Open access • Journal Article • DOI:10.1137/140988590

Symbol-Based Multigrid Methods for Galerkin B-Spline Isogeometric Analysis — Source link

Marco Donatelli, Carlo Garoni, Carla Manni, Stefano Serra-Capizzano ...+1 more authors

Published on: 05 Jan 2017 - SIAM Journal on Numerical Analysis (Society for Industrial and Applied Mathematics)

Topics: Multigrid method, Isogeometric analysis, Rate of convergence, B-spline and Galerkin method

Symbol-based multigrid methods for Galerkin B-spline isogeometric analysis

Marco Donatelli   Carlo Garoni   Carla Manni
Stefano Serra-Capizzano   Hendrik Speleers

Report TW 650, July 2014

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

Symbol-based multigrid methods for Galerkin B-spline isogeometric analysis

Marco Donatelli *Carlo Garoni* *Carla Manni*
Stefano Serra-Capizzano *Hendrik Speleers*

Report TW 650, July 2014

Department of Computer Science, KU Leuven

Abstract

We consider the stiffness matrices coming from the Galerkin B-spline isogeometric analysis approximation of classical elliptic problems. By exploiting specific spectral properties compactly described by a symbol, we design efficient multigrid methods for the fast solution of the related linear systems. We prove the optimality of the two-grid methods (in the sense that their convergence rate is independent of the matrix size) for spline degrees up to 3, both in the 1D and 2D case. Despite the theoretical optimality, the convergence rate of the two-grid methods with classical stationary smoothers worsens exponentially when the spline degrees increase. With the aid of the symbol, we provide a theoretical explanation of this exponential worsening and by a proper factorization of the symbol we provide a preconditioned conjugate gradient 'smoother', in the spirit of the multi-iterative strategy, that allows us to obtain a good convergence rate independent both of the matrix size and of the spline degrees. A selected set of numerical experiments confirms the effectiveness of our proposal and the numerical optimality with a uniformly high convergence rate, also for the V-cycle multigrid method and large spline degrees.

Keywords : Multigrid methods, preconditioning, isogeometric analysis, B-splines, Toeplitz matrices.

SYMBOL-BASED MULTIGRID METHODS FOR GALERKIN B-SPLINE ISOGEOMETRIC ANALYSIS

MARCO DONATELLI [†], CARLO GARONI [†], CARLA MANNI [‡], STEFANO SERRA-CAPIZZANO [†], AND HENDRIK SPELEERS [‡§]

Abstract. We consider the stiffness matrices coming from the Galerkin B-spline isogeometric analysis approximation of classical elliptic problems. By exploiting specific spectral properties compactly described by a symbol, we design efficient multigrid methods for the fast solution of the related linear systems. We prove the optimality of the two-grid methods (in the sense that their convergence rate is independent of the matrix size) for spline degrees up to 3, both in the 1D and 2D case. Despite the theoretical optimality, the convergence rate of the two-grid methods with classical stationary smoothers worsens exponentially when the spline degrees increase. With the aid of the symbol, we provide a theoretical explanation of this exponential worsening and by a proper factorization of the symbol we provide a preconditioned conjugate gradient 'smoother', in the spirit of the multi-iterative strategy, that allows us to obtain a good convergence rate independent both of the matrix size and of the spline degrees. A selected set of numerical experiments confirms the effectiveness of our proposal and the numerical optimality with a uniformly high convergence rate, also for the V-cycle multigrid method and large spline degrees.

Key words. Multigrid methods, preconditioning, isogeometric analysis, B-splines, Toeplitz matrices.

AMS subject classifications. 65N55, 65N30, 65F08.

1. Introduction.

We consider the model problem

$$\begin{cases} -\Delta u + \boldsymbol{\beta} \cdot \nabla u + \gamma u = \mathrm{f}, & \text{in } \Omega, \\ u = 0, & \text{on } \partial\Omega, \end{cases} \tag{1.1}$$

with $\Omega := (0,1)^d$, $\mathrm{f} \in L^2(\Omega)$, $\boldsymbol{\beta} := (\beta_1, \dots, \beta_d) \in \mathbb{R}^d$, $\gamma \geq 0$. Our aim is the design of fast solvers for large linear systems coming from the Galerkin B-spline Isogeometric Analysis (IgA) discretization of problem (1.1), see [12]. IgA was introduced in [26] aiming to reduce the gap between the worlds of 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. [9]), this approach offers some interesting advantages from the geometric, the analytic, and the computational point of view, see [12, 26] and references therein.

In the recent work [23], the spectral properties of the Galerkin B-spline IgA stiffness matrices have been studied in some detail. In particular, the spectral localization and the conditioning were investigated, while the asymptotic spectral distribution, as the matrix size tends to infinity, has been compactly characterized in terms of a d-variate trigonometric polynomial, denoted by $f := f_{\boldsymbol{p}}$, with $\boldsymbol{p} := (p_1, \dots, p_d)$ and p_j

[†] Department of Science and High Technology, University of Insubria, Via Valleggio 11, 22100 Como, Italy (email addresses: marco.donatelli@uninsubria.it, carlo.garoni@uninsubria.it, stefano.serrac@uninsubria.it).

[‡] Department of Mathematics, University of Rome 'Tor Vergata', Via della Ricerca Scientifica, 00133 Rome, Italy (email addresses: manni@mat.uniroma2.it, speleers@mat.uniroma2.it).

