)} \leq \frac{1+\sqrt{2}}{3\left(\varepsilon - \frac{\varepsilon^2}{2}\right)} \leq \frac{1+\sqrt{2}}{3\varepsilon} + 1. \tag{82}$$

From (81)–(82) it follows that (77) 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 (73) and (77), we are able to find an upper bound for $q_n(\varepsilon)$. Indeed, $\forall \varepsilon \in (0,1)$ and $\forall n \geq \max\left(5, \frac{11|\beta|}{12\varepsilon}, \frac{\sqrt{2\gamma}}{\varepsilon}\right) = O\left(\frac{1}{\varepsilon}\right)$,

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

5 The 2D setting

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

$$\begin{cases} -\Delta u(x,y) + \boldsymbol{\beta} \cdot \nabla u(x,y) + \gamma u(x,y) = \mathrm{f}(x,y), & \forall (x,y) \in \Omega, \\ u(x,y) = 0, & \forall (x,y) \in \partial\Omega, \end{cases} \tag{83}$$

with $\mathrm{f} \in L_2((0,1)^2)$, $\boldsymbol{\beta} = [\beta_1\ \beta_2]^T \in \mathbb{R}^2$, $\gamma \geq 0$. In order to approximate the weak solution of problem (83) 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 Sect. 4 (for the x and y directions):

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

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

where

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

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

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

where

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

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

$$\mathscr{W}_{n_1,n_2}^{[p_1,p_2]} := \langle N_{i,j,[p_1,p_2]} :\ i = 2, \ldots, n_1 + p_1 - 1,\ j = 2, \ldots, n_2 + p_2 - 1 \rangle, \tag{84}$$

and we consider the elements of the basis (84) ordered as follows:

$$\varphi_{(n_1+p_1-2)(j-1)+i} = N_{i+1,j+1,[p_1,p_2]}, \tag{85}$$

with $i = 1, \ldots, n_1 + p_1 - 2,\ j = 1, \ldots, n_2 + p_2 - 2$.

Once we have fixed the tensor-product B-spline basis (84) ordered as in (85), 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:

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

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

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

Using the integration rules described in Sect. 4.1, we obtain that

$$A_{n_1,n_2}^{[p_1,p_2]} = \frac{n_1}{n_2} \widehat{K}_{n_1,n_2}^{[p_1,p_2]} + \frac{n_2}{n_1} \widetilde{K}_{n_1,n_2}^{[p_1,p_2]} + \frac{\beta_1}{n_2} \widehat{H}_{n_1,n_2}^{[p_1,p_2]} + \frac{\beta_2}{n_1} \widetilde{H}_{n_1,n_2}^{[p_1,p_2]} + \frac{\gamma}{n_1 n_2} M_{n_1,n_2}^{[p_1,p_2]}, \tag{86}$$

where

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

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

$$A_{n,n}^{[p,p]} = K_{n,n}^{[p,p]} + \frac{\beta_1}{n} \widehat{H}_{n,n}^{[p,p]} + \frac{\beta_2}{n} \widetilde{H}_{n,n}^{[p,p]} + \frac{\gamma}{n^2} M_{n,n}^{[p,p]}, \tag{87}$$

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 (86) under the additional mild assumption that the ratio $\frac{n_2}{n_1} =: \nu$ is constant as $n_1 \to \infty$.[3] With this assumption we have

$$A_{n_1,n_2}^{[p_1,p_2]} = \frac{1}{\nu} \widehat{K}_{n_1,n_2}^{[p_1,p_2]} + \nu \widetilde{K}_{n_1,n_2}^{[p_1,p_2]} + \frac{\beta_1}{\nu n_1} \widehat{H}_{n_1,n_2}^{[p_1,p_2]}$$
$$+ \frac{\beta_2}{n_1} \widetilde{H}_{n_1,n_2}^{[p_1,p_2]} + \frac{\gamma}{\nu (n_1)^2} M_{n_1,n_2}^{[p_1,p_2]}. \tag{88}$$

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

$$\widehat{K}_{n_1,n_2}^{[p_1,p_2]} = \widehat{B}_{n_1,n_2}^{[p_1,p_2]} + \widehat{R}_{n_1,n_2}^{[p_1,p_2]}, \quad \widetilde{K}_{n_1,n_2}^{[p_1,p_2]} = \widetilde{B}_{n_1,n_2}^{[p_1,p_2]} + \widetilde{R}_{n_1,n_2}^{[p_1,p_2]}, \tag{89}$$

where

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

and

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

[3] 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.

