S.I.: EMG 2014

Two-grid optimality for Galerkin linear systems based on B-splines

Marco Donatelli[1] · Carlo Garoni[1,2] · Carla Manni[2] · Stefano Serra-Capizzano[1,3] · Hendrik Speleers[2]

Received: 12 November 2014 / Accepted: 18 December 2015 / Published online: 18 January 2016

Abstract A multigrid method for linear systems stemming from the Galerkin B-spline discretization of classical second-order elliptic problems is considered. The spectral features of the involved stiffness matrices, as the fineness parameter h tends to zero, have been deeply studied in previous works, with particular attention to the dependencies of the spectrum on the degree p of the B-splines used in the discretization process. Here, by exploiting this information in connection with τ-matrices, we describe a multigrid strategy and we prove that the corresponding two-grid iterations have a convergence rate independent of h for $p = 1, 2, 3$. For larger p, the proof may be obtained through algebraic manipulations. Unfortunately, as confirmed by the numerical experiments, the dependence on p is bad and hence other techniques have to be considered for large p.

Keywords Multigrid methods · Isogeometric analysis · B-splines · τ-matrices

Mathematics Subject Classification 65N55 · 65N30 · 65F08

1 Introduction

In this paper we consider the differential problem

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

with $\Omega := (0, 1)^d$, $\mathrm{f} \in L^2(\Omega)$, $\gamma \geq 0$. We are interested in designing a multigrid strategy for the fast numerical solution of large linear systems stemming from the discretization of (1) by the Galerkin B-spline isogeometric analysis (IgA) technique; see [7,18].

In [16] we studied in detail the spectral properties of the resulting stiffness matrices based on uniform tensor-product B-splines. Not only the spectral localization and the conditioning were investigated, but also the global spectral behavior. This spectral behavior can be described in the Weyl sense (see [24] and the literature therein) in terms of a d-variate trigonometric polynomial $f_{\boldsymbol{p}}$, the so-called (spectral) symbol; here, $\boldsymbol{p} := (p_1, \dots, p_d)$ and p_i is the spline degree in the i-th direction, $i = 1, \dots, d$. It turns out that the symbol $f_{\boldsymbol{p}}$ is equivalent to the classical symbol

$$z_d(\boldsymbol{\theta}) = z_d(\theta_1, \dots, \theta_d) := \sum_{j=1}^{d} (2 - 2\cos\theta_j), \tag{2}$$

Communicated by Artem Napov, Yvan Notay, and Stefan Vandewalle.

✉ Hendrik Speleers
speleers@mat.uniroma2.it

Marco Donatelli
marco.donatelli@uninsubria.it

Carlo Garoni
carlo.garoni@uninsubria.it; garoni@mat.uniroma2.it

Carla Manni
manni@mat.uniroma2.it

Stefano Serra-Capizzano
stefano.serrac@uninsubria.it; stefano.serra@it.uu.se

[1] Department of Science and High Technology, University of Insubria, Via Valleggio 11, 22100 Como, Italy

[2] Department of Mathematics, University of Rome 'Tor Vergata', Via della Ricerca Scientifica, 00133 Rome, Italy

[3] Division of Scientific Computing, Department of Information Technology, Uppsala University, Box 337, 751 05 Uppsala, Sweden

which is obtained when approximating (1) by standard uniform centered second-order Finite Differences (FD) or by piecewise linear Finite Elements (FE) on uniform triangulations. In other words, there exist positive constants $c_{\boldsymbol{p}}, C_{\boldsymbol{p}}$ such that

$$c_{\boldsymbol{p}} z_d(\boldsymbol{\theta}) \le f_{\boldsymbol{p}}(\boldsymbol{\theta}) \le C_{\boldsymbol{p}} z_d(\boldsymbol{\theta}), \quad \forall \boldsymbol{\theta} \in [0,\pi]^d. \tag{3}$$

From (3) we expect that the conditioning of the Galerkin B-spline IgA stiffness matrices grows like $m^{2/d}$, where m is the matrix size, d is the dimensionality of the elliptic problem, and 2 is the order of the elliptic operator in (1). The approximation parameters $\boldsymbol{p}$ play a limited role, and characterize the constant in the expression $O(m^{2/d})$. We refer the reader to [15, Theorem 4.6] for a rigorous proof of the above statements.[1]

Moreover, in view of the equivalence (3) and given the τ-like (resp., Toeplitz-like) structure of the considered stiffness matrices, we expect that a standard multigrid procedure designed for τ (resp., Toeplitz) linear systems with symbol z_d has to be optimal also in our context. The optimality was predicted and validated through numerical experiments in [9], but a formal proof was not yet given in the literature. In this paper, we formally prove the optimality for the two-grid method and certain values of $\boldsymbol{p}$, and hence also for the W-cycle and k-grid method (with k independent of the matrix size).

In order to design our optimal multigrid solver, we heavily rely on the spectral and structural information of the coefficient matrices analyzed in detail in [16]. More precisely, the stiffness matrices arising from the Galerkin B-spline IgA discretization of problem (1) are:

- banded in a d-level sense with partial bandwidths proportional to p_j, $j = 1, \ldots, d$;
- a small perturbation of a d-level τ-matrix (or Toeplitz matrix) generated by $f_{\boldsymbol{p}}$, and are spectrally distributed like the symbol $f_{\boldsymbol{p}}$ in the Weyl sense.

The first item implies that optimal methods should have a total cost which is linear with respect to the matrix size and with a constant proportional to $\|\boldsymbol{p}\|_\infty$. The second item suggests to look for optimal methods in the wide literature of multilevel τ (or Toeplitz) solvers [19]. In this paper, inspired by [12,22], we follow a sort of 'canonical procedure' for creating—on the basis of the symbol—a two-grid method from which we expect optimal convergence properties. The design of the method is based on analogous optimal techniques for τ and Toeplitz matrices with symbol $f_{\boldsymbol{p}}$ or, equivalently, with symbol z_d, because of the relation (3). The optimality of the method was already predicted in [9] on the basis of heuristic arguments, but no rigorous proof was given. Here, we provide the formal proof, at least for $p_j \le 3$, $j = 1, \ldots, d$. When proving the optimality result, we arrive at a matrix inequality, see (26), which is useful not only in a multigrid setting, but also in a preconditioning context for designing optimal preconditioners for Krylov-type techniques, in particular for the Conjugate Gradient (CG) method; see Remark 6.

It is worth mentioning that our proof of optimality is based on standard tools from algebraic multigrid analysis [20], applied within the framework of the theory of τ and Toeplitz matrices [22]. Our choice to consider the formalism of multigrid methods for τ and Toeplitz matrices instead of the classical Local Fourier Analysis (LFA) is motivated by the possibility to apply directly the results in [16]. Nevertheless, as proved in [8], the convergence analysis based on the symbol of τ and Toeplitz matrices arising in the discretization of differential problems is equivalent to the LFA. The reader who is familiar with the LFA can follow the proof just being aware of two facts:

- the symbol is not scaled by the discretization step;
- the information on the order of the differential problem survives in the order of the zero of the symbol.

We refer to [8] for further details.

Unfortunately, it turns out that the equivalence constant $c_{\boldsymbol{p}}$ in (3) tends to 0 when $p_j \to \infty$ for some j, i.e., when $\|\boldsymbol{p}\|_\infty \to \infty$. Even worse, two further facts occur [9,11]:

- the convergence to 0 is exponential with respect to each p_j, $j = 1, \ldots, d$;
- the values of $\boldsymbol{\theta}$ for which $f_{\boldsymbol{p}}(\boldsymbol{\theta})$ tends to 0 exponentially are localized at the frontier of $[0,\pi]^d$, where at least one variable θ_j equals π.

As a consequence, despite the m-independence of the two-grid convergence rate (recalling that m is the matrix size), the method is unsatisfactory when $\boldsymbol{p}$ has large entries: we have theoretical optimality, but the spectral radius of the two-grid iteration matrix is close to 1. For instance, in the 1D case the spectral radius of the two-grid iteration matrix tends to 1 exponentially as p increases and a similar phenomenon is observed for any dimensionality d. This unpleasant behavior is due to the analytical properties of the symbol $f_{\boldsymbol{p}}$ and is essentially related to the existence of a subspace of high frequencies associated with very small eigenvalues. This explains the numerical results in [9,14]. In particular, the considered two-grid and the associated V/W-cycle methods converge very fast in low frequencies, but they are slow, for

[1] Take into account that these statements hold when the discretization steps $h_i = 1/n_i$ in each direction x_i tend to 0 with the same speed. This happens, for instance, when $n_i = \nu_i n$, $\boldsymbol{\nu} := (\nu_1, \ldots, \nu_d)$ is fixed and $n \to \infty$.

large $\boldsymbol{p}$, in high frequencies. This fact is nontrivial, but it can be understood in terms of the theory of multilevel τ-matrices (or Toeplitz matrices) and is related to specific analytic features of the symbol. We refer the reader to [11] for a deeper theoretical insight in these statements.

The above intrinsic difficulty can be addressed by following the multi-iterative idea from [21]. Indeed, the m-independent two-grid/multigrid method was successfully combined in [9,11] with a $\boldsymbol{p}$-independent smoother of preconditioned Krylov type (namely a PCG), where the preconditioner was suggested by the symbol. This results in a multi-iterative multigrid solver with m- and $\boldsymbol{p}$-independent convergence rate. A very similar multi-iterative multigrid solver was also constructed for IgA collocation methods in [10].

The remainder of this paper is organized as follows. In Sect. 2 we detail the considered model problem; we define d-level τ-matrices; and we describe the two-grid method for such kind of matrices. In Sect. 3 we recall the spectral properties obtained in [16] for the matrices arising from the discretization of (1) by IgA based on uniform B-splines. In Sect. 4 the optimality of the two-grid applied to our problem is proved for $p = 1, 2, 3$ in 1D and for $1 \le p_1, p_2 \le 3$ in 2D. The proof can be trivially extended to any dimensionality for $1 \le p_1, \dots, p_d \le 3$. In Sect. 5 we provide some numerical examples to support our theoretical analysis. Sect. 6 concludes the work.

2 Preliminaries

We start with a brief description of the Galerkin method applied to (1) in the IgA context. Then, we introduce some auxiliary structures, namely d-level τ-matrices, which are used for designing multigrid algorithms and for studying their convergence features.

2.1 The d-dimensional problem setting

IgA was introduced in [18] aiming to reduce the gap between Finite Element Analysis and Computer-Aided Design (CAD). 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. Thanks to the well-known properties of the B-spline basis (see, e.g., [6]), this approach offers some interesting advantages from the geometric, the analytic, and the computational point of view; see [7,18] and references therein.

Let us now consider our model problem (1). The corresponding weak form reads as follows: find $u \in H_0^1(\Omega)$ such that

$$a(u, v) = \mathrm{F}(v), \quad \forall v \in H_0^1(\Omega), \tag{4}$$

where $a(u, v) := \int_\Omega (\nabla u \cdot \nabla v + \gamma u v)$ and $\mathrm{F}(v) := \int_\Omega f v$. It is well-known that there exists a unique solution u of (4), the so-called weak solution of (1).

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

$$a(u_{\mathcal{W}}, v) = \mathrm{F}(v), \quad \forall v \in \mathcal{W}. \tag{5}$$

Let $\dim \mathcal{W} = N$, and fix a basis $\{\varphi_1, \dots, \varphi_N\}$ for $\mathcal{W}$. It is known that problem (5) always has a unique solution $u_{\mathcal{W}}$, which can be written as $u_{\mathcal{W}} = \sum_{j=1}^N u_j \varphi_j$ and can be computed as follows: find $\mathbf{u} := (u_1, \dots, u_N)^T \in \mathbb{R}^N$ such that

$$A\mathbf{u} = \mathbf{b}, \tag{6}$$

where $A := \left[a(\varphi_j, \varphi_i)\right]_{i,j=1}^N \in \mathbb{R}^{N \times N}$ is the stiffness matrix and $\mathbf{b} := [\mathrm{F}(\varphi_i)]_{i=1}^N \in \mathbb{R}^N$.