[§] Department of Computer Science, University of Leuven, Celestijnenlaan 200A, 3001 Heverlee (Leuven), Belgium (email address: hendrik.speleers@cs.kuleuven.be).

being the spline degree in the direction x_j, $j = 1, \dots, d$. In analogy with Finite Difference (FD) and Finite Element (FE) cases, the conditioning grows as $m^{2/d}$, where m is the matrix size, d is the dimensionality of the elliptic problem, and 2 is the order of the elliptic operator in (1.1). As expected, the approximation parameters $\boldsymbol{p}$ play a limited role, since they only characterize the constant in the expression $O(m^{2/d})$.

The growth of the condition number implies that all classical stationary iterative methods (if convergent) and the Krylov methods are not optimal, in the sense that the number of iterations for reaching a preassigned accuracy ϵ is a function diverging to infinity as the matrix size m tends to infinity. We specify that the notion of optimality for an iterative method is twofold. First, the number of iterations for reaching a preassigned accuracy ϵ must be bounded by a constant $c(\epsilon)$ independent of the matrix size: the latter is also known as the *optimal convergence rate condition*. For stationary iterative methods, it translates into the requirement that the spectral radius of the iteration matrix is bounded by a constant $c < 1$ independent of the matrix size. Second, when the matrix size goes to infinity, the cost per iteration must be asymptotically of the same order as the cost of multiplying the matrix by a vector.[1]

In order to design optimal methods, we heavily rely on the spectral and structural information of the coefficient matrices analyzed in detail in [23]. More precisely, the coefficient matrices coming from the IgA approximation to equation (1.1) are

- banded in a d-level sense with partial bandwidths proportional to p_j, $j = 1, \dots, d$;
- a low rank correction of a d-level Toeplitz matrix generated by $f_{\boldsymbol{p}}$ and are spectrally distributed as $f_{\boldsymbol{p}}$.

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 Toeplitz solvers [27]. Concerning preconditioned Krylov solvers, for $d > 1$ it has been proved that matrix algebra preconditioners (like circulants, Hartley matrices, matrices associated to trigonometric transforms [27], etc) cannot be optimal (see [31, 28] and references therein), when there is asymptotic ill-conditioning as in our setting. This restricts us to considering and adapting multigrid methods (V- and W-cycles) of the kind devised in [1, 20] to our case, with the aim of designing optimal iterative solvers. As a first step, we follow (see [19, 32]) a sort of ‘canonical procedure’ for creating – based on the symbol – a two-grid method from which we expect optimal convergence properties. We also refer to [16] for another application in a Discontinuous Galerkin context. The optimality result is proved formally for the two-grid method and some values of $\boldsymbol{p}$ and hence also for the W-cycle k-grid method (with k independent of the matrix size), whereas for the V-cycle the result is numerically observed. When proving the optimality result for the two-grid method, we arrive at a matrix inequality, see (3.8), which is useful not only in a multigrid setting, but can be also employed in a preconditioning context for designing optimal preconditioners for Krylov-type techniques, in particular for the Conjugate Gradient (CG) method, see the discussion in Section 8.

Despite the m-independence of the convergence rate in the two-grid case, the method is, however, not really satisfactory 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

[1] The optimal cost requirement is often easily satisfied and thus, throughout this paper, we take the convergence rate independent of m as a synonymous of optimality.

tends to 1 exponentially as p increases and a similar phenomenon is observed for any dimensionality d. This catastrophic 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. Therefore, the considered two-grid and the associated V/W-cycle methods converge very fast in low frequencies but they are slow, for large $\boldsymbol{p}$, in high frequencies. This fact is nontrivial: it can be understood in terms of the theory of multilevel Toeplitz matrices and is related to interesting analytic features of the symbol. We refer to Section 4 for a theoretical proof of these facts.

In order to address the above intrinsic difficulty, we enrich our two-grid procedure by varying the choice of the smoothers, in the sense of the multi-iterative idea [30]. More precisely, we consider a Preconditioned Conjugate Gradient (PCG) technique for our specific linear algebra problem, designed ad hoc for reducing the error in high frequencies. In fact, the related preconditioner is chosen as the Toeplitz matrix generated by a specific function coming from a factorization of the symbol. This method induces a convergence which is independent of $\boldsymbol{p}$. Unfortunately, the conditioning of the preconditioned matrix still grows as $m^{2/d}$, so that the number of PCG iterations grows as $m^{1/d}$, and the error is slowly reduced in the low frequency space.

In other words, we have identified an optimal two-grid procedure and a $\boldsymbol{p}$-independent PCG technique: the former is especially effective in low frequencies, whereas the latter is very slow in low frequencies but effective in high frequencies. Following [30], our multi-iterative proposal consists in using few steps of the proposed PCG technique as a smoother in our multigrid method and only at the finest level. The combination of these two techniques with complementary spectral features – one converging well in low frequencies, the other converging well in high frequencies – leads to a global iteration which is optimal and whose convergence speed turns out to be substantially independent of all the relevant parameters.

The relevant literature on fast solvers for IgA linear systems seems to be very recent and quite limited [5, 6, 11, 21, 22]. Even though in some of the contributions (see e.g. [21]) the bad dependency on the parameter $\boldsymbol{p}$ was observed, there was no understanding that spurious small eigenvalues are present already for $p_j \geq 4$ and that the related eigenspace largely intersects the high frequencies. The latter phenomenon is indeed unexpected in this context, since high frequency eigenspaces related to small eigenvalues are typical of matrices coming from the approximation of integral operators, like in the setting of blurring models in imaging and signal processing (see e.g. [18]). Although the combination of multigrid and Krylov methods was already investigated in [22], in order to obtain a robust solver with a convergence rate independent of p, our approach follows a different strategy using a much simpler PCG inside an elementary geometric multigrid method.