We recall that the matrices $B_n^{[p]}$, $R_n^{[p]}$, $C_n^{[p]}$, $S_n^{[p]}$ were introduced in Sect. 4.4, see (65)–(68). Finally, we define

$$B_{n_1,n_2}^{[p_1,p_2]} := \frac{1}{\nu}\widehat{B}_{n_1,n_2}^{[p_1,p_2]} + \nu\widetilde{B}_{n_1,n_2}^{[p_1,p_2]}, \tag{90}$$

$$R_{n_1,n_2}^{[p_1,p_2]} := \frac{1}{\nu}\widehat{R}_{n_1,n_2}^{[p_1,p_2]} + \nu\widetilde{R}_{n_1,n_2}^{[p_1,p_2]}. \tag{91}$$

From Lemmas 9 and 10 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

$$B_{n_1,n_2}^{[p_1,p_2]} = T_{n_2+p_2-2,n_1+p_1-2}\left(\frac{1}{\nu}h_{p_2} \otimes f_{p_1} + \nu f_{p_2} \otimes h_{p_1}\right). \tag{92}$$

Hence, by Theorem 5,

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

and

$$\{B_{n_1,n_2}^{[p_1,p_2]}\} \overset{\lambda}{\sim} \frac{1}{\nu}h_{p_2} \otimes f_{p_1} + \nu f_{p_2} \otimes h_{p_1}. \tag{93}$$

By Lemma 1 and the inequalities (66) and (68), we have

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

and similarly, we also have $\operatorname{rank}(\widetilde{R}_{n_1,n_2}^{[p_1,p_2]}) = o((n_1 + p_1 - 2)(\nu n_1 + p_2 - 2))$, as $n_1 \to \infty$. Thus,

$$\begin{aligned}
\operatorname{rank}(R_{n_1,n_2}^{[p_1,p_2]}) &\leq \operatorname{rank}(\widehat{R}_{n_1,n_2}^{[p_1,p_2]}) + \operatorname{rank}(\widetilde{R}_{n_1,n_2}^{[p_1,p_2]})\\
&= o((n_1 + p_1 - 2)(\nu n_1 + p_2 - 2)),
\end{aligned} \tag{94}$$

as $n_1 \to \infty$. Note that $(n_1 + p_1 - 2)(\nu n_1 + p_2 - 2)$ is the dimension of the matrix $A_{n_1,n_2}^{[p_1,p_2]}$. Moreover, using Lemmas 1, 9, 10 and the fact that the matrices $K_n^{[p]}$, $H_n^{[p]}$, $M_n^{[p]}$, $B_n^{[p]}$, $C_n^{[p]}$ are normal by Theorem 7, we obtain

$$\begin{aligned}
\|R_{n_1,n_2}^{[p_1,p_2]}\| &\le \frac{1}{\nu}\|\widehat{R}_{n_1,n_2}^{[p_1,p_2]}\| + \nu\|\widetilde{R}_{n_1,n_2}^{[p_1,p_2]}\| \\
&= \frac{1}{\nu}\|\widehat{K}_{n_1,n_2}^{[p_1,p_2]} - \widehat{B}_{n_1,n_2}^{[p_1,p_2]}\| + \nu\|\widetilde{K}_{n_1,n_2}^{[p_1,p_2]} - \widetilde{B}_{n_1,n_2}^{[p_1,p_2]}\| \\
&\le \frac{1}{\nu}\|M_{n_2}^{[p_2]}\|\|K_{n_1}^{[p_1]}\| + \frac{1}{\nu}\|C_{n_2}^{[p_2]}\|\|B_{n_1}^{[p_1]}\| + \nu\|K_{n_2}^{[p_2]}\|\|M_{n_1}^{[p_1]}\| \\
&\quad + \nu\|B_{n_2}^{[p_2]}\|\|C_{n_1}^{[p_1]}\| \\
&\le \frac{1}{\nu}\|M_{n_2}^{[p_2]}\|_\infty\|K_{n_1}^{[p_1]}\|_\infty + \frac{1}{\nu}M_{f_{p_1}} + \nu\|K_{n_2}^{[p_2]}\|_\infty\|M_{n_1}^{[p_1]}\|_\infty + \nu M_{f_{p_2}}.
\end{aligned}$$