In classical FE methods the approximation space $\mathcal{W}$ is usually a space of C^0 piecewise polynomials vanishing on $\partial\Omega$, whereas in the IgA framework $\mathcal{W}$ is a spline space with higher continuity; see [7,18]. In this paper we only consider the IgA setting without any geometry map.

2.2 d-Level τ-matrices

For every $m \in \mathbb{N}$ we denote by $\mathcal{Q}_m$ the symmetric unitary discrete sine transform,

$$\mathcal{Q}_m := \sqrt{\frac{2}{m+1}} \left[\sin\left(\frac{ij\pi}{m+1}\right)\right]_{i,j=1}^m,$$

and for every multi-index $\boldsymbol{m} := (m_1, \dots, m_d) \in \mathbb{N}^d$ we set $\mathcal{Q}_{\boldsymbol{m}} := \mathcal{Q}_{m_1} \otimes \cdots \otimes \mathcal{Q}_{m_d}$.

Definition 1 Given a d-variate function $g : [0, \pi]^d \to \mathbb{R}$ and a multi-index $\boldsymbol{m} \in \mathbb{N}^d$, the d-level τ-matrix $\tau_{\boldsymbol{m}}(g)$ of partial orders $m_1, \dots, m_d$ (and order $m_1 \cdots m_d$) associated with g is defined as

$$\tau_{\boldsymbol{m}}(g) := \mathcal{Q}_{\boldsymbol{m}} \cdot \underset{j_1=1,\dots,m_1}{\operatorname{diag}} \left[\cdots \left[\underset{j_d=1,\dots,m_d}{\operatorname{diag}} g\left(\frac{j_1\pi}{m_1+1}, \dots, \frac{j_d\pi}{m_d+1}\right)\right] \cdots\right] \mathcal{Q}_{\boldsymbol{m}}.$$

The function g is called the generating function of the τ-family $\{\tau_{\boldsymbol{m}}(g)\}_{\boldsymbol{m} \in \mathbb{N}^d}$.

We denote by $C_c(\mathbb{R})$ the set of all continuous functions on $\mathbb{R}$ with compact support. Given $g : [0, \pi]^d \to \mathbb{R}$ in $C([0, \pi]^d)$, one can check that, $\forall F \in C_c(\mathbb{R})$,

$$\lim_{\boldsymbol{m}\to\infty} \frac{1}{m_1\dots m_d} \sum_{j=1}^{m_1\cdots m_d} F(\lambda_j(\tau_{\boldsymbol{m}}(g))) = \frac{1}{\pi^d}\int_{[0,\pi]^d} F(g(\theta_1,\dots,\theta_d))\,\mathrm{d}\theta_1\dots\mathrm{d}\theta_d, \tag{7}$$

where, for a multi-index $\boldsymbol{m} \in \mathbb{N}^d$, $\boldsymbol{m} \to \infty$ means that $\min(m_1,\dots,m_d) \to \infty$. Due to the limit relation (7), the function g is called the symbol of the τ-family $\{\tau_{\boldsymbol{m}}(g)\}_{\boldsymbol{m}\in\mathbb{N}^d}$.

Note that, for every $\boldsymbol{m} \in \mathbb{N}^d$, if g is a linear d-variate cosine trigonometric polynomial then the d-level τ-matrix $\tau_{\boldsymbol{m}}(g)$ coincides with the d-level Toeplitz matrix $T_{\boldsymbol{m}}(g)$ associated with g (see [9, Definition 3.2] for its definition).

2.3 Multigrid methods

Given a linear system of dimension m,

$$A_m\mathbf{u} = \mathbf{b}, \tag{8}$$

we assume to have a convergent stationary iterative method

$$\mathbf{u}^{(k+1)} = S_m\mathbf{u}^{(k)} + (I - S_m)A_m^{-1}\mathbf{b},$$

called smoother, for the solution of (8), and a full-rank matrix $P_m \in \mathbb{R}^{l\times m}$ with $l \le m$, called projector or grid-transfer operator. Moreover, we define the coarse matrix as $P_m A_m P_m^T$, following the Galerkin approach. Then, given an approximation $\mathbf{u}^{(k)}$ to the solution $\mathbf{u} = A_m^{-1}\mathbf{b}$, the corresponding Two-Grid Method (TGM) for solving (8) computes a new approximation $\mathbf{u}^{(k+1)}$ by applying a coarse-grid correction and a smoothing iteration as follows:

Algorithm 1 [TGM]

```
1. compute the residual: $\mathbf{r} \leftarrow \mathbf{b} - A_m\mathbf{u}^{(k)}$;
2. project the residual: $\mathbf{r} \leftarrow P_m\mathbf{r}$;
3. compute the correction: $\mathbf{e} \leftarrow \left(P_m A_m P_m^T\right)^{-1}\mathbf{r}$;
4. extend the coarse error: $\mathbf{e} \leftarrow P_m^T\mathbf{e}$;
5. correct the given approximation: $\mathbf{u}^{(k+1)} \leftarrow \mathbf{u}^{(k)} + \mathbf{e}$;
6. relax one time: $\mathbf{u}^{(k+1)} \leftarrow S_m\mathbf{u}^{(k+1)} + (I - S_m)A_m^{-1}\mathbf{b}$.
```

The iteration matrix of this two-grid scheme is

$$TG(S_m, P_m) := S_m\left(I - P_m^T\left(P_m A_m P_m^T\right)^{-1} P_m A_m\right).$$

Note that Algorithm 1 only considers a single post-smoothing iteration in order to stay in the framework of [20], so as to simplify the presentation of the theoretical analysis, but it is clear that one can add a convergent pre-smoother and/or more smoothing iterations to improve the convergence rate of the TGM.

In practice, the coarser linear system of the TGM could be too large to be solved directly. Hence, the third step in Algorithm 1 is usually replaced by one recursive call, obtaining a multigrid V-cycle algorithm, or by two recursive calls, obtaining a multigrid W-cycle algorithm.

The optimality proofs for the two-grid methods, presented in this paper, heavily rely on a classical result for the two-grid convergence rate, stated in Theorem 2. For its proof, we refer the reader to [20, Theorem 5.2] and [2, Remark 2.2]. Given a Symmetric Positive Definite (SPD) matrix $X \in \mathbb{R}^{m\times m}$, we denote by $\|\cdot\|_X$ both the vector-norm and the matrix-norm induced by X, i.e.,

$$\|\mathbf{x}\|_X := \|X^{1/2}\mathbf{x}\|_2, \qquad \mathbf{x} \in \mathbb{R}^m,$$
$$\|Y\|_X := \|X^{1/2}YX^{-1/2}\|_2, \qquad Y \in \mathbb{R}^{m\times m},$$

where $\|\cdot\|_2$ stands for both the classical 2-norm (the Euclidean norm) and its induced matrix-norm.

Theorem 2 ([20]) *Let $A_m \in \mathbb{R}^{m\times m}$ be SPD, let $S_m \in \mathbb{R}^{m\times m}$, and let $P_m \in \mathbb{R}^{l\times m}$ be full-rank ($l \le m$). Assume*

(a) $\exists\, a_m > 0 : \|S_m\mathbf{x}\|^2_{A_m} \le \|\mathbf{x}\|^2_{A_m} - a_m\|\mathbf{x}\|^2_{A_m^2}$,
(b) $\exists\, b_m > 0 : \min_{\mathbf{y}\in\mathbb{R}^l} \|\mathbf{x} - P_m^T\mathbf{y}\|_2^2 \le b_m\|\mathbf{x}\|^2_{A_m}$,

for all $\mathbf{x} \in \mathbb{R}^m$. Then $b_m \ge a_m$ and

$$\rho\left(TG(S_m, P_m)\right) \le \|TG(S_m, P_m)\|_{A_m} \le \sqrt{1 - \frac{a_m}{b_m}}.$$

The first condition (a) in Theorem 2 is referred to as the *smoothing condition*, whereas the second condition (b) as the *approximation condition*. In the following, we discuss the values of the constants a_m and b_m for specific smoothers and projectors. For the smoothers, the discussion will be completely general, independent of the matrix A_m (see Lemma 1), while for the projectors we will restrict our attention to matrices $A_m = A_{\boldsymbol{m}}$ (of size $m = m_1\cdots m_d$) that 'majorize', in the sense of (13), the τ-matrix $\tau_{\boldsymbol{m}}(z_d)$ generated by the trigonometric polynomial (2).

When using the Richardson iteration, the smoothing condition can be easily satisfied and the next lemma can be proved in the same way as [20, Theorem 4.4] (with $D = I$ and $Q = I/\omega$).

Lemma 1 *Let $A_m \in \mathbb{R}^{m\times m}$ be SPD, let $S_m := I - \omega A_m$ ($\omega \in \mathbb{R}$), and let $\mu_m \ge \rho(A_m)$. If $0 < \omega < 2/\mu_m$, then the smoothing condition (a) in Theorem 2 holds with $a_m := \omega(2 - \omega\mu_m) > 0$ and moreover $\rho(S_m) < 1$.*

2.4 Multigrid methods for τ-matrices

The definition of multigrid methods for τ-matrices requires a proper choice of the grid-transfer operators, in order to guarantee fast convergence speed and to preserve the same structure of the matrices at the coarser levels [1,2,12,13].

We now define our grid-transfer operator (or projector) $P_{\boldsymbol{m}}$ for multi-indices $\boldsymbol{m} \in \mathbb{N}^d$ satisfying certain additional constraints. For any odd $m \geq 3$, let us denote by U_m the cutting matrix of size $\frac{m-1}{2} \times m$ given by

$$U_m := \begin{bmatrix} 0 & 1 & & & & & 0 \\ & & 0 & 1 & & & 0 \\ & & & & \ddots & & \vdots \\ & & & & & 0\;1 & 0 \end{bmatrix} \in \mathbb{R}^{\frac{m-1}{2} \times m}.$$

For any $\boldsymbol{m} \in \mathbb{N}^d$ with odd $m_1, \ldots, m_d \geq 3$, we define $U_{\boldsymbol{m}} := U_{m_1} \otimes \cdots \otimes U_{m_d}$. Then, we set

$$P_{\boldsymbol{m}} := U_{\boldsymbol{m}}\, \tau_{\boldsymbol{m}}(q_d), \tag{9}$$

with

$$q_d(\theta_1, \ldots, \theta_d) := \prod_{j=1}^{d} (1 + \cos\theta_j).$$

By the properties of τ-matrices and Kronecker tensor-products, we have

$$P_{\boldsymbol{m}} = \bigotimes_{j=1}^{d} U_{m_j}\, \tau_{m_j}(1 + \cos\theta_j),$$

and

$$P_{\boldsymbol{m}} = \bigotimes_{j=1}^{d} \frac{1}{2} \underbrace{\begin{bmatrix} 1 & 2 & 1 & & & & \\ & & 1 & 2 & 1 & & \\ & & & & \ddots & & \\ & & & & & 1\;2 & 1 \end{bmatrix}}_{m_j}.$$

The matrix $P_{\boldsymbol{m}}$ has full rank $M := \prod_{j=1}^{d} \frac{m_j - 1}{2}$ and is the standard restriction operator or the so-called full-weighting projector. Its transpose $P_{\boldsymbol{m}}^T$ is the traditional linear interpolation operator.