As a matter of fact, the basic ingredients of the techniques used so far are not different from the ones in our proposal: different kinds of preconditioning and various types of multigrid algorithms. However, the novelty of our proposal is that the choice of the ingredients of the global solver (in fact a multi-iterative solver) is guided by the knowledge of the symbol which in turn offers an approximate understanding of the subspaces where the stiffness matrix is ill-conditioned. By exploiting the information given by the symbol, we are able to design a cheap (indeed optimal) solver of multi-iterative type, whose convergence speed is independent of all the relevant parameters of the problem: the fineness parameter (related to the size of the matrices), the approximation parameters (i.e. the degrees $\boldsymbol{p}$), and the dimensionality d.

The paper is organized as follows. In Section 2 we detail the considered model problem; we define d-level τ-matrices and d-level Toeplitz matrices; and we describe the canonical scheme of the two-grid method. Section 3 is devoted to the one-dimensional setting, where optimality results are proved and a poor behavior (for increasing p) is observed. Section 4 provides a local Fourier analysis of the two-grid method, which explains why the method does not work satisfactorily for large p. In order to maintain the optimality and to add robustness with respect to p, in Section 5 we introduce a multi-iterative strategy, by using a specialized PCG smoothing. Section 6 deals with extensions of the analysis and of the proposals to the two-dimensional setting. In Section 7 we give a numerical evidence that the proposed V-cycle multigrid algorithm is also optimal and effective in practice. Finally, Section 8 concludes the work, by emphasizing perspectives and open problems. We refer to the twin paper [14] for an extensive numerical testing and comparison of the different kind of fast solvers mentioned above and beyond (two-grid and multigrid methods with different smoothing and size reducing strategies) in accordance with the given spectral analysis.

2. Preliminaries.

We first present the coefficient matrices coming from the Galerkin B-spline IgA approximation of (1.1) and then we introduce some auxiliary structures, namely d-level Toeplitz and τ matrices, which are used for designing the proposed algorithms and for studying their converge features.

2.1. The d-dimensional problem setting.

Our model problem is the elliptic problem (1.1), which can be solved in the weak form as follows: find $u \in H_0^1(\Omega)$ such that

$$a(u,v) = \mathrm{F}(v), \qquad \forall v \in H_0^1(\Omega), \tag{2.1}$$

where $a(u,v) := \int_\Omega (\nabla u \cdot \nabla v + \boldsymbol{\beta} \cdot \nabla u\, v + \gamma u v)$ and $\mathrm{F}(v) := \int_\Omega \mathrm{f} v$. It is known [10] that there exists a unique solution u of (2.1), the so-called weak solution of (1.1).

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 (Galerkin) problem: find $u_{\mathcal{W}} \in \mathcal{W}$ such that

$$a(u_{\mathcal{W}}, v) = \mathrm{F}(v), \qquad \forall v \in \mathcal{W}. \tag{2.2}$$

Let $\dim \mathcal{W} = N$, and fix a basis $\{\varphi_1, \ldots, \varphi_N\}$ for $\mathcal{W}$. It is known that the problem (2.2) always has a unique solution $u_{\mathcal{W}} \in \mathcal{W}$, which can be written as $u_{\mathcal{W}} = \sum_{j=1}^N u_j \varphi_j$ and can be computed as follows: find $\mathbf{u} := (u_1, \ldots, u_N)^T \in \mathbb{R}^N$ such that

$$A\mathbf{u} = \mathbf{b}, \tag{2.3}$$

where $A := [a(\varphi_j, \varphi_i)]_{i,j=1}^N \in \mathbb{R}^{N\times N}$ is the stiffness matrix, and $\mathbf{b} := [\mathrm{F}(\varphi_i)]_{i=1}^N \in \mathbb{R}^N$. The matrix A is positive definite in the sense that $\mathbf{v}^T A \mathbf{v} > 0$, $\forall \mathbf{v} \in \mathbb{R}^N \setminus \{\mathbf{0}\}$.

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 [12].

2.2. d-level τ-matrices and d-level Toeplitz matrices.

In this paper, for every $m \in \mathbb{N}$ we denote by $\mathcal{S}_m$ the unitary discrete sine transform,

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

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

Definition 2.1. *Given a d-variate function $g : [0, \pi]^d \to \mathbb{R}$ and a multi-index $\boldsymbol{m} \in \mathbb{N}^d$, $\tau_{\boldsymbol{m}}(g)$ is the d-level τ-matrix of partial orders $m_1, \ldots, m_d$ (and order $m_1 \cdots m_d$) associated with g, i.e.,*

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

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

$$\lim_{\boldsymbol{m}\to\infty} \frac{1}{m_1 \cdots m_d} \sum_{j=1}^{m_1\cdots m_d} F[\lambda_j(\tau_{\boldsymbol{m}}(g))] = \frac{1}{\pi^d} \int_{[0,\pi]^d} F[g(\theta_1, \ldots, \theta_d)]\, \mathrm{d}\theta_1 \cdots \mathrm{d}\theta_d, \tag{2.4}$$

where, for a multi-index $\boldsymbol{m} \in \mathbb{N}^d$, $\boldsymbol{m} \to \infty$ means that $\min(m_1, \ldots, m_d) \to \infty$. Due to (2.4), the function g is called the (spectral) symbol of the τ-family $\{\tau_{\boldsymbol{m}}(g)\}_{\boldsymbol{m}\in\mathbb{N}^d}$.