From Lemma 8 it follows

$$\|R_{n_1,n_2}^{[p_1,p_2]}\| \le Q_{p_1,p_2}, \tag{95}$$

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

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

$$g_{p_1,p_2} := \frac{1}{\nu}h_{p_2} \otimes f_{p_1} + \nu f_{p_2} \otimes h_{p_1}, \tag{96}$$

in the sense of the eigenvalues.

Proof Let

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

Then, by (88)–(91), we have

$$A_{n_1,n_2}^{[p_1,p_2]} = B_{n_1,n_2}^{[p_1,p_2]} + R_{n_1,n_2}^{[p_1,p_2]} + U_{n_1,n_2}^{[p_1,p_2]}.$$

We now prove that all 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 $Y_{n_1} = R_{n_1,n_2}^{[p_1,p_2]} + U_{n_1,n_2}^{[p_1,p_2]}$. We have seen in (93) that $\{B_{n_1,n_2}^{[p_1,p_2]}\} \overset{\lambda}{\sim} g_{p_1,p_2}$.

We note that $B_{n_1,n_2}^{[p_1,p_2]}$ is symmetric and that g_{p_1,p_2} is nonnegative over its domain $[-\pi,\pi]^2$. 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

$$\|U_{n_1,n_2}^{[p_1,p_2]}\| \le \frac{|\beta_1|}{\nu n_1}\|M_{n_2}^{[p_2]}\|_\infty \|H_{n_1}^{[p_1]}\|_\infty + \frac{|\beta_2|}{n_1}\|H_{n_2}^{[p_2]}\|_\infty \|M_{n_1}^{[p_1]}\|_\infty + \frac{\gamma}{\nu(n_1)^2}\|M_{n_2}^{[p_2]}\|_\infty \|M_{n_1}^{[p_1]}\|_\infty,$$

and from Lemma 8 it follows that

$$\|U_{n_1,n_2}^{[p_1,p_2]}\| = O\left(\frac{1}{n_1}\right). \tag{97}$$

Combining (95) and (97), we obtain

$$\|R_{n_1,n_2}^{[p_1,p_2]} + U_{n_1,n_2}^{[p_1,p_2]}\| \le \bar{Q}_{p_1,p_2},$$

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

On the other hand, by using (94)–(95) and (97), we get

$$\begin{aligned}
&\|R_{n_1,n_2}^{[p_1,p_2]} + U_{n_1,n_2}^{[p_1,p_2]}\|_1 \\
&\le \|R_{n_1,n_2}^{[p_1,p_2]}\|_1 + \|U_{n_1,n_2}^{[p_1,p_2]}\|_1 \\
&\le \operatorname{rank}(R_{n_1,n_2}^{[p_1,p_2]})\, Q_{p_1,p_2} + (n_1+p_1-2)(\nu n_1+p_2-2)\, O\left(\frac{1}{n_1}\right) \\
&= o((n_1+p_1-2)(\nu n_1+p_2-2)), \quad \text{as } n_1 \to \infty.
\end{aligned}$$

Hence, all the hypotheses of Theorem 1 are satisfied, and it follows that the function (96) is the spectral distribution symbol of the sequence $\{A_{n_1,n_2}^{[p_1,p_2]}\}_{n_1}$. □

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 $n_1 = n_2 = n \ge 4$, the matrix $A_{n,n}^{[1,1]}$ is of size $(n-1)^2 \times (n-1)^2$, see (87). Theorem 18 reads in this case 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 $n^2 > \frac{\gamma}{3}$*

$$\sigma(A_{n,n}^{[1,1]}) \subset \left(\max\left(\frac{\gamma}{n^2}, \frac{8}{3}\left(\sin\frac{\pi}{2n}\right)^2 + \frac{\gamma}{9n^2}\right), \min\left(4 + \frac{\gamma}{n^2}, \frac{16}{3} - \frac{\gamma}{9n^2}\right)\right) \times \left[-\frac{|\beta_1| + |\beta_2|}{n}, \frac{|\beta_1| + |\beta_2|}{n}\right] \subset \mathbb{C}. \tag{98}$$