Let z_d be defined as in (2). Note that z_d is a linear nonnegative d-variate cosine trigonometric polynomial with a unique zero at $(0, \ldots, 0)$ over $[0, \pi]^d$. The next lemma (Lemma 2) addresses the approximation condition in Theorem 2 when $A_m = A_{\boldsymbol{m}}$ is the d-level τ-matrix $\tau_{\boldsymbol{m}}(z_d)$ (of size $m = m_1 \cdots m_d$) and $P_m = P_{\boldsymbol{m}}$ is the projector in (9). The lemma is a direct consequence of [22, Lemma 8.2] thanks to the following two properties of q_d and z_d: given the set of mirror points of $\boldsymbol{\theta} := (\theta_1, \ldots, \theta_d)$ as defined in [22, p. 454], namely

$$\mathcal{M}(\boldsymbol{\theta}) := \left\{ \widehat{\boldsymbol{\theta}} := (\widehat{\theta}_1, \ldots, \widehat{\theta}_d) \in [0, \pi]^d : \widehat{\theta}_i \in \{\theta_i, \pi - \theta_i\},\ \forall i = 1, \ldots, d \right\} \setminus \{\boldsymbol{\theta}\},$$

we have[2]

$$\sum_{\widehat{\boldsymbol{\theta}} \in \mathcal{M}(\boldsymbol{\theta}) \cup \{\boldsymbol{\theta}\}} q_d^2(\widehat{\boldsymbol{\theta}}) > 0, \quad \forall \boldsymbol{\theta} \in [0, \pi]^d, \tag{10}$$

$$\limsup_{\boldsymbol{\theta} \to \mathbf{0}} \max_{\widehat{\boldsymbol{\theta}} \in \mathcal{M}(\boldsymbol{\theta})} \frac{q_d^2(\widehat{\boldsymbol{\theta}})}{z_d(\boldsymbol{\theta})} < \infty. \tag{11}$$

Lemma 2 ([22]) *For $\boldsymbol{m} \in \mathbb{N}^d$ with odd $m_1, \ldots, m_d \geq 3$, let $A_{\boldsymbol{m}} = \tau_{\boldsymbol{m}}(z_d)$ and let $P_{\boldsymbol{m}}$ be the full-rank projector given by (9). Then, the matrix $A_{\boldsymbol{m}}$ is SPD and the approximation condition (b) in Theorem 2 holds with a constant depending only on d, i.e.,*

$$\exists\, \widetilde{b}_d > 0 : \min_{\mathbf{y} \in \mathbb{R}^M} \|\mathbf{x} - P_{\boldsymbol{m}}^T \mathbf{y}\|_2^2 \leq \widetilde{b}_d \|\mathbf{x}\|_{A_{\boldsymbol{m}}}^2, \tag{12}$$

for all $\mathbf{x} \in \mathbb{R}^{m_1 \cdots m_d}$. Moreover, if $d = 1$ then (12) holds with $\widetilde{b}_1 = 1/2$.

The specific value $\widetilde{b}_1$ has been found by looking carefully at the proof of [22, Lemma 3.2].

From Lemma 2 we deduce the result in Lemma 3. Given $X, Y \in \mathbb{C}^{m \times m}$, we write $X \leq Y$ if and only if X, Y are both Hermitian and $Y - X$ is nonnegative definite.

Lemma 3 *For $\boldsymbol{m} \in \mathbb{N}^d$ with odd $m_1, \ldots, m_d \geq 3$, let $A_{\boldsymbol{m}} \in \mathbb{R}^{(m_1 \cdots m_d) \times (m_1 \cdots m_d)}$ be SPD and let $P_{\boldsymbol{m}}$ be given by (9). Let $\delta_{\boldsymbol{m}} > 0$ such that*

$$A_{\boldsymbol{m}} \geq \delta_{\boldsymbol{m}}\, \tau_{\boldsymbol{m}}(z_d). \tag{13}$$

Then, the approximation condition (b) in Theorem 2 holds, i.e.,

$$\exists\, b_{\boldsymbol{m},d} := \frac{\widetilde{b}_d}{\delta_{\boldsymbol{m}}} > 0 : \min_{\mathbf{y} \in \mathbb{R}^M} \|\mathbf{x} - P_{\boldsymbol{m}}^T \mathbf{y}\|_2^2 \leq b_{\boldsymbol{m},d} \|\mathbf{x}\|_{A_{\boldsymbol{m}}}^2,$$

for all $\mathbf{x} \in \mathbb{R}^{m_1 \cdots m_d}$, where $\widetilde{b}_d$ is defined in Lemma 2.

Proof We use the same monotonicity argument as in [22, proof of Lemmas 4.2 and 9.2]. Assuming (13), we have

$$\|\mathbf{x}\|_{\tau_{\boldsymbol{m}}(z_d)}^2 = \mathbf{x}^T \tau_{\boldsymbol{m}}(z_d) \mathbf{x} \leq \frac{1}{\delta_{\boldsymbol{m}}} \mathbf{x}^T A_{\boldsymbol{m}} \mathbf{x} = \frac{1}{\delta_{\boldsymbol{m}}} \|\mathbf{x}\|_{A_{\boldsymbol{m}}}^2,$$

[2] The first property holds because q_d is nonnegative and, by a direct computation,

$\sum_{\widehat{\boldsymbol{\theta}} \in \mathcal{M}(\boldsymbol{\theta}) \cup \{\boldsymbol{\theta}\}} q_d(\widehat{\boldsymbol{\theta}}) = 2^d > 0, \quad \forall \boldsymbol{\theta} \in [0, \pi]^d.$

for all $\mathbf{x} \in \mathbb{R}^{m_1 \cdots m_d}$. By Lemma 2 we get

$$\min_{\mathbf{y}\in\mathbb{R}^M} \|\mathbf{x} - P_{\boldsymbol{m}}^T \mathbf{y}\|_2^2 \le \widetilde{b}_d \|\mathbf{x}\|^2_{\tau_{\boldsymbol{m}}(z_d)} \le \frac{\widetilde{b}_d}{\delta_{\boldsymbol{m}}} \|\mathbf{x}\|^2_{A_{\boldsymbol{m}}},$$

for all $\mathbf{x} \in \mathbb{R}^{m_1 \cdots m_d}$, which completes the proof. □

The next corollary follows immediately from Theorem 2 in combination with Lemmas 1 and 3.

Corollary 1 *Let $\mathcal{I}$ be a set of multi-indices such that $\mathcal{I} \subseteq \{\boldsymbol{m} \in \mathbb{N}^d : m_1, \ldots, m_d \ge 3 \text{ odd}\}$. $\forall \boldsymbol{m} \in \mathcal{I}$, let $A_{\boldsymbol{m}} \in \mathbb{R}^{(m_1\cdots m_d)\times(m_1\cdots m_d)}$ be SPD, let $S_{\boldsymbol{m}} := I - \omega A_{\boldsymbol{m}}$ and let $P_{\boldsymbol{m}} := U_{\boldsymbol{m}}\, \tau_{\boldsymbol{m}}(q_d)$. Assume that $\mu := \sup_{\boldsymbol{m}\in\mathcal{I}} \rho(A_{\boldsymbol{m}}) < \infty$ and that the inequality (13) holds with $\delta := \inf_{\boldsymbol{m}\in\mathcal{I}} \delta_{\boldsymbol{m}} > 0$, and take $0 < \omega < 2/\mu$. Then,*

$$\rho(TG(S_{\boldsymbol{m}}, P_{\boldsymbol{m}})) \le \sqrt{1 - \frac{a\,\delta}{\widetilde{b}_d}}, \quad \forall \boldsymbol{m} \in \mathcal{I},$$

where $a := \omega(2 - \omega\mu)$ and $\widetilde{b}_d$ is defined in Lemma 2.

In the case where $A_{\boldsymbol{m}} = \tau_{\boldsymbol{m}}(z_d)$ and the projector $P_{\boldsymbol{m}}$ is taken as in (9), the coarser matrix $P_{\boldsymbol{m}} A_{\boldsymbol{m}} P_{\boldsymbol{m}}^T$ is again a τ-matrix generated by z_d up to a multiplicative constant. More generally, in a multigrid perspective, if we fix multi-indices $\boldsymbol{m}_0 := \boldsymbol{m} > \boldsymbol{m}_1 > \boldsymbol{m}_2 > \cdots > \boldsymbol{m}_l > 0$, where the inequalities are componentwise; if we take at each multigrid level $i = 0, \ldots, l-1$ the projector $P_{\boldsymbol{m}_i}$, as given by (9) for $\boldsymbol{m} = \boldsymbol{m}_i$; and if we define the coefficient matrix at level $i+1$ as $A_{\boldsymbol{m}_{i+1}} := P_{\boldsymbol{m}_i} A_{\boldsymbol{m}_i} P_{\boldsymbol{m}_i}^T$ for $i = 0, \ldots, l-1$; then, from the results in [22] or by direct computation, we have $A_{\boldsymbol{m}_i} = \tau_{\boldsymbol{m}_i}(r_i z_d)$ for all $i = 0, \ldots, l$, where r_i is a positive constant. This observation shows that the τ structure is preserved on the coarser levels, which is fundamental for the construction of multigrid algorithms with more than one recursion level.

In this regard, we remark that the conditions (10)–(11) are not sufficient to obtain the V-cycle optimality; see [2]. Nevertheless, taking into account the properties of q_d and z_d, we see that condition (11) can be replaced by the following stronger version:

$$\limsup_{\boldsymbol{\theta}\to\mathbf{0}} \max_{\widehat{\boldsymbol{\theta}}\in\mathcal{M}(\boldsymbol{\theta})} \frac{q_d(\widehat{\boldsymbol{\theta}})}{z_d(\boldsymbol{\theta})} < \infty. \tag{14}$$

According to the analysis in [1], conditions (10) and (14) lead to the V-cycle optimality, when the V-cycle algorithm is applied to the τ-matrix $A_{\boldsymbol{m}} = \tau_{\boldsymbol{m}}(z_d)$. Unfortunately, Lemma 3 does not suffice to extend the optimality proof provided in [1,2] for τ-matrices (and matrix algebras in general) to more general matrix structures like our matrices in the IgA context. Such a proof will be a (difficult) task to be addressed in future work.

3 Galerkin discretization using B-splines

We now detail the Galerkin discretization based on uniform B-splines of our model problem (1), and we devote special attention to the symbol which describes the spectral behavior of the discretization matrices.

3.1 The 1D setting

In this section we focus on our model problem for $d = 1$:

$$\begin{cases} -u'' + \gamma u = \mathrm{f}, & \text{in } (0,1), \\ u(0) = 0, \quad u(1) = 0, \end{cases} \tag{15}$$

with $\mathrm{f} \in L^2(0,1)$ and $\gamma \ge 0$. We approximate the (weak) solution u of (15) in the space $\mathcal{W}$ of polynomial splines with maximal smoothness represented in the B-spline basis. More precisely, for $p \ge 1$ and $n \ge 2$, let

$$\mathcal{V}_n^{[p]} := \left\{ s \in C^{p-1}([0,1]) : s_{\big|\left[\frac{i}{n}, \frac{i+1}{n}\right)} \in \mathbb{P}_p,\ 0 \le i < n \right\},$$
$$\mathcal{W}_n^{[p]} := \left\{ s \in \mathcal{V}_n^{[p]} : s(0) = s(1) = 0 \right\} \subset H_0^1(0,1).$$

It is known that $\dim \mathcal{V}_n^{[p]} = n + p$ and $\dim \mathcal{W}_n^{[p]} = n + p - 2$. In the Galerkin method (5) we choose $\mathcal{W} = \mathcal{W}_n^{[p]}$, for some $p \ge 1$ and $n \ge 2$, and for $\mathcal{W}_n^{[p]}$ we choose the B-spline basis $\{N_{2,[p]}, \ldots, N_{n+p-1,[p]}\}$, which is defined as follows (see [6] or [16]).