Definition 2.2. *Given a d-variate function $g : [-\pi, \pi]^d \to \mathbb{R}$ in $L^1([-\pi, \pi]^d)$ and a multi-index $\boldsymbol{m} \in \mathbb{N}^d$, $T_{\boldsymbol{m}}(g)$ is the d-level Toeplitz matrix of partial orders $m_1, \ldots, m_d$ (and order $m_1 \cdots m_d$) associated with g, i.e.,*

$$T_{\boldsymbol{m}}(g) := \left[\cdots \left[[g_{i_1-j_1, i_2-j_2, \ldots, i_d-j_d}]_{i_d,j_d=1}^{m_d} \right]_{i_{d-1},j_{d-1}=1}^{m_{d-1}} \cdots \right]_{i_1,j_1=1}^{m_1},$$

where $g_{i_1,i_2,\ldots,i_d}$, $i_1, i_2, \ldots, i_d \in \mathbb{Z}$, are the Fourier coefficients of g,

$$g_{i_1,i_2,\ldots,i_d} = \frac{1}{(2\pi)^d} \int_{[-\pi,\pi]^d} g(\theta_1, \ldots, \theta_d) \mathrm{e}^{-\mathrm{i}(i_1\theta_1 + i_2\theta_2 + \ldots + i_d\theta_d)} \mathrm{d}\theta_1 \cdots \mathrm{d}\theta_d.$$

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

By the Szegö-Tilli theorem [35], a relation similar to (2.4) holds for $\{T_{\boldsymbol{m}}(g)\}_{\boldsymbol{m}\in\mathbb{N}^d}$: $\forall F \in C_c(\mathbb{C})$,

$$\lim_{\boldsymbol{m}\to\infty} \frac{1}{m_1 \cdots m_d} \sum_{j=1}^{m_1\cdots m_d} F[\lambda_j(T_{\boldsymbol{m}}(g))] = \frac{1}{(2\pi)^d} \int_{[-\pi,\pi]^d} F[g(\theta_1, \ldots, \theta_d)]\, \mathrm{d}\theta_1 \cdots \mathrm{d}\theta_d. \tag{2.5}$$

For this reason, g is called the symbol of the Toeplitz family $\{T_{\boldsymbol{m}}(g)\}_{\boldsymbol{m}\in\mathbb{N}^d}$.

Suppose that $g : [-\pi, \pi]^d \to \mathbb{R}$ is continuous over $[-\pi, \pi]^d$ and symmetric in each variable, in the sense that $g(\varepsilon_1\theta_1, \ldots, \varepsilon_d\theta_d) = g(\theta_1, \ldots, \theta_d)$ for $(\theta_1, \ldots, \theta_d) \in [-\pi, \pi]^d$ and $(\varepsilon_1, \ldots, \varepsilon_d) \in \{-1, 1\}^d$. Then, the right-hand sides of (2.4) and (2.5) coincide and g is simultaneously the symbol of $\{\tau_{\boldsymbol{m}}(g)\}_{\boldsymbol{m}\in\mathbb{N}^d}$ and $\{T_{\boldsymbol{m}}(g)\}_{\boldsymbol{m}\in\mathbb{N}^d}$.

It can also be shown that, if g is a linear d-variate cosine trigonometric polynomial, i.e. $g(\theta_1, \ldots, \theta_d) = \sum_{j_1=0}^{1} \cdots \sum_{j_d=0}^{1} a_{j_1,\ldots,j_d} \cos(j_1\theta_1) \cdots \cos(j_d\theta_d)$ for some coefficients $a_{j_1,\ldots,j_d} \in \mathbb{R}$, then $\tau_{\boldsymbol{m}}(g) = T_{\boldsymbol{m}}(g)$, $\forall \boldsymbol{m} \in \mathbb{N}^d$.

2.3. Two-grid methods.

Given a linear system of dimension m,

$$A_m\mathbf{u} = \mathbf{b}, \tag{2.6}$$

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 (2.6), and a full-rank matrix $P_m \in \mathbb{R}^{l\times m}$ with $l \leq m$, called projector or grid transfer operator. Moreover, we define the coarse matrix as $(P_m A_m P_m^T)^{-1}$ 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 (2.6) computes a new approximation $\mathbf{u}^{(k+1)}$ by applying a coarse-grid correction and a smoothing iteration as follows:

Algorithm 2.3 (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. solve the coarse error equation: $\mathbf{e} \leftarrow \left(P_m A_m P_m^T\right)^{-1}\mathbf{r}$;
4. extend the coarse error: $\mathbf{e} \leftarrow P_m^T\mathbf{e}$;
5. correct the initial 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 the above two-grid scheme is

$$TG(S_m, P_m) := S_m\left(I - P_m^T\left(P_m A_m P_m^T\right)^{-1} P_m A_m\right). \tag{2.7}$$

Note that Algorithm 2.3 only considers a single post-smoothing iteration for the sake of simplicity in the presentation of the theoretical analysis according to the framework [29], but it is clear that one can add a convergent pre-smoother or more smoothing iterations improving the convergence rate of the TGM.