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

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

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

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

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

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 (90). Therefore, by (92) 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, from Lemmas 1 and 11 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

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

In addition, again by the minimax principle, by Lemmas 1 and 11, and by the fact that $\lambda_{\min}(K_n^{[1]}) = 4\left(\sin\frac{\pi}{2n}\right)^2$, we obtain

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

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

$$\sigma(\operatorname{Re} A_{n,n}^{[1,1]}) \subset \left(\max\left(\frac{\gamma}{n^2}, \frac{8}{3}\left(\sin\frac{\pi}{2n}\right)^2 + \frac{\gamma}{9n^2}\right), \min\left(4 + \frac{\gamma}{n^2}, \frac{16}{3} - \frac{\gamma}{9n^2}\right)\right). \tag{99}$$

We now localize the spectrum $\sigma(\mathrm{Im}\, A_{n,n}^{[1,1]})$. By Lemmas 1 and 11, we get $\sigma(\widehat{H}_{n,n}^{[1,1]}) = \sigma(\widetilde{H}_{n,n}^{[1,1]}) \subset \{0\} \times (-1, 1)$. By means of the minimax principle, it follows that

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

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

$$\sigma(\mathrm{Im}\, A_{n,n}^{[1,1]}) \subseteq \left[-\frac{|\beta_1| + |\beta_2|}{n}, \frac{|\beta_1| + |\beta_2|}{n}\right]. \tag{100}$$

Combining (6) with (99)–(100), we get (98). □

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 2 *$\forall \varepsilon \in (0, 1)$ 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 $[0, 4]_\varepsilon$.

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 $n^2 > \frac{\gamma}{3}$, $\frac{\gamma}{n^2} \leq \varepsilon$ and $\frac{|\beta_1|+|\beta_2|}{n} \leq \varepsilon$, by Theorem 19 we have

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

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 of size $n^2 \times n^2$, see (87). Theorem 18 reads in this case as $\{A_{n,n}^{[2,2]}\} \overset{\lambda}{\sim} g_{2,2}$, with

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

5.4.1 Localization of the eigenvalues

Theorem 20 *For every $n \geq 5$ such that $n^2 > \frac{5}{4}\gamma$*

$$\sigma(A_{n,n}^{[2,2]}) \subset \left(\max\left(\frac{\pi^2+10\gamma}{100n^2}, \frac{2\pi^2+\gamma}{100n^2}\right), \frac{49}{24}+\frac{\gamma}{n^2}\right) \times \left[-\frac{11}{12}\frac{|\beta_1|+|\beta_2|}{n}, \frac{11}{12}\frac{|\beta_1|+|\beta_2|}{n}\right] \subset \mathbb{C}. \tag{101}$$

Proof Fix $n \geq 5$ such that the condition $n^2 > \frac{5}{4}\gamma$ is met. From (87) we know that

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

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

First we localize the spectrum of $\text{Re } A_{n,n}^{[2,2]}$. From the minimax principle it follows that $\lambda_{\min}(\text{Re } A_{n,n}^{[2,2]}) \geq \lambda_{\min}(M_n^{[2]} \otimes K_n^{[2]}) + \lambda_{\min}(K_n^{[2]} \otimes M_n^{[2]}) + \frac{\gamma}{n^2}\lambda_{\min}(M_n^{[2]} \otimes M_n^{[2]})$. Then, by using Lemmas 1 and 12, and by (72), we get

$$\lambda_{\min}(\text{Re } A_{n,n}^{[2,2]}) > 2 \cdot \frac{\pi^2}{10n^2}\frac{1}{10} + \frac{\gamma}{100n^2} = \frac{2\pi^2+\gamma}{100n^2}. \tag{102}$$

In addition, the minimax principle also implies that

$$\lambda_{\min}(\text{Re } A_{n,n}^{[2,2]}) \geq \lambda_{\min}\left(M_n^{[2]} \otimes K_n^{[2]}\right) + \lambda_{\min}\left(\left(K_n^{[2]} + \frac{\gamma}{n^2}M_n^{[2]}\right) \otimes M_n^{[2]}\right).$$