Definition 2 Consider the knot sequence

$$t_1 = \cdots = t_{p+1} = 0 < t_{p+2} < \cdots$$
$$\cdots < t_{p+n} < 1 = t_{p+n+1} = \cdots = t_{2p+n+1}, \tag{16}$$

with $t_{p+i+1} := i/n$, $i = 0, \ldots, n$. The B-splines $N_{i,[p]} : [0,1] \to \mathbb{R}$, $i = 1, \ldots, n+p$, are defined recursively over the knot sequence (16) as follows: for $1 \le i \le n + 2p$,

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

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

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

where we assume that a fraction with zero denominator is zero.

With these choices of the approximation space $\mathcal{W}_n^{[p]}$ and of the basis functions $\{N_{2,[p]}, \ldots, N_{n+p-1,[p]}\}$, we obtain in

(6) the $(n+p-2)\times(n+p-2)$ stiffness matrix $A=A_n^{[p]}$ given by

$$A_n^{[p]}=\left[a(N_{j,[p]},N_{i,[p]})\right]_{i,j=2}^{n+p-1}=nK_n^{[p]}+\frac{\gamma}{n}M_n^{[p]},\quad (17)$$

where

$$nK_n^{[p]}:=\left[\int_{(0,1)}N'_{j,[p]}N'_{i,[p]}\right]_{i,j=2}^{n+p-1},\quad (18)$$

$$\frac{1}{n}M_n^{[p]}:=\left[\int_{(0,1)}N_{j,[p]}N_{i,[p]}\right]_{i,j=2}^{n+p-1}.\quad (19)$$

The above matrices have the following properties.

Lemma 4 [16] *For every $p\geq 1$ and $n\geq 2$,*

- *$K_n^{[p]}$ is SPD and $\|K_n^{[p]}\|_\infty\leq 4p$;*
- *$M_n^{[p]}$ is SPD, $\|M_n^{[p]}\|_\infty\leq 1$ and $\exists\, C^{[p]}>0$, depending only on p, such that $\lambda_{\min}(M_n^{[p]})>C^{[p]}$.*

We now recall the spectral properties of the sequence $\{\frac{1}{n}A_n^{[p]}\}_n$ obtained in [16]. For $p\geq 0$, 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:

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

and

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

We point out that the 'central' basis functions $N_{i,[p]}(x)$, $i=p+1,\ldots,n$, are uniformly shifted and scaled versions of the cardinal B-spline $\phi_{[p]}$, because we have

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

Let us denote by $\ddot{\phi}_{[p]}(t)$ the second derivative of $\phi_{[p]}(t)$ with respect to its argument t (for $p\geq 3$). For $p\geq 0$, let $h_p:[-\pi,\pi]\to\mathbb{R}$,

$$h_p(\theta):=\phi_{[2p+1]}(p+1)+2\sum_{k=1}^{p}\phi_{[2p+1]}(p+1-k)\cos(k\theta).\quad (20)$$

In particular, we have $h_0(\theta)=1$. For $p\geq 1$, let $f_p:[-\pi,\pi]\to\mathbb{R}$,

$$f_p(\theta):=-\ddot{\phi}_{[2p+1]}(p+1)-2\sum_{k=1}^{p}\ddot{\phi}_{[2p+1]}(p+1-k)\cos(k\theta).\quad (21)$$

Let $m_n^{[p]}:=n+p-2$ and fix $p\geq 1$. It has been proved in [16, Theorem 12] that, $\forall F\in C_c(\mathbb{R})$,

$$\lim_{n\to\infty}\frac{1}{m_n^{[p]}}\sum_{j=1}^{m_n^{[p]}}F\left(\lambda_j\left(\frac{1}{n}A_n^{[p]}\right)\right)=\frac{1}{2\pi}\int_{-\pi}^{\pi}F(f_p(\theta))\,\mathrm{d}\theta.\quad (22)$$

Due to this limit relation, f_p is called the symbol of the sequence of matrices $\{\frac{1}{n}A_n^{[p]}\}_n$. Note that f_p is symmetric on $[-\pi,\pi]$, so it is also the symbol of $\{\tau_{n+p-2}(f_p)\}_n$, meaning that (22) holds with $\tau_{n+p-2}(f_p)$ instead of $\frac{1}{n}A_n^{[p]}$.

Remark 1 The symbol f_p is independent of γ. Moreover, when modifying our problem (15) by adding an advection term $\beta u'$, the resulting symbol is again f_p, independent of β and γ; see [16]. The independence of f_p from the advection/reaction terms is not a surprise, as it is known in the literature (see, e.g., [17,23,24]) that the symbol has a canonical structure in which only the coefficient of the higher-order operator is present (in our specific case, the higher-order operator is the Laplacian). If $|\beta|$ is very large, however, the fineness parameter n has to be chosen extremely large in order to dampen the effects of the advection term. In such cases, the theoretical spectral distribution (22) is reasonably attained only for very large n. This implies that, for small values of n, the advection term affects significantly the spectrum of $\frac{1}{n}A_n^{[p]}$ (and also the performance of multigrid solvers like the one presented in this paper).

Lemma 5 ([16]) *For all $p\geq 1$ and $\theta\in[-\pi,\pi]$,*

$$f_p(\theta)=z_1(\theta)h_{p-1}(\theta),$$

with $z_1(\theta)=2-2\cos\theta$ as defined by (2) for $d=1$, and h_p as defined by (20). Moreover,

$$\left(\frac{4}{\pi^2}\right)^p\leq h_{p-1}(\theta)\leq h_{p-1}(0)=1,$$

and f_p has a unique zero of order two at $\theta=0$ (like the function z_1).

The properties in Lemma 5 have been proved in [16, Lemma 7 and Remark 2] for $p\geq 2$, but they can be easily checked for $p=1$ by direct computation. Lemma 5 immediately provides two constants c_p and C_p for the inequalities (3) in the case $d=1$. Figure 1 shows the graph of f_p normalized by its maximum M_{f_p} for $p=1,\ldots,5$.

We conclude this section by pointing out that h_p is the symbol of the sequence of matrices $\{M_n^{[p]}\}_n$: $\forall F\in C_c(\mathbb{C})$,

$$\lim_{n\to\infty}\frac{1}{m_n^{[p]}}\sum_{j=1}^{m_n^{[p]}}F(\lambda_j(M_n^{[p]}))=\frac{1}{2\pi}\int_{-\pi}^{\pi}F(h_p(\theta))\,\mathrm{d}\theta.$$

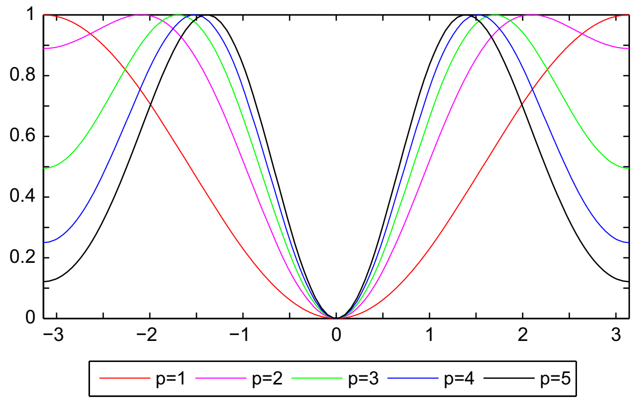

Fig. 1 Graph of f_p/M_{f_p} for $p = 1, \dots, 5$

We omit the formal proof of this result; it can be proved using the same arguments as [16, Theorem 12].

3.2 The 2D setting

In this section we focus on our model problem (1) in the case $d = 2$, and we follow the same scheme as the 1D setting. Given any two functions $f, g : [a, b] \to \mathbb{R}$, we denote by $f \otimes g$ the tensor-product function

$$f \otimes g : [a, b]^2 \to \mathbb{R}, \quad (f \otimes g)(x, y) := f(x)g(y).$$

We approximate the weak solution u of (1) by means of tensor-product B-splines. More precisely, we choose $\mathcal{W} = \mathcal{W}^{[p_1,p_2]}_{n_1,n_2}$, for some $p_1, p_2 \geq 1$, $n_1, n_2 \geq 2$, where

$$\mathcal{W}^{[p_1,p_2]}_{n_1,n_2} := \langle N_{j_1,[p_1]} \otimes N_{j_2,[p_2]} : \\ j_1 = 2, \dots, n_1 + p_1 - 1,\ j_2 = 2, \dots, n_2 + p_2 - 1\rangle,$$

and $N_{j,[p]}$, $j = 2, \dots, n + p - 1$, are the basis functions considered in Sect. 3.1 (see Definition 2). We order the tensor-product B-spline basis $\{N_{j_1,[p_1]} \otimes N_{j_2,[p_2]} : j_1 = 2, \dots, n_1+p_1-1,\ j_2 = 2, \dots, n_2+p_2-1\}$ in the following way:

$$\left[\left[N_{j_1,[p_1]} \otimes N_{j_2,[p_2]}\right]_{j_1=2,\dots,n_1+p_1-1}\right]_{j_2=2,\dots,n_2+p_2-1}.$$

With these choices of the approximation space and of the basis functions, we obtain in (6) the stiffness matrix $A = A^{[p_1,p_2]}_{n_1,n_2}$ given by

$$A^{[p_1,p_2]}_{n_1,n_2} := K^{[p_1,p_2]}_{n_1,n_2} + \frac{\gamma}{n_1 n_2} M^{[p_2]}_{n_2} \otimes M^{[p_1]}_{n_1}, \tag{23}$$

where

$$K^{[p_1,p_2]}_{n_1,n_2} := \frac{n_1}{n_2} M^{[p_2]}_{n_2} \otimes K^{[p_1]}_{n_1} + \frac{n_2}{n_1} K^{[p_2]}_{n_2} \otimes M^{[p_1]}_{n_1},$$

and the matrices $K^{[p]}_n$, $M^{[p]}_n$ are defined in (18)–(19).

Remark 2 By Lemma 4 and by the fact that $X \otimes Y$ is SPD whenever X, Y are SPD, we know that $A^{[p_1,p_2]}_{n_1,n_2}$ is SPD for all $p_1, p_2 \geq 1$ and $n_1, n_2 \geq 2$.

We now recall from [16] the spectral properties of the sequence $\{A^{[p_1,p_2]}_{\nu_1 n,\nu_2 n}\}_n$. For $p_1, p_2 \geq 1$ and $\nu_1, \nu_2 \in \mathbb{Q}_+ := \{r \in \mathbb{Q} : r > 0\}$, let $f^{(\nu_1,\nu_2)}_{p_1,p_2} : [-\pi, \pi]^2 \to \mathbb{R}$,

$$f^{(\nu_1,\nu_2)}_{p_1,p_2} := \frac{\nu_1}{\nu_2} h_{p_2} \otimes f_{p_1} + \frac{\nu_2}{\nu_1} f_{p_2} \otimes h_{p_1},$$

where h_p and f_p are given in (20)–(21). From now on we assume that $n \in \mathbb{N}$ is chosen so that $\nu_1 n, \nu_2 n \in \mathbb{N}$. Let us consider the sequence of matrices $\{A^{[p_1,p_2]}_{\nu_1 n,\nu_2 n}\}_n$. It was proved in [16, Section 5.2] that, $\forall F \in C_c(\mathbb{R})$,

$$\lim_{n\to\infty} \frac{1}{N_n} \sum_{j=1}^{N_n} F\left(\lambda_j\left(A^{[p_1,p_2]}_{\nu_1 n,\nu_2 n}\right)\right) \\ = \frac{1}{(2\pi)^2} \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} F(f^{(\nu_1,\nu_2)}_{p_1,p_2}(\theta_1, \theta_2))\, \mathrm{d}\theta_1 \mathrm{d}\theta_2, \tag{24}$$

where $N_n := (\nu_1 n + p_1 - 2)(\nu_2 n + p_2 - 2)$ is the size of $A^{[p_1,p_2]}_{\nu_1 n,\nu_2 n}$. Due to this limit relation, $f^{(\nu_1,\nu_2)}_{p_1,p_2}$ is called the symbol of the sequence $\{A^{[p_1,p_2]}_{\nu_1 n,\nu_2 n}\}_n$. Since we have $f^{(\nu_1,\nu_2)}_{p_1,p_2}(\pm\theta_1, \pm\theta_2) = f^{(\nu_1,\nu_2)}_{p_1,p_2}(\theta_1, \theta_2)$, the relation (24) continues to hold if $A^{[p_1,p_2]}_{\nu_1 n,\nu_2 n}$ were replaced by the two-level τ-matrix $\tau_{\nu_2 n+p_2-2,\nu_1 n+p_1-2}(f^{(\nu_1,\nu_2)}_{p_1,p_2})$. This means that $f^{(\nu_1,\nu_2)}_{p_1,p_2}$ is also the symbol of the sequence of τ-matrices $\{\tau_{\nu_2 n+p_2-2,\nu_1 n+p_1-2}(f^{(\nu_1,\nu_2)}_{p_1,p_2})\}_n$.