The optimality proofs for the two-grid methods, discussed in this paper, heavily rely on Theorem 2.4. For its proof, we refer to [29, Theorem 5.2] and [2, Remark 2.2]. If $X \in \mathbb{R}^{m\times m}$ is a Symmetric Positive Definite (SPD) matrix, then we denote by $\|\cdot\|_X$ both the vector-norm and the matrix-norm induced by X, i.e. $\|\mathbf{x}\|_X = \|X^{1/2}\mathbf{x}\|_2$, $\mathbf{x} \in \mathbb{R}^m$ and $\|Y\|_X = \|X^{1/2}YX^{-1/2}\|_2$, $Y \in \mathbb{R}^{m\times m}$, where $\|\cdot\|_2$ denotes both the classical 2-norm (the Euclidean norm) and its induced matrix-norm. Moreover, 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.

Theorem 2.4. *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 ($l \leq m$). Assume*

(a) $\exists\, a_m > 0 : \|S_m\mathbf{x}\|_{A_m}^2 \leq \|\mathbf{x}\|_{A_m}^2 - a_m\|\mathbf{x}\|_{A_m^2}^2, \quad \forall \mathbf{x} \in \mathbb{R}^m$;

(b) $\exists\, b_m > 0 : \min_{\mathbf{y}\in\mathbb{R}^l}\|\mathbf{x} - P_m^T\mathbf{y}\|_2^2 \leq b_m\|\mathbf{x}\|_{A_m}^2, \quad \forall \mathbf{x} \in \mathbb{R}^m$.

Then $b_m \geq a_m$, and

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

The condition (a) in Theorem 2.4 is usually referred to as *smoothing condition*, and the condition (b) as *approximation condition*. In the following, we discuss the values of the constants a_m and b_m for specific smoothers and projectors.

When using the Richardson iteration, the smoothing condition can be easily satisfied and the next lemma can be proved in the same way as [29, Theorem 4.4] (with $D = I$ and $Q = I/\omega$).

Lemma 2.5. *Let $A_m \in \mathbb{R}^{m\times m}$ be SPD, let $S_m := I - \omega A_m$ ($\omega \in \mathbb{R}$), and assume $\mu_m \geq \rho(A_m)$. Then, the smoothing condition (a) in Theorem 2.4 holds if $0 < \omega < 2/\mu_m$. Moreover, in this case we also have $\rho(S_m) < 1$ and the smoothing condition (a) in Theorem 2.4 holds with $a_m := \omega(2 - \omega\mu_m) > 0$.*

We now define the projector $P_{\boldsymbol{m}}$ for multi-indices $\boldsymbol{m} \in \mathbb{N}^d$ satisfying certain additional constraints to be seen later. 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

$$P_{\boldsymbol{m}} := U_{\boldsymbol{m}} T_{\boldsymbol{m}}(q_d) = U_{\boldsymbol{m}} \tau_{\boldsymbol{m}}(q_d), \qquad q_d(\theta_1, \ldots, \theta_d) := \prod_{j=1}^{d} (1 + \cos\theta_j). \tag{2.8}$$

The equality $T_{\boldsymbol{m}}(q_d) = \tau_{\boldsymbol{m}}(q_d)$ in (2.8) holds because q_d is a linear d-variate cosine trigonometric polynomial. By the properties of the Kronecker tensor-product it holds that

$$P_{\boldsymbol{m}} = \bigotimes_{j=1}^{d} P_{m_j}, \qquad P_{m_j} = U_{m_j} \tau_{m_j}(q_1), \qquad q_1(\theta) = 1 + \cos\theta,$$

resulting in

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

which has full rank $\prod_{j=1}^{d} \frac{m_j - 1}{2}$ and it is the standard d-linear interpolation operator.

Let

$$z_d(\theta_1, \ldots, \theta_d) := \sum_{j=1}^{d} (2 - 2\cos\theta_j), \tag{2.9}$$

which is a linear nonnegative d-variate cosine trigonometric polynomial with a unique zero at $(0, \ldots, 0)$ over $[0, \pi]^d$. Lemma 2.6 addresses the approximation condition in Theorem 2.4 in the case where A_m is the particular τ-matrix $\tau_{\boldsymbol{m}}(z_d)$ (of size $m = m_1 \cdots m_d$) and $P_m = P_{\boldsymbol{m}}$. The lemma is a direct consequence of [32, Lemma 8.2] thanks to the following two properties of q_d and z_d reported in (2.8) and (2.9), respectively. Let $\mathcal{M}(\boldsymbol{\theta})$ be the set of mirror points of $\boldsymbol{\theta} := (\theta_1, \ldots, \theta_d)$ as defined in [32, p. 454], namely

$$\mathcal{M}(\boldsymbol{\theta}) := \left\{ \widehat{\boldsymbol{\theta}} := (\widehat{\theta}_1, \ldots, \widehat{\theta}_d) \in [0, \pi]^d : \ \widehat{\theta}_i \in \{\theta_i, \pi - \theta_i\}, \ \forall i = 1, \ldots, d \right\} \setminus \{\boldsymbol{\theta}\}, \tag{2.10}$$

then [2]

$$\sum_{\widehat{\boldsymbol{\theta}} \in \mathcal{M}(\boldsymbol{\theta}) \cup \{\boldsymbol{\theta}\}} q_d^2(\widehat{\boldsymbol{\theta}}) > 0, \ \forall \boldsymbol{\theta} \in [0, \pi]^d, \quad \text{and} \quad \limsup_{\boldsymbol{\theta} \to \mathbf{0}} \max_{\widehat{\boldsymbol{\theta}} \in \mathcal{M}(\boldsymbol{\theta})} \frac{q_d^2(\widehat{\boldsymbol{\theta}})}{z_d(\boldsymbol{\theta})} < \infty. \tag{2.11}$$