Because $n^2 > \frac{5}{4}\gamma$, we can use the bound given in Lemma 12 for the spectrum of the matrix $K_n^{[2]} + \frac{\gamma}{n^2}M_n^{[2]}$. Then, by Lemmas 1 and 12, and by (72), we get

$$\lambda_{\min}(\text{Re } A_{n,n}^{[2,2]}) > \frac{1}{10}\frac{\pi^2}{10n^2} + \frac{\gamma}{n^2}\frac{1}{10} = \frac{\pi^2+10\gamma}{100n^2}. \tag{103}$$

Furthermore, since $K_{n,n}^{[2,2]} = B_{n,n}^{[2,2]} + R_{n,n}^{[2,2]}$, we know that $\text{Re } A_{n,n}^{[2,2]} = B_{n,n}^{[2,2]} + R_{n,n}^{[2,2]} + \frac{\gamma}{n^2}M_{n,n}^{[2,2]}$. We recall from (92) 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 the first Gershgorin theorem that $\sigma(R_{n,n}^{[2,2]}) \subset \left[-\frac{269}{360}, \frac{13}{24}\right]$. Using Lemmas 1 and 12, we also find that $\sigma(M_{n,n}^{[2,2]}) \subset (\frac{1}{100}, 1)$. Then, we apply again the minimax principle to obtain the upper bound $\lambda_{\max}(\text{Re } A_{n,n}^{[2,2]}) \leq \lambda_{\max}(B_{n,n}^{[2,2]}) + \lambda_{\max}(R_{n,n}^{[2,2]}) + \frac{\gamma}{n^2}\lambda_{\max}(M_{n,n}^{[2,2]})$ so that

$$\lambda_{\max}(\text{Re } A_{n,n}^{[2,2]}) < \frac{3}{2} + \frac{13}{24} + \frac{\gamma}{n^2} = \frac{49}{24} + \frac{\gamma}{n^2}. \tag{104}$$

Now we localize the spectrum of $\mathrm{Im}\, A_{n,n}^{[2,2]}$. By Lemmas 1 and 12, we have $\sigma(\widehat{H}_{n,n}^{[2,2]}) = \sigma(\widetilde{H}_{n,n}^{[2,2]}) \subset \{0\} \times (-\frac{11}{12}, \frac{11}{12})$, and hence, by the minimax principle,

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

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,

$$\sigma(\mathrm{Im}\, A_{n,n}^{[2,2]}) \subseteq \left[-\frac{11}{12}\frac{|\beta_1|+|\beta_2|}{n}, \frac{11}{12}\frac{|\beta_1|+|\beta_2|}{n}\right]. \tag{105}$$

By using (6) in combination with (102)–(104) and (105), we obtain (101). □

6 Conclusions

We have studied 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 have provided a spectral characterization of the involved matrices. In particular, we have given an asymptotic analysis of

1. the eigenvalue of minimum modulus and the eigenvalue of maximum 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. These generalizations yield the loss of the Toeplitz structure. Nevertheless, we expect that the global symbol of the associated matrix sequences can 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 [19]), 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., Lindskog, G.: On the rate of convergence of the preconditioned conjugate gradient method. Numer. Math. **48**, 499–523 (1986)
3. 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. Eng. **199**, 229–263 (2010)