The symbol $f^{(\nu_1,\nu_2)}_{p_1,p_2}$ possesses the following properties, which are consequences of Lemma 5.

Lemma 6 *Let $p_1, p_2 \geq 1$ and $\nu_1, \nu_2 \in \mathbb{Q}_+$. Then, for all $(\theta_1, \theta_2) \in [-\pi, \pi]^2$,*

$$f^{(\nu_1,\nu_2)}_{p_1,p_2}(\theta_1, \theta_2) \geq \left(\frac{4}{\pi^2}\right)^{p_1+p_2+1} \min\left(\frac{\nu_2}{\nu_1}, \frac{\nu_1}{\nu_2}\right) z_2(\theta_1, \theta_2),$$

$$f^{(\nu_1,\nu_2)}_{p_1,p_2}(\theta_1, \theta_2) \leq \max\left(\frac{\nu_2}{\nu_1}, \frac{\nu_1}{\nu_2}\right) z_2(\theta_1, \theta_2),$$

with $z_2(\theta_1, \theta_2) = \sum_{j=1}^{2}(2 - 2\cos\theta_j)$ as defined by (2) for $d = 2$. In particular, $f^{(\nu_1,\nu_2)}_{p_1,p_2}$ has a unique zero at $(\theta_1, \theta_2) = (0, 0)$ of order two (like the function z_2).

Lemma 6 immediately provides two constants $c_{\boldsymbol{p}}$ and $C_{\boldsymbol{p}}$ for the inequalities (3) in the case $d = 2$.

4 TGM for Galerkin B-spline matrices: proof of optimality

In this section we prove that the standard TGM with relaxed Richardson smoother and full-weighting projector (described in Sects. 2.3 and 2.4) is optimal for linear systems with coefficient matrix $\frac{1}{n}A_n^{[p]}$ $(1 \leq p \leq 3)$ in the 1D case and $A_{\nu_1 n,\nu_2 n}^{[p_1,p_2]}$ $(1 \leq p_1, p_2 \leq 3)$ in the 2D case. The results can be easily extended to higher dimensionalities $d > 2$, by replicating the argument used in the 2D case.

4.1 Optimality of the TGM for $\frac{1}{n}A_n^{[p]}$

Let us start by giving an intuitive motivation based on the symbol f_p why the proposed TGM with relaxed Richardson smoother and full-weighting projector is expected to be optimal for $\frac{1}{n}A_n^{[p]}$. Because f_p is the symbol of both $\{\frac{1}{n}A_n^{[p]}\}_n$ and $\{\tau_{n+p-2}(f_p)\}_n$, these sequences of matrices share the same spectral distribution. The convergence properties of two-grid/multigrid methods strongly depend on the spectrum of the matrices to which they are applied. Therefore, it is reasonable to use for $\frac{1}{n}A_n^{[p]}$ the same TGM as proposed in [22] which was proved to be optimal for $\tau_{n+p-2}(f_p)$, and to expect that it is optimal also for $\frac{1}{n}A_n^{[p]}$.

Fix $p \geq 1$ and consider the sequence of matrices $\{\frac{1}{n}A_n^{[p]} : n \in \mathcal{I}_p\}$, with $\mathcal{I}_p \subseteq \{n \geq 2 : n+p-2 \geq 3 \text{ odd}\}$ an infinite set of indices. The requirement on the matrix size is due to our projector choice; see (25). We want to solve $\frac{1}{n}A_n^{[p]}\mathbf{u} = \mathbf{b}$, with $n \in \mathcal{I}_p$ and $\mathbf{b} \in \mathbb{R}^{n+p-2}$, by means of the TGM. As smoother we take the relaxed Richardson method with iteration matrix

$$S_n^{[p]} := I - \omega^{[p]}\frac{1}{n}A_n^{[p]},$$

where $\omega^{[p]} \in \mathbb{R}$ is a relaxation parameter chosen as a function of p and independent of n. The projector is taken to be

$$P_n^{[p]} := U_{n+p-2}\,\tau_{n+p-2}(1+\cos\theta), \tag{25}$$

as defined in (9) for $d = 1$ and $\boldsymbol{m} = n + p - 2$.

For the sake of simplicity, we now assume $\gamma = 0$, so $\frac{1}{n}A_n^{[p]} = K_n^{[p]}$. Under this assumption and under suitable conditions on the relaxation parameter $\omega^{[p]}$, we show that, for $p = 1, 2, 3$, the TGM with iteration matrix $TG(S_n^{[p]}, P_n^{[p]})$ is optimal, i.e., $\exists\, c_p < 1$ such that $\rho(TG(S_n^{[p]}, P_n^{[p]})) \leq c_p$ for all $n \in \mathcal{I}_p$. Corollary 1 can be reformulated in our 1D context as follows.

Corollary 2 *Assume that for fixed* $p \geq 1$,

$$\exists\, \delta^{[p]} > 0 : \ K_n^{[p]} \geq \delta^{[p]}\tau_{n+p-2}(2-2\cos\theta), \quad \forall n \geq 2, \tag{26}$$

and let $\mu^{[p]} := \sup_{n\in\mathcal{I}_p} \rho(K_n^{[p]})$. *Then, for any* $\omega^{[p]} \in (0, 2/\mu^{[p]})$ *it holds that*

$$\rho(TG(S_n^{[p]}, P_n^{[p]})) \leq \sqrt{1 - 2\,a^{[p]}\,\delta^{[p]}}, \quad \forall n \in \mathcal{I}_p,$$

where $a^{[p]} := \omega^{[p]}(2 - \omega^{[p]}\mu^{[p]})$.

From Lemma 4 we know that $\mu^{[p]} \leq 4p$ for all $p \geq 1$. We also have $\mu^{[1]} = 4$ and $\mu^{[2]} \leq 3/2 + (1 + \sqrt{2})/6$; see [16, Eq. (79)]. Moreover, from some numerical experiments it seems that $\mu^{[2]} = 3/2$ and $\mu^{[3]} \leq 1.80$.

In the next theorem we prove that (26) holds for $p = 1, 2, 3$.

Theorem 3 *For* $1 \leq p \leq 3$, *condition* (26) *is satisfied with* $\delta^{[1]} = 1$, $\delta^{[2]} = 1/3$ *and* $\delta^{[3]} = 28/465$. *Hence, for* $1 \leq p \leq 3$ *and for any* $\omega^{[p]} \in (0, 2/\mu^{[p]})$, $\exists\, c_p < 1$ *such that* $\rho(TG(S_n^{[p]}, P_n^{[p]})) \leq c_p$ *for all* $n \in \mathcal{I}_p$.

Proof Since $K_n^{[1]} = \tau_{n-1}(2 - 2\cos\theta)$ for any $n \geq 2$, it is obvious that (26) holds for $p = 1$ with $\delta^{[1]} = 1$.

In the case $p = 2$, for $n \geq 5$ we have

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

and one can check that the matrix $K_n^{[2]} - \delta\,\tau_n(2 - 2\cos\theta)$ is nonnegative definite for $\delta = 1/3$, thanks to the Gershgorin theorems [5]. As it can be directly verified that $K_n^{[2]} \geq (1/3)\tau_n(2 - 2\cos\theta)$ for $n = 2, \ldots, 4$, we conclude that (26) holds for $p = 2$ with $\delta^{[2]} = 1/3$.

In the case $p = 3$, for $n \geq 8$ we have

$$K_n^{[3]} = \frac{1}{240}\cdot \begin{bmatrix} 360 & 9 & -60 & -3 & & & & & & & \\ 9 & 162 & -8 & -47 & -2 & & & & & & \\ -60 & -8 & 160 & -30 & -48 & -2 & & & & & \\ -3 & -47 & -30 & 160 & -30 & -48 & -2 & & & & \\ & -2 & -48 & -30 & 160 & -30 & -48 & -2 & & & \\ & & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & & \\ & & & -2 & -48 & -30 & 160 & -30 & -48 & -2 & \\ & & & & -2 & -48 & -30 & 160 & -30 & -47 & -3 \\ & & & & & -2 & -48 & -30 & 160 & -8 & -60 \\ & & & & & & -2 & -47 & -8 & 162 & 9 \\ & & & & & & & -3 & -60 & 9 & 360 \end{bmatrix}.$$

Since

$$f_3(\theta) = (\cos^2\theta + 13\cos\theta + 16)(2 - 2\cos\theta)/30$$
$$\geq (2/15)(2 - 2\cos\theta), \qquad \forall\theta \in [-\pi, \pi],$$

we have $\tau_m(f_3) \geq (2/15)\tau_m(2 - 2\cos\theta)$, $\forall m \geq 1$. By the Gershgorin theorems we find that $K_n^{[3]} \geq \varepsilon\,\tau_{n+1}(f_3)$ for all $n \geq 8$ with $\varepsilon = 14/31$. As a consequence, the inequality $K_n^{[3]} \geq (28/465)\tau_{n+1}(2 - 2\cos\theta)$ holds for all $n \geq 8$. A direct verification shows that it also holds for $n = 2, \ldots, 7$. □

Remark 3 There are at least two reasons why condition (26) is likely to be satisfied for all $p \geq 1$.

- The condition would hold if we had $\tau_{n+p-2}(f_p)$ instead of $K_n^{[p]}$. Indeed, from Lemma 5 it follows that $f_p(\theta) \geq (4/\pi^2)^p\,(2 - 2\cos\theta)$, and this implies that $\tau_m(f_p) \geq (4/\pi^2)^p\,\tau_m(2 - 2\cos\theta)$, $\forall m \geq 1$. On the other hand, $K_n^{[p]}$ mimics $\tau_{n+p-2}(f_p)$, because these matrices share the same symbol f_p and they differ from each other only by a small-rank correction term [16].
- The matrices $K_n^{[p]}$ and $\tau_{n+p-2}(2 - 2\cos\theta)$ are both associated with particular approximations of the elliptic problem (15) in the case $\gamma = 0$.

Multiplying (26) by $(\tau_{n+p-2}(2 - 2\cos\theta))^{-1/2}$ on the left and the right, and observing that

$$(\tau_{n+p-2}(2 - 2\cos\theta))^{-1/2} K_n^{[p]} (\tau_{n+p-2}(2 - 2\cos\theta))^{-1/2}$$

is similar to

$$(\tau_{n+p-2}(2 - 2\cos\theta))^{-1} K_n^{[p]},$$

we obtain that (26) is equivalent to the following:

$$\exists\delta^{[p]} > 0:$$
$$\lambda_{\min}((\tau_{n+p-2}(2 - 2\cos\theta))^{-1} K_n^{[p]}) \geq \delta^{[p]}, \quad \forall n \geq 2. \tag{27}$$

The inequality (27) is certainly satisfied for $p = 1$ (with $\delta^{[1]} = 1$), and numerical experiments reveal that (27) is also satisfied for $p = 2, \ldots, 6$, with the best value $\delta^{[p],*} := \inf_{n\geq 2} \lambda_{\min}((\tau_{n+p-2}(2 - 2\cos\theta))^{-1} K_n^{[p]})$ given by $\delta^{[2],*} \approx 0.3333$, $\delta^{[3],*} \approx 0.1333$, $\delta^{[4],*} \approx 0.0537$, $\delta^{[5],*} \approx 0.0177$, $\delta^{[6],*} \approx 0.0054$. Note that the value $\delta^{[p]}$ obtained in Theorem 3 coincides with $\delta^{[p],*}$ not only for $p = 1$ but also for $p = 2$.