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

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

Lemma 2.6. *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 (2.8). Then, the matrix $A_{\boldsymbol{m}}$ is SPD and the approximation condition (b) in Theorem 2.4 holds with a constant depending only on d, i.e.,*

$$\exists\, \widetilde{b}_d > 0 : \quad \min_{\mathbf{y} \in \mathbb{R}^{\Pi_{j=1}^d \left(\frac{m_j-1}{2}\right)}} \|\mathbf{x} - P_{\boldsymbol{m}}^T \mathbf{y}\|_2^2 \leq \widetilde{b}_d \|\mathbf{x}\|_{A_{\boldsymbol{m}}}^2, \quad \forall \mathbf{x} \in \mathbb{R}^{m_1 \cdots m_d}. \tag{2.12}$$

Moreover, if $d = 1$ then (2.12) holds with $\widetilde{b}_1 = 1/2$.

The specific value $\widetilde{b}_1$ has been found by looking carefully at the proof of [32, Lemma 3.2]. From Lemma 2.6 we deduce the following result.

Lemma 2.7. *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 (2.8). Let $\delta_{\boldsymbol{m}} > 0$ such that*

$$A_{\boldsymbol{m}} \geq \delta_{\boldsymbol{m}}\, \tau_{\boldsymbol{m}}(z_d). \tag{2.13}$$

Then, the approximation condition (b) in Theorem 2.4 holds, i.e.,

$$\exists\, b_{\boldsymbol{m},d} := \frac{\widetilde{b}_d}{\delta_{\boldsymbol{m}}} > 0 : \quad \min_{\mathbf{y} \in \mathbb{R}^{\Pi_{j=1}^d \left(\frac{m_j-1}{2}\right)}} \|\mathbf{x} - P_{\boldsymbol{m}}^T \mathbf{y}\|_2^2 \leq b_{\boldsymbol{m},d} \|\mathbf{x}\|_{A_{\boldsymbol{m}}}^2, \quad \forall \mathbf{x} \in \mathbb{R}^{m_1 \cdots m_d},$$

where $\widetilde{b}_d$ is defined in Lemma 2.6.

Proof. We use the same monotonicity argument as in [32, proof of Lemmas 4.2 and 9.2]. Assuming (2.13), we have

$$\|\mathbf{x}\|_{\tau_{\boldsymbol{m}}(z_d)}^2 = \mathbf{x}^T \tau_{\boldsymbol{m}}(z_d) \mathbf{x} \leq \frac{1}{\delta_{\boldsymbol{m}}} \mathbf{x}^T A_{\boldsymbol{m}} \mathbf{x} = \frac{1}{\delta_{\boldsymbol{m}}} \|\mathbf{x}\|_{A_{\boldsymbol{m}}}^2, \quad \forall \mathbf{x} \in \mathbb{R}^{m_1 \cdots m_d}.$$

By Lemma 2.6 we get

$$\min_{\mathbf{y} \in \mathbb{R}^{\Pi_{j=1}^d \left(\frac{m_j-1}{2}\right)}} \|\mathbf{x} - P_{\boldsymbol{m}}^T \mathbf{y}\|_2^2 \leq \widetilde{b}_d \|\mathbf{x}\|_{\tau_{\boldsymbol{m}}(z_d)}^2 \leq \frac{\widetilde{b}_d}{\delta_{\boldsymbol{m}}} \|\mathbf{x}\|_{A_{\boldsymbol{m}}}^2, \quad \forall \mathbf{x} \in \mathbb{R}^{m_1 \cdots m_d},$$

which completes the proof. □

The next corollary follows immediately from Theorem 2.4 in combination with Lemmas 2.5 and 2.7.

Corollary 2.8. *Let $\mathcal{I}$ be a set of multi-indices such that $\mathcal{I} \subseteq \{\boldsymbol{m} \in \mathbb{N}^d : m_1, \ldots, m_d \geq 3 \text{ odd}\}$. $\forall \boldsymbol{m} \in \mathcal{I}$, 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 $\mu := \sup_{\boldsymbol{m} \in \mathcal{I}} \rho(A_{\boldsymbol{m}}) < \infty$, that (2.13) holds with $\delta := \inf_{\boldsymbol{m} \in \mathcal{I}} \delta_{\boldsymbol{m}} > 0$, and that $0 < \omega < 2/\mu$. Then,*

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

where $a := \omega(2 - \omega\mu)$ and $\widetilde{b}_d$ is defined in Lemma 2.6.

2.4. Multigrid methods. In practice, the coarser linear system of the TGM could be too large to be solved directly. Hence, the third step in Algorithm 2.3 is usually replaced by one recursive call obtaining a multigrid (V-cycle) algorithm.