4. Beckermann, B., Kuijlaars, A.B.J.: Superlinear convergence of Conjugate Gradients. SIAM J. Numer. Anal. **39**, 300–329 (2001)
5. Beckermann, B., Kuijlaars, A.B.J.: On the sharpness of an asymptotic error estimate for Conjugate Gradients. BIT **41**, 856–867 (2001)
6. Beckermann, B., Serra-Capizzano, S.: On the asymptotic spectrum of Finite Elements matrices. SIAM J. Numer. Anal. **45**, 746–769 (2007)
7. 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)
8. Bhatia, R.: Matrix Analysis. Springer, New York (1997)
9. de Boor, C.: A Practical Guide to Splines. Springer, New York (2001)
10. Böttcher, A., Silbermann, B.: Introduction to Large Truncated Toeplitz Matrices. Springer, New York (1999)
11. Böttcher, A., Widom, H.: From Toeplitz eigenvalues through Green's kernels to higher-order Wirtinger–Sobolev inequalities. Oper. Theory Adv. Appl. **171**, 73–87 (2007)
12. Buffa, A., Harbrecht, H., Kunoth, A., Sangalli, G.: BPX-preconditioning for isogeometric analysis. Comput. Methods Appl. Mech. Eng. **265**, 63–70 (2013)
13. Chui, C.K.: An Introduction to Wavelets. Academic Press, San Diego (1992)
14. Cottrell, J.A., Hughes, T.J.R., Bazilevs, Y.: Isogeometric Analysis: Toward Integration of CAD and FEA. Wiley, Chichester (2009)
15. Gahalaut, K.P.S., Kraus, J.K., Tomar, S.K.: Multigrid methods for isogeometric discretization. Comput. Methods Appl. Mech. Eng. **253**, 413–425 (2013)
16. Garoni, C., Manni, C., Pelosi, F., Serra-Capizzano, S., Speleers, H.: On the spectrum of stiffness matrices arising from isogeometric analysis. Technical Report TW632, Department of Computer Science, KU Leuven (2013)
17. Golinskii, L., Serra-Capizzano, S.: The asymptotic properties of the spectrum of nonsymmetrically perturbed Jacobi matrix sequences. J. Approx. Theory **144**, 84–102 (2007)
18. Grenander, U., Szegö, G.: Toeplitz Forms and Their Applications, 2nd edn. Chelsea, New York (1984)
19. Hörmander, L.: Pseudo-differential operators and non-elliptic boundary problems. Ann. Math. **2**, 129–209 (1966)
20. Hughes, T.J.R., Cottrell, J.A., Bazilevs, Y.: Isogeometric analysis: CAD, finite elements, NURBS, exact geometry and mesh refinement. Comput. Methods Appl. Mech. Eng. **194**, 4135–4195 (2005)
21. Manni, C., Pelosi, F., Sampoli, M.L.: Generalized B-splines as a tool in isogeometric analysis. Comput. Methods Appl. Mech. Eng. **200**, 867–881 (2011)
22. Parter, S.V.: On the extreme eigenvalues of truncated Toeplitz matrices. Bull. Am. Math. Soc. **67**, 191–197 (1961)
23. Parter, S.V.: On the eigenvalues of certain generalizations of Toeplitz matrices. Arch. Ration. Math. Mech. **3**, 244–257 (1962)
24. Quarteroni, A.: Numerical Models for Differential Problems. Springer, Italy (2009)
25. 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)
26. Saad, Y.: Iterative Methods for Sparse Linear Systems. PWS Publishing, Boston (1996)
27. Schumaker, L.L.: Spline Functions: Basic Theory, 3rd edn. Cambridge Mathematical Library, Cambridge (2007)
28. Serra-Capizzano, S.: Preconditioning strategies for asymptotically ill-conditioned block Toeplitz systems. BIT **34**, 579–594 (1994)
29. Serra-Capizzano, S.: Generalized Locally Toeplitz sequences: spectral analysis and applications to discretized partial differential equations. Linear Algebra Appl. **366**, 371–402 (2003)
30. Serra-Capizzano, S.: GLT sequences as a generalized Fourier analysis and applications. Linear Algebra Appl. **419**, 180–233 (2006)
31. 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)
32. Smith, G.D.: Numerical Solution of Partial Differential Equations: Finite Difference Methods, 3rd edn. Clarendon Press, Oxford (1985)

33. Speleers, H., Manni, C., Pelosi, F., Sampoli, M.L.: Isogeometric analysis with Powell–Sabin splines for advection–diffusion–reaction problems. Comput. Methods Appl. Mech. Eng. **221–222**, 132–148 (2012)
34. Tilli, P.: A note on the spectral distribution of Toeplitz matrices. Linear Multilinear Algebra **45**, 147–159 (1998)
35. Tilli, P.: Locally Toeplitz sequences: spectral properties and applications. Linear Algebra Appl. **278**, 91–120 (1998)
36. van der Sluis, A., van der Vorst, H.A.: The rate of convergence of conjugate gradients. Numer. Math. **48**, 543–560 (1986)