Remark 4 From a theoretical viewpoint, the normalized symbol $f_p(\theta)/M_{f_p}$ has a unique zero at $\theta = 0$. On the other hand, it has been proved in [11] that the value $f_p(\pi)/M_{f_p} \leq 2^{2-p}$ decreases exponentially to zero as $p \to \infty$. This means that, from a numerical viewpoint, for large p, the normalized symbol f_p/M_{f_p} possesses two zeros over $[0, \pi]$: one at $\theta = 0$ and another at the corresponding mirror point $\theta = \pi$. Because of this, we expect a slow (though optimal) convergence rate, when solving a linear system of the form $\frac{1}{n}A_n^{[p]}\mathbf{u} = \mathbf{b}$ for large p, by means of the TGM described above; see [11] for a rigorous analysis in the case of the τ-matrix $\tau_{n+p-2}(f_p)$. Possible ways to overcome this problem are the choice of a different size reduction at the lower level and/or adopting a multi-iterative strategy, involving a variation of the smoothers. Both approaches have been extensively numerically tested in [9] and the second one turned out to be more efficient, especially for large values of p and higher dimensionalities.

4.2 Optimality of the TGM for $A^{[p_1,p_2]}_{\nu_1 n,\nu_2 n}$

Throughout this section, we use the notation $m^{[p_j]}_{\nu_j n} := \nu_j n + p_j - 2$, $j = 1, 2$. The symbol-based motivation why the TGM with relaxed Richardson smoother and full-weighting projector is expected to be optimal for $A^{[p_1,p_2]}_{\nu_1 n,\nu_2 n}$ is essentially the same as in the 1D case: if this TGM were applied to the τ-matrix $\tau_{m^{[p_2]}_{\nu_2 n}, m^{[p_1]}_{\nu_1 n}}(f^{(\nu_1,\nu_2)}_{p_1,p_2})$, which has the same symbol $f^{(\nu_1,\nu_2)}_{p_1,p_2}$ as $A^{[p_1,p_2]}_{\nu_1 n,\nu_2 n}$, then it would be optimal by the results in [22].

Fix $p_1, p_2 \geq 1$, $\nu_1, \nu_2 \in \mathbb{Q}_+$, and consider the sequence of matrices $\{A^{[p_1,p_2]}_{\nu_1 n,\nu_2 n} : n \in \mathcal{I}^{(\nu_1,\nu_2)}_{p_1,p_2}\}$, with

$$\mathcal{I}^{(\nu_1,\nu_2)}_{p_1,p_2} \subseteq \Big\{n : \nu_1 n \geq 2,\ \nu_2 n \geq 2,\ m^{[p_1]}_{\nu_1 n} \geq 3 \text{ odd},$$
$$m^{[p_2]}_{\nu_2 n} \geq 3 \text{ odd}\Big\}, \qquad \#\mathcal{I}^{(\nu_1,\nu_2)}_{p_1,p_2} = \infty.$$

The requirements on $m^{[p_1]}_{\nu_1 n}$ and $m^{[p_2]}_{\nu_2 n}$ are due to our projector choice; see (28). We want to solve $A^{[p_1,p_2]}_{\nu_1 n,\nu_2 n}\mathbf{u} = \mathbf{b}$, with $n \in \mathcal{I}^{(\nu_1,\nu_2)}_{p_1,p_2}$ and $\mathbf{b} \in \mathbb{R}^{m^{[p_1]}_{\nu_1 n} m^{[p_2]}_{\nu_2 n}}$, by means of the TGM. Like in the 1D setting, we take the relaxed Richardson iteration as smoother,

$$S^{[p_1,p_2]}_{\nu_1 n,\nu_2 n} := I - \omega^{[p_1,p_2,\nu_1,\nu_2]} A^{[p_1,p_2]}_{\nu_1 n,\nu_2 n},$$

where $\omega^{[p_1,p_2,\nu_1,\nu_2]}$ is the relaxation parameter (independent of n). The projector is taken to be

$$P^{[p_1,p_2]}_{\nu_1 n,\nu_2 n} := U_{m^{[p_2]}_{\nu_2 n}, m^{[p_1]}_{\nu_1 n}}\,\tau_{m^{[p_2]}_{\nu_2 n}, m^{[p_1]}_{\nu_1 n}}(q_2), \tag{28}$$

as defined in (9) for $d = 2$ and $\boldsymbol{m} = (m^{[p_2]}_{\nu_2 n}, m^{[p_1]}_{\nu_1 n})$.

For the sake of simplicity, we now assume $\gamma = 0$, so $A^{[p_1,p_2]}_{\nu_1 n,\nu_2 n} = K^{[p_1,p_2]}_{\nu_1 n,\nu_2 n}$. In the bilinear case $p_1 = p_2 = 1$, it can be shown that

$$K^{[1,1]}_{\nu_1 n,\nu_2 n} = \tau_{\nu_2 n-1,\nu_1 n-1}(f^{(\nu_1,\nu_2)}_{1,1}),$$

and the eigenvalues of $K^{[1,1]}_{\nu_1 n,\nu_2 n}$ are given by

$$f^{(\nu_1,\nu_2)}_{1,1}\left(\frac{j_2\pi}{\nu_2 n}, \frac{j_1\pi}{\nu_1 n}\right),$$

for $j_2 = 1,\ldots,m^{[1]}_{\nu_2 n}$, $j_1 = 1,\ldots,m^{[1]}_{\nu_1 n}$, and

$$\begin{aligned}\mu^{[1,1,\nu_1,\nu_2]} &:= \sup_{n\in\mathcal{I}^{(\nu_1,\nu_2)}_{1,1}} \rho(K^{[1,1]}_{\nu_1 n,\nu_2 n}) = \lim_{n\to\infty} \rho(K^{[1,1]}_{\nu_1 n,\nu_2 n}) \\ &= \max_{(\theta_1,\theta_2)\in[0,\pi]^2} f^{(\nu_1,\nu_2)}_{1,1}(\theta_1,\theta_2) = 4\max\left(\frac{\nu_1}{\nu_2}, \frac{\nu_2}{\nu_1}\right).\end{aligned}$$

In this case, for any $\omega^{[1,1,\nu_1,\nu_2]} \in \left(0, 2/\mu^{[1,1,\nu_1,\nu_2]}\right)$ and $\nu_1,\nu_2 \in \mathbb{Q}_+$, the optimality of the TGM with iteration matrix $TG(S^{[1,1]}_{\nu_1 n,\nu_2 n}, P^{[1,1]}_{\nu_1 n,\nu_2 n})$ was proved in [22] and the optimality of the related V-cycle in [1].

More generally, for $1 \le p_1, p_2 \le 3$ and $\nu_1,\nu_2 \in \mathbb{Q}_+$, we are going to show that the TGM with iteration matrix $TG(S^{[p_1,p_2]}_{\nu_1 n,\nu_2 n}, P^{[p_1,p_2]}_{\nu_1 n,\nu_2 n})$ is optimal under the assumption $\omega^{[p_1,p_2,\nu_1,\nu_2]} \in (0, 2/\mu^{[p_1,p_2,\nu_1,\nu_2]})$, with $\mu^{[p_1,p_2,\nu_1,\nu_2]} := \sup_{n\in\mathcal{I}^{(\nu_1,\nu_2)}_{p_1,p_2}} \rho(K^{[p_1,p_2]}_{\nu_1 n,\nu_2 n})$. From Lemma 4 we know that, $\forall n \ge 2$,

$$\begin{aligned}&\rho(K^{[p_1,p_2]}_{\nu_1 n,\nu_2 n}) \\ &\quad = \|K^{[p_1,p_2]}_{\nu_1 n,\nu_2 n}\|_2 \\ &\quad \le \frac{\nu_1}{\nu_2}\|M^{[p_2]}_{\nu_2 n}\|_2\|K^{[p_1]}_{\nu_1 n}\|_2 + \frac{\nu_2}{\nu_1}\|K^{[p_2]}_{\nu_2 n}\|_2\|M^{[p_1]}_{\nu_1 n}\|_2 \\ &\quad \le \frac{\nu_1}{\nu_2}\|M^{[p_2]}_{\nu_2 n}\|_\infty\|K^{[p_1]}_{\nu_1 n}\|_\infty + \frac{\nu_2}{\nu_1}\|K^{[p_2]}_{\nu_2 n}\|_\infty\|M^{[p_1]}_{\nu_1 n}\|_\infty \\ &\quad \le \frac{4p_1\nu_1}{\nu_2} + \frac{4p_2\nu_2}{\nu_1},\end{aligned}$$

where we used the fact that if X, Y are normal matrices then $\|X\otimes Y\|_2 = \|X\|_2\|Y\|_2$ and $\|X\|_2 \le \|X\|_\infty$.

In our 2D context, condition (13) reads as

$$\begin{aligned}\exists\,\delta^{[p_1,p_2,\nu_1,\nu_2]} > 0 :\ K^{[p_1,p_2]}_{\nu_1 n,\nu_2 n} \ge \delta^{[p_1,p_2,\nu_1,\nu_2]}\cdot \\ \tau_{m^{[p_2]}_{\nu_2 n},m^{[p_1]}_{\nu_1 n}}(4 - 2\cos\theta_1 - 2\cos\theta_2).\end{aligned} \qquad (29)$$

In the next theorem we show that (29) holds for $1 \le p_1, p_2 \le 3$, yielding the optimality of the TGM with iteration matrix $TG(S^{[p_1,p_2]}_{\nu_1 n,\nu_2 n}, P^{[p_1,p_2]}_{\nu_1 n,\nu_2 n})$ for these values of p_1 and p_2.

Theorem 4 *Let $1 \le p_1, p_2 \le 3$. Then, (29) holds with*

$$\delta^{[p_1,p_2,\nu_1,\nu_2]} = \min\left(\frac{\nu_1}{\nu_2}C^{[p_2]}\delta^{[p_1]}, \frac{\nu_2}{\nu_1}C^{[p_1]}\delta^{[p_2]}\right),$$

where $C^{[p]}$ is given in Lemma 4 and $\delta^{[p]}$ is specified in Theorem 3 for $1 \le p \le 3$. Hence, the TGM with iteration matrix $TG(S^{[p_1,p_2]}_{\nu_1 n,\nu_2 n}, P^{[p_1,p_2]}_{\nu_1 n,\nu_2 n})$ is optimal for $1 \le p_1, p_2 \le 3$ and for any $\omega^{[p_1,p_2,\nu_1,\nu_2]} \in (0, 2/\mu^{[p_1,p_2,\nu_1,\nu_2]})$.