In the case where $A_{\boldsymbol{m}} = \tau_{\boldsymbol{m}}(z_d)$ and the projector $P_{\boldsymbol{m}}$ is defined in (2.8), then the coarser matrix is again a τ matrix generated by z_d up to a constant scaling. More precisely, fix the multi-indices $\boldsymbol{m}_0 := \boldsymbol{m} > \boldsymbol{m}_1 > \boldsymbol{m}_2 > \cdots > \boldsymbol{m}_l > 0$, where the inequalities are component-wise; take at each level the projector $P_{\boldsymbol{m}_i} \in \mathbb{R}^{\boldsymbol{m}_{i+1}\times \boldsymbol{m}_i}$; and define the coefficient matrix at the i-th level as $A_{\boldsymbol{m}_{i+1}} := P_{\boldsymbol{m}_i} A_{\boldsymbol{m}_i} P_{\boldsymbol{m}_i}^T$ for $i = 0, \ldots, l-1$. From the results in [32] (or by direct computation), we know that, if $P_{\boldsymbol{m}_i} := U_{\boldsymbol{m}_i} \tau_{\boldsymbol{m}_i}(q_d)$, $i = 0, \ldots, l-1$, then the coarser matrix is $A_{\boldsymbol{m}_i} = \tau_{\boldsymbol{m}_i}(r_i z_d)$, where r_i is a constant, for all coarser levels $i = 1, \ldots, l$.

Finally, we observe that the condition (2.11) is not sufficient to obtain the V-cycle optimality, see [2]. Nevertheless, it can be strengthened as follows

$$\limsup_{\boldsymbol{\theta}\to\boldsymbol{0}} \max_{\widehat{\boldsymbol{\theta}}\in\mathcal{M}(\boldsymbol{\theta})} \frac{q_d(\widehat{\boldsymbol{\theta}})}{z_d(\boldsymbol{\theta})} < \infty, \tag{2.14}$$

which leads to the V-cycle optimality according to the results in [1]. Unfortunately, Lemma 2.7 does not suffice to extend the optimality proof provided in [1] for matrix algebras to more general matrix structures. This could be a difficult task to be considered in a future research.

3. The 1D setting.

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

$$\begin{cases} -u'' + \beta u' + \gamma u = \mathrm{f}, & \text{in } (0,1), \\ u(0) = 0, \quad u(1) = 0, \end{cases} \tag{3.1}$$

with $\mathrm{f} \in L^2(0,1)$, $\beta \in \mathbb{R}$, $\gamma \geq 0$. In the framework of Galerkin B-spline IgA, we approximate the (weak) solution u of (3.1) in the space $\mathcal{W}$ of polynomial splines with maximal smoothness represented in the B-spline basis. More precisely, for $p \geq 1$ and $n \geq 2$, let

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

It is known that $\dim \mathcal{V}_n^{[p]} = n + p$ and $\dim \mathcal{W}_n^{[p]} = n + p - 2$. Then we choose $\mathcal{W} = \mathcal{W}_n^{[p]}$, for some $p \geq 1$ and $n \geq 2$, and the corresponding uniform B-spline basis $\{N_{2,[p]}, \ldots, N_{n+p-1,[p]}\}$ described in [23, Section 4]. With these choices, we obtain in (2.3) the stiffness matrix $A_n^{[p]} \in \mathbb{R}^{(n+p-2)\times(n+p-2)}$ such that

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

where

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

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

The above matrices have the following properties, see [23].

Lemma 3.1. *For every $p \geq 1$ and $n \geq 2$,*

- *$K_n^{[p]}$ is SPD and $\|K_n^{[p]}\|_\infty \leq 4p$;*
- *$H_n^{[p]}$ is skew-symmetric and $\|H_n^{[p]}\|_\infty \leq 2$;*
- *$M_n^{[p]}$ is SPD, $\|M_n^{[p]}\|_\infty \leq 1$ and $\exists\, C^{[p]} > 0$, depending only on p, such that $\lambda_{\min}(M_n^{[p]}) > C^{[p]}$.*

3.1. The symbol of the sequence $\{\frac{1}{n}A_n^{[p]}\}_n$. For $p \geq 0$, let $\phi_{[p]}$ be the cardinal B-spline of degree p over the uniform knot sequence $\{0, 1, \dots, p+1\}$, which is defined recursively as follows [9]:

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

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 $N_{i,[p]}(x)$, $i = p+1, \dots, n$, are cardinal B-splines, namely

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

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 $h_p : [-\pi, \pi] \to \mathbb{R}$,

$$h_0(\theta) := 1, \qquad h_p(\theta) := \phi_{[2p+1]}(p+1) + 2\sum_{k=1}^{p} \phi_{[2p+1]}(p+1-k)\cos(k\theta), \tag{3.2}$$

and, for $p \geq 1$, let $f_p : [-\pi, \pi] \to \mathbb{R}$,

$$f_p(\theta) := -\ddot{\phi}_{[2p+1]}(p+1) - 2\sum_{k=1}^{p} \ddot{\phi}_{[2p+1]}(p+1-k)\cos(k\theta). \tag{3.3}$$

It has been proved in [23, Theorem 12] that, for each fixed $p \geq 1$,

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

Hence, we can say that f_p is 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 both $\{\tau_{n+p-2}(f_p)\}_n$ and $\{T_{n+p-2}(f_p)\}_n$, see Section 2.2. The symbol f_p is independent of β and γ, and possesses the properties collected in Lemma 3.2, see [23, Section 3]. Recall that the modulus of the Fourier transform of the cardinal B-spline $\phi_{[p]}$ is given by

$$\left|\widehat{\phi_{[p]}}(\theta)\right|^2 = \left(\frac{2 - 2\cos\theta}{\theta^2}\right)^{p+1}. \tag{3.4}$$

Lemma 3.2. *The following properties hold for all $p \geq 1$ and $\theta \in [-\pi, \pi]$:*