Proof Recall that if X, X', Y, Y' are SPD with $X \ge X'$ and $Y \ge Y'$, then $X\otimes Y$ and $X'\otimes Y'$ are SPD with $X\otimes Y \ge X'\otimes Y'$. Hence, for every $\nu_1 n, \nu_2 n \ge 2$ integer, from Theorem 3 and the properties of τ-matrices we deduce that

$$\begin{aligned}&K^{[p_1,p_2]}_{\nu_1 n,\nu_2 n} \\ &\quad = \frac{\nu_1}{\nu_2}M^{[p_2]}_{\nu_2 n}\otimes K^{[p_1]}_{\nu_1 n} + \frac{\nu_2}{\nu_1}K^{[p_2]}_{\nu_2 n}\otimes M^{[p_1]}_{\nu_1 n} \\ &\quad \ge \frac{\nu_1}{\nu_2}C^{[p_2]}I_{m^{[p_2]}_{\nu_2 n}}\otimes \delta^{[p_1]}\tau_{m^{[p_1]}_{\nu_1 n}}(2-2\cos\theta_1) \\ &\qquad + \frac{\nu_2}{\nu_1}\,\delta^{[p_2]}\tau_{m^{[p_2]}_{\nu_2 n}}(2-2\cos\theta_2)\otimes C^{[p_1]}I_{m^{[p_1]}_{\nu_1 n}} \\ &\quad \ge \delta^{[p_1,p_2,\nu_1,\nu_2]}\,\tau_{m^{[p_2]}_{\nu_2 n},m^{[p_1]}_{\nu_1 n}}\left(\sum_{j=1}^{2}(2-2\cos\theta_j)\right).\end{aligned}$$

□

The proof of Theorem 4 can be extended in a straightforward way to higher dimensionalities $d > 2$.

Remark 5 The key result that allowed us to prove the optimality of the two-grid methods both in the 1D and 2D setting is the matrix inequality (26). If (26) were true for all $p \ge 1$, then it would be easy to give a proof of optimality for all $p \ge 1$ (in the 1D setting) and for all $p_1, p_2 \ge 1$ (in the 2D setting). Indeed, the former would be a direct consequence of Corollary 2, whereas the latter would follow by replicating the argument used in Theorem 4. We point out that the matrix inequality (26) for larger p could be handled by using a dyadic decomposition argument; see [4] and references therein or [25,26] for further insights on this subject.

Remark 6 The inequality (26) would be also of interest in the context of preconditioning related to the CG method and the GMRES method. Indeed, in the light of the Axelsson–Lindskog theorems [3], it can be shown that (26), which is equivalent to (27) by Remark 3, ensures that $\tau_{n+p-2}(2-2\cos\theta)$ is an optimal CG preconditioner for $K^{[p]}_n$. Hence, for $p = 1, 2, 3$, Theorem 3 ensures $\tau_{n+p-2}(2-2\cos\theta)$ to be an optimal CG preconditioner for $K^{[p]}_n$.

Remark 7 Let $M_{f^{(\nu_1,\nu_2)}_{p_1,p_2}} := \max_{\boldsymbol{\theta}\in[0,\pi]^2} f^{(\nu_1,\nu_2)}_{p_1,p_2}(\boldsymbol{\theta})$. It follows from Lemma 6 that the normalized symbol $f^{(\nu_1,\nu_2)}_{p_1,p_2}/M_{f^{(\nu_1,\nu_2)}_{p_1,p_2}}$ has a unique zero at $\boldsymbol{\theta} = \mathbf{0}$. However, from [11] we know that $f^{(\nu_1,\nu_2)}_{p_1,p_2}(\theta_1,\pi) \le 2^{2-p_1}M_{f^{(\nu_1,\nu_2)}_{p_1,p_2}}$ and $f^{(\nu_1,\nu_2)}_{p_1,p_2}(\pi,\theta_2) \le 2^{2-p_2}M_{f^{(\nu_1,\nu_2)}_{p_1,p_2}}$. Hence, when p_1, p_2 are large, the normalized symbol also has infinitely many small values (they can be seen as numerical zeros) over $[0,\pi]^2$, located at the edge points

$$\{(\theta_1,\pi) : 0 \le \theta_1 \le \pi\} \cup \{(\pi,\theta_2) : 0 \le \theta_2 \le \pi\}.$$

Due to this behavior, the TGM described above for the matrix $A_{\nu_1 n, \nu_2 n}^{[p_1, p_2]}$ is expected to show a bad (though optimal) convergence rate for large p_1, p_2. A possible way to overcome this problem has been proposed in [9,11] and consists in adopting a multi-iterative strategy involving a specialized PCG smoother.

5 Numerical examples

Numerical experiments addressing the pure Laplacian ($\gamma = 0$) can be found in [9, Sections 6.1 and 7.1]. They confirm the optimality of the TGM analyzed in Sect. 4. However, they also reveal that the spectral radii of the corresponding iteration matrices rapidly approach 1 for increasing $p \geq 2$ (resp., $p_1, p_2 \geq 2$). This poor behavior is related to the fact that $f_p(\pi)/M_{f_p}$ (resp., $f_{p_1,p_2}^{(\nu_1,\nu_2)}/M_{f_{p_1,p_2}^{(\nu_1,\nu_2)}}$) converges exponentially to 0 for increasing degree, as discussed in Remarks 4 and 7. Such a worsening was observed in [9] not only in the presence of the relaxed Richardson smoother, but also in the case of the relaxed Gauss-Seidel smoother; in particular, we refer the reader to [9, Table 4]. Actually, this worsening is an intrinsic feature of the problem that arises whenever a classical smoother is employed. So, the proposed TGM can be used only when dealing with small values of p (resp., p_1, p_2). Nevertheless, we point out that other techniques can be considered for large p (resp., p_1, p_2), as illustrated in [9,11].

In this section we show that the same conclusions also hold when addressing more general second-order elliptic problems, involving nonzero advection/reaction terms.

5.1 1D Examples

Table 1 shows the results of some numerical experiments for $TG(S_n^{[p]}, P_n^{[p]})$ applied to a system with coefficient matrix $\frac{1}{n}A_n^{[p]}$ and $\gamma = 1000$. The value of the parameter $\omega^{[p]}$ for the relaxed Richardson smoother $S_n^{[p]}$ was taken as in [9, Table 2]; it was determined in order to approximately minimize the asymptotic spectral radius in the pure Laplacian case ($\gamma = 0$). Then, we computed the spectral radii $\rho_n^{[p]} := \rho(TG(S_n^{[p]}, P_n^{[p]}))$ for $p = 1, \ldots, 6$ and increasing values of n. In all the considered experiments, the proposed TGM is optimal. Moreover, when $n \to \infty$, $\rho_n^{[p]}$ converges to a limit $\rho_\infty^{[p]}$, which is minimal for $p = 2$. We also observe that $\rho_\infty^{[p]}$ increases for increasing $p \geq 2$, in such a way that even for moderate values of p (such as $p = 5, 6$) the value $\rho_\infty^{[p]}$ is not really satisfactory.

To show that the numerical behavior observed for $TG(S_n^{[p]}, P_n^{[p]})$ is common to all classical smoothers, we perform the same test using as smoother the relaxed Gauss-Seidel method, whose iteration matrix is denoted by $\widehat{S}_n^{[p]}$. Table 2 illustrates the behavior of the spectral radius $\widehat{\rho}_n^{[p]} := \rho(TG(\widehat{S}_n^{[p]}, P_n^{[p]}))$. The relaxation parameter $\omega^{[p]}$ for $\widehat{S}_n^{[p]}$ was chosen as in [9, Table 3], which again approximately minimizes the asymptotic spectral radius in the pure Laplacian case ($\gamma = 0$). It follows from Table 2 that, except for the particular case $p = 2$, the use of the Gauss-Seidel smoother improves the convergence rate of the two-grid. However, we also observe that $\widehat{\rho}_n^{[p]}$ presents the same dependence on p as $\rho_n^{[p]}$: the scheme is optimal but its asymptotic convergence rate (if existing) attains its minimum for $p = 2$ and then worsens as p increases from 2 to 6.

Table 1 Values of $\rho_n^{[p]} := \rho(TG(S_n^{[p]}, P_n^{[p]}))$ in the case $\gamma = 1000$, for the specified parameter $\omega^{[p]}$

n	$\rho_n^{[1]}$ $[\omega^{[1]} = 1/3]$	$\rho_n^{[3]}$ $[\omega^{[3]} = 1.0368]$	$\rho_n^{[5]}$ $[\omega^{[5]} = 1.2576]$
80	0.3501889	0.5069841	0.9238604
160	0.3375447	0.4666843	0.9013911
320	0.3343860	0.4540155	0.8952219
640	0.3335965	0.4497422	0.8934305
1280	0.3333991	0.4481536	0.8928593
2560	0.3333498	0.4475009	0.8926547
n	$\rho_n^{[2]}$ $[\omega^{[2]} = 0.7311]$	$\rho_n^{[4]}$ $[\omega^{[4]} = 1.2229]$	$\rho_n^{[6]}$ $[\omega^{[6]} = 1.2235]$
81	0.0791465	0.7783691	0.9823774
161	0.0423859	0.7490167	0.9662849
321	0.0308670	0.7406819	0.9616686
641	0.0271624	0.7382560	0.9602275
1281	0.0259266	0.7374988	0.9597231
2561	0.0254774	0.7372365	0.9595247

Table 2 Values of $\widehat{\rho}_n^{[p]} := \rho(TG(\widehat{S}_n^{[p]}, P_n^{[p]}))$ in the case $\gamma = 1000$, for the specified parameter $\omega^{[p]}$

n	$\widehat{\rho}_n^{[1]}$ $[\omega^{[1]} = 0.9065]$	$\widehat{\rho}_n^{[3]}$ $[\omega^{[3]} = 0.9483]$	$\widehat{\rho}_n^{[5]}$ $[\omega^{[5]} = 1.1999]$
80	0.1649162	0.1497176	0.4269339
160	0.1946180	0.1324016	0.3803359
320	0.2151965	0.1392187	0.4151028
640	0.2295672	0.1495306	0.4440683
1280	0.2379982	0.1555952	0.4622802
2560	0.2425720	0.1587062	0.4721079
n	$\widehat{\rho}_n^{[2]}$ $[\omega^{[2]} = 0.9109]$	$\widehat{\rho}_n^{[4]}$ $[\omega^{[4]} = 1.0602]$	$\widehat{\rho}_n^{[6]}$ $[\omega^{[6]} = 1.3292]$
81	0.0565397	0.2831668	0.5866458
161	0.0561352	0.2613614	0.5249935
321	0.0593355	0.2849307	0.5494735
641	0.0618100	0.3051810	0.5818468
1281	0.0632679	0.3177150	0.6027278
2561	0.0640522	0.3247125	0.6128128

Table 3 Values of $\rho_n^{[p]} := \rho(TG(S_n^{[p]}, P_n^{[p]}))$ in the case $\beta = -30$ and $\gamma = 1$, for the specified parameter $\omega^{[p]}$

n	$\rho_n^{[1]}$ $[\omega^{[1]} = 1/3]$	$\rho_n^{[3]}$ $[\omega^{[3]} = 1.0368]$	$\rho_n^{[5]}$ $[\omega^{[5]} = 1.2576]$
80	0.3230098	0.4790377	0.9108503
160	0.3307569	0.4575274	0.8976670
320	0.3326895	0.4506303	0.8940369
640	0.3331724	0.4483575	0.8930053
1280	0.3332931	0.4475402	0.8926882
2560	0.3333233	0.4472142	0.8925937
n	$\rho_n^{[2]}$ $[\omega^{[2]} = 0.7311]$	$\rho_n^{[4]}$ $[\omega^{[4]} = 1.2229]$	$\rho_n^{[6]}$ $[\omega^{[6]} = 1.2235]$
81	0.1048347	0.7610977	0.9726925
161	0.0534626	0.7440238	0.9632887
321	0.0335327	0.7391183	0.9606204
641	0.0267034	0.7377120	0.9598142
1281	0.0251805	0.7372869	0.9595439
2561	0.0251951	0.7371457	0.9594419

We also investigated the behavior of the TGM in the case of the diffusion-advection-reaction problem

$$\begin{cases} -u'' + \beta u' + \gamma u = \mathrm{f}, \text{ in } (0,1), \\ u(0) = 0, \quad u(1) = 0, \end{cases}$$

with $\beta = -30$ and $\gamma = 1$. It is known (see Remark 1) that the corresponding sequence of Galerkin B-spline matrices has the same symbol f_p as in (21), which is independent of β and γ. The results of some numerical experiments (with a similar setup as in the previous test) are collected in Tables 3 and 4. We may conclude that the nonzero advection and reaction terms do not have a major influence on the *asymptotic* spectral radii of the TGM both in the case of the Richardson and the Gauss-Seidel smoother. However, as explained in Remark 1, the presence of a very large $|\beta|$ may affect (negatively) the two-grid convergence rate for relatively small values of n.

5.2 2D examples

Table 5 shows the results of some numerical experiments for $TG(S_{n,n}^{[p,p]}, P_{n,n}^{[p,p]})$ applied to a system with coefficient matrix $A_{n,n}^{[p,p]}$ and $\gamma = 1$; see (23). The value of the relaxation parameter $\omega^{[p,p]}$ for the relaxed Richardson smoother $S_{n,n}^{[p,p]}$ was taken as in [9, Table 12], and was determined in order to approximately minimize the asymptotic spectral radius in the pure Laplacian case ($\gamma = 0$). Then, we computed the spectral radii $\rho_{n,n}^{[p,p]} := \rho(TG(S_{n,n}^{[p,p]}, P_{n,n}^{[p,p]}))$ for $p = 1, \ldots, 6$ and increasing values of n. In all the consid-

Table 4 Values of $\widehat{\rho}_n^{[p]} := \rho(TG(\widehat{S}_n^{[p]}, P_n^{[p]}))$ in the case $\beta = -30$ and $\gamma = 1$, for the specified parameter $\omega^{[p]}$

n	$\widehat{\rho}_n^{[1]}$ [$\omega^{[1]} = 0.9065$]	$\widehat{\rho}_n^{[3]}$ [$\omega^{[3]} = 0.9483$]	$\widehat{\rho}_n^{[5]}$ [$\omega^{[5]} = 1.1999$]
80	0.2628693	0.2662531	0.3914766
160	0.2414011	0.1807720	0.3528311
320	0.2374313	0.1638459	0.4034783
640	0.2402989	0.1608873	0.4393023
1280	0.2432890	0.1609092	0.4607042
2560	0.2452101	0.1611828	0.4707694
n	$\widehat{\rho}_n^{[2]}$ [$\omega^{[2]} = 0.9109$]	$\widehat{\rho}_n^{[4]}$ [$\omega^{[4]} = 1.0602$]	$\widehat{\rho}_n^{[6]}$ [$\omega^{[6]} = 1.3292$]
81	0.1724812	0.2852895	0.5864728
161	0.1132497	0.2574381	0.4942180
321	0.0837635	0.2831336	0.5396466
641	0.0731110	0.3053799	0.5793680
1281	0.0688985	0.3194229	0.6005087
2561	0.0670204	0.3264244	0.6121435

Table 5 Values of $\rho_{n,n}^{[p,p]} := \rho(TG(S_{n,n}^{[p,p]}, P_{n,n}^{[p,p]}))$ in the case $\gamma = 1$, for the specified parameter $\omega^{[p,p]}$

n	$\rho_{n,n}^{[1,1]}$ [$\omega^{[1,1]} = 0.3335$]	$\rho_{n,n}^{[3,3]}$ [$\omega^{[3,3]} = 1.3739$]	$\rho_{n,n}^{[5,5]}$ [$\omega^{[5,5]} = 1.3293$]
16	0.3278650	0.9250838	0.9984588
28	0.3313190	0.9245759	0.9983433
40	0.3321758	0.9233720	0.9983185
52	0.3325122	0.9231215	0.9983133
n	$\rho_{n,n}^{[2,2]}$ [$\omega^{[2,2]} = 1.1009$]	$\rho_{n,n}^{[4,4]}$ [$\omega^{[4,4]} = 1.4000$]	$\rho_{n,n}^{[6,6]}$ [$\omega^{[6,6]} = 1.2505$]
17	0.6085012	0.9885328	0.9997976
29	0.6085456	0.9881167	0.9997766
41	0.6085572	0.9880109	0.9997724
53	0.6085998	0.9879838	0.9997715

Table 6 Values of $\widehat{\rho}_{n,n}^{[p,p]} := \rho(TG(\widehat{S}_{n,n}^{[p,p]}, P_{n,n}^{[p,p]}))$ in the case $\gamma = 1$, for the specified parameter $\omega^{[p,p]}$

n	$\widehat{\rho}_{n,n}^{[1,1]}$ [$\omega^{[1,1]} = 1.0035$]	$\widehat{\rho}_{n,n}^{[3,3]}$ [$\omega^{[3,3]} = 1.3143$]	$\widehat{\rho}_{n,n}^{[5,5]}$ [$\omega^{[5,5]} = 1.3990$]
16	0.1586321	0.6423071	0.9630879
28	0.1677612	0.6412603	0.9633778
40	0.1749603	0.6418210	0.9626884
52	0.1802092	0.6464128	0.9620643
n	$\widehat{\rho}_{n,n}^{[2,2]}$ [$\omega^{[2,2]} = 1.1695$]	$\widehat{\rho}_{n,n}^{[4,4]}$ [$\omega^{[4,4]} = 1.3248$]	$\widehat{\rho}_{n,n}^{[6,6]}$ [$\omega^{[6,6]} = 1.4914$]
17	0.2661695	0.8798789	0.9913530
29	0.2683807	0.8780234	0.9903296
41	0.2896229	0.8773965	0.9898868
53	0.3041406	0.8778226	0.9897379

ered numerical experiments, the proposed TGM is optimal. However, for $p \geq 3$, the (asymptotic) spectral radius is very close to 1, and this is not satisfactory for practical purposes.

The numerical experiments in Table 6, obtained as those in Table 5, show a certain improvement in the two-grid convergence rate when using the relaxed Gauss-Seidel smoother instead of Richardson's, but again the results worsen for increasing p.

An effective smoother for large p based on a preconditioned Krylov method has been proposed in [11], whereas an extensive numerical testing can be found in [9].

6 Conclusion and perspectives

In this paper we have proposed two-grid (and multigrid) methods for the solution of linear systems arising from the Galerkin B-spline IgA approximation of 1D and 2D elliptic problems. The optimality of the two-grid scheme, already predicted in [9], has been formally proved for some values of the spline degrees $\boldsymbol{p}$. It is important to point out that:

– the proposal of the methods is motivated and based on the spectral symbol and on the corresponding techniques for τ-matrices [1,2,12,13,22]; we could also have opted for Toeplitz matrices;
– the optimality proofs are based on classical tools from algebraic multigrid analysis, applied within the framework of the theory of τ-matrices [20,22]; again, we could have opted for Toeplitz matrices;
– the spectral properties of the considered matrices, as well as the properties of the associated symbol, were analyzed in a previous work [16].

A plan for future research could include the proof of relation (26) for all $p \geq 1$. It would give at once the optimality proof of the two-grid and—with a little more effort—the optimality proof of the W-cycle multigrid method.

Acknowledgments This work was partially supported by INdAM-GNCS Gruppo Nazionale per il Calcolo Scientifico, by the MIUR 'Futuro in Ricerca 2013' Programme through the project DREAMS, by the MIUR-PRIN 2012 N. 2012MTE38N, and by the Donation KAW 2013.0341 from the Knut & Alice Wallenberg Foundation in collaboration with the Royal Swedish Academy of Sciences, supporting Swedish research in mathematics.

References

1. Aricò, A., Donatelli, M.: A V-cycle multigrid for multilevel matrix algebras: proof of optimality. Numer. Math. **105**, 511–547 (2007)
2. Aricò, 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)
3. Axelsson, O., Lindskog, G.: On the rate of convergence of the preconditioned conjugate gradient method. Numer. Math. **48**, 499–523 (1986)
4. Beckermann, B., Serra-Capizzano, S.: On the asymptotic spectrum of finite element matrix sequences. SIAM J. Numer. Anal. **45**, 746–769 (2007)
5. Bhatia, R.: Matrix Analysis. Springer, New York (1997)
6. Boor, C. de: A Practical Guide to Splines. Springer, New York (2001)
7. Cottrell, J.A., Hughes, T.J.R., Bazilevs, Y.: Isogeometric Analysis: Toward Integration of CAD and FEA. Wiley, Chichester (2009)
8. Donatelli, M.: An algebraic generalization of local Fourier analysis for grid transfer operators in multigrid based on Toeplitz matrices. Numer. Linear Algebra Appl. **17**, 179–197 (2010)
9. Donatelli, M., Garoni, C., Manni, C., Serra-Capizzano, S., Speleers, H.: Robust and optimal multi-iterative techniques for IgA Galerkin linear systems. Comput. Methods Appl. Mech. Eng. **284**, 230–264 (2015)
10. Donatelli, M., Garoni, C., Manni, C., Serra-Capizzano, S., Speleers, H.: Robust and optimal multi-iterative techniques for IgA collocation linear systems. Comput. Methods Appl. Mech. Eng. **284**, 1120–1146 (2015)
11. Donatelli, M., Garoni, C., Manni, C., Serra-Capizzano, S., Speleers, H.: Symbol-based multigrid methods for Galerkin B-spline isogeometric analysis, submitted
12. Fiorentino, G., Serra, S.: Multigrid methods for Toeplitz matrices. Calcolo **28**, 283–305 (1991)
13. Fiorentino, G., Serra, S.: Multigrid methods for symmetric positive definite block Toeplitz matrices with nonnegative generating functions. SIAM J. Sci. Comput. **17**, 1068–1081 (1996)
14. Gahalaut, K.P.S., Kraus, J.K., Tomar, S.K.: Multigrid methods for isogeometric discretization. Comput. Methods Appl. Mech. Eng. **253**, 413–425 (2013)
15. Garoni, C.: Structured matrices coming from PDE approximation theory: spectral analysis, spectral symbol and design of fast iterative solvers. Ph.D. Thesis in Mathematics of Computation, University of Insubria, Como, Italy. http://hdl.handle.net/10277/568 (2015)
16. Garoni, C., Manni, C., Pelosi, F., Serra-Capizzano, S., Speleers, H.: On the spectrum of stiffness matrices arising from isogeometric analysis. Numer. Math. **127**, 751–799 (2014)
17. Garoni, C., Manni, C., Serra-Capizzano, S., Sesana, D., Speleers, H.: Spectral analysis and spectral symbol of matrices in isogeometric Galerkin methods. Technical Report 2015-005, Department of Information Technology, Uppsala University, Sweden (2015)
18. 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)
19. Jin, X.Q.: Developments and Applications of Block Toeplitz Iterative Solvers. Kluwer Academic Publishers, Dordrecht (2002)
20. Ruge, J.W., Stüben, K.: Algebraic multigrid, Chapter 4 of the book Multigrid Methods by S. McCormick. SIAM Publications, Philadelphia (1987)
21. Serra, S.: Multi-iterative methods. Comput. Math. Appl. **26**, 65–87 (1993)
22. Serra-Capizzano, S.: Convergence analysis of two-grid methods for elliptic Toeplitz and PDEs matrix-sequences. Numer. Math. **92**, 433–465 (2002)
23. Serra-Capizzano, S.: Generalized locally Toeplitz sequences: spectral analysis and applications to discretized partial differential equations. Linear Algebra Appl. **366**, 371–402 (2003)
24. Serra-Capizzano, S.: The GLT class as a generalized Fourier analysis and applications. Linear Algebra Appl. **419**, 180–233 (2006)
25. 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)
26. Serra-Capizzano, S., Tablino-Possio, C.: Positive representation formulas for finite difference discretizations of (elliptic) second order PDEs. Contemp. Math. **281**, 295–318 (2001)