- $f_p(\theta) = (2 - 2\cos\theta)h_{p-1}(\theta)$;
- $h_{p-1}(\theta) = \sum_{k\in\mathbb{Z}} \left|\widehat{\phi_{[p-1]}}(\theta + 2k\pi)\right|^2$;
- $\left(\frac{4}{\pi^2}\right)^p \leq h_{p-1}(\theta) \leq h_{p-1}(0) = 1.$

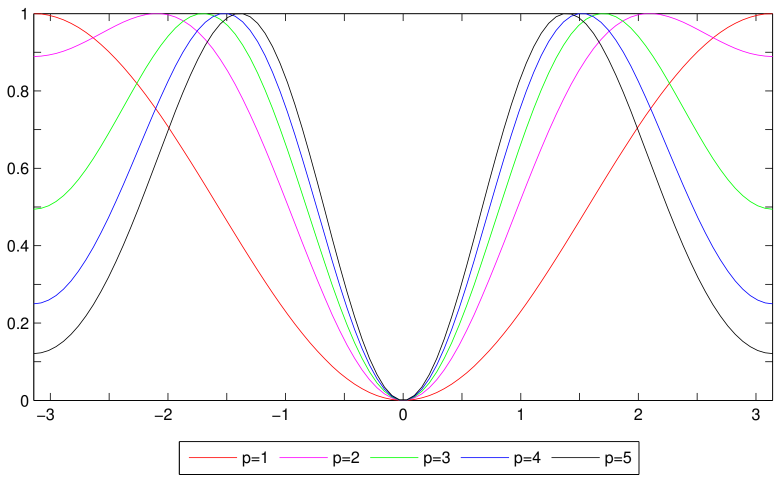

FIGURE 3.1. *Graph of f_p/M_{f_p} for $p = 1, \ldots, 5$.*

TABLE 3.1
Computation of $f_p(\pi)/M_{f_p}$ for $p = 1, \ldots, 9$.

p	1	2	3	4	5	6	7	8	9
$f_p(\pi)/M_{f_p}$	1.000	0.889	0.494	0.249	0.129	0.057	0.026	0.012	0.005

The properties in Lemma 3.2 have been proved in [23, Lemma 7 and Remark 2] for $p \geq 2$, but it can be checked that they also hold for $p = 1$. Moreover, we recall from [23] that h_p is the symbol of the sequence of matrices $\{M_n^{[p]}\}_n$.

Figure 3.1 shows the graph of f_p normalized by its maximum M_{f_p}, for $p = 1, \ldots, 5$. The value $f_p(\pi)/M_{f_p}$ decreases exponentially to zero as $p \to \infty$, see Table 3.1. This will be formally proved in Proposition 4.4. From a numerical viewpoint, we can say that, for large p, the normalized symbol f_p/M_{f_p} possesses two zeros over $[0, \pi]$: one at $\theta = 0$ and the other at the corresponding mirror point $\theta = \pi$. Because of this, and in view of the theoretical results in Section 4 (see Observation 7.2), we expect intrinsic difficulties, in particular a slow (though optimal) convergence rate, when solving for large p a linear system of the form $\frac{1}{n}A_n^{[p]}\mathbf{u} = \mathbf{b}$ by means of the two-grid method described in Section 3.2. Indeed, if $z_d(\theta)$ is replaced by a function which vanishes both at $\theta = 0$ and $\theta = \pi$, then the conditions (2.11) cannot be satisfied, independently of the choice of any different function instead of $q_d(\theta)$. Possible ways to overcome this problem are choosing a different size reduction at the lower level and/or adopting a multi-iterative strategy involving a variation of the smoothers. The first possibility was described in [17] for Toeplitz matrices, whereas the second one has been considered in [30]. Both approaches have been extensively numerically tested in [14]. A specialized multi-iterative strategy turned out to be the only optimal and totally robust solver, so we just focus on this strategy in Section 5.

3.2. Symbol-based construction and optimality of the TGM for $\frac{1}{n}A_n^{[p]}$. We now design a specific two-grid method for linear systems with coefficient matrix $\frac{1}{n}A_n^{[p]}$. Since the symbol f_p of the sequence $\{\frac{1}{n}A_n^{[p]}\}_n$ is known, we can adopt from [32] a sort of 'canonical two-grid procedure' for which we expect optimal convergence properties. The underlying idea is to treat $\frac{1}{n}A_n^{[p]}$ as if it were the τ-matrix $\tau_{n+p-2}(f_p)$ or the Toeplitz matrix $T_{n+p-2}(f_p)$ associated with the symbol f_p, and to design a two-grid method that has been proved in [32] to be optimal for both the sequences of τ-matrices and Toeplitz matrices.

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. Note that we require $n+p-2$ to be odd and greater than or equal to 3 in view of the definition of the projector $P_n^{[p]}$, see (3.7). We are looking for an optimal two-grid method for solving

$$\frac{1}{n}A_n^{[p]}\mathbf{u} = \mathbf{b}, \tag{3.5}$$

with $n \in \mathcal{I}_p$ and $\mathbf{b} \in \mathbb{R}^{n+p-2}$. As smoother we take the relaxed Richardson iteration,

$$S_n^{[p]} := I - \omega^{[p]}\frac{1}{n}A_n^{[p]}, \tag{3.6}$$

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

$$P_n^{[p]} := U_{n+p-2}\, \tau_{n+p-2}(1+\cos\theta), \tag{3.7}$$

as defined in (2.8) for $d = 1$ and $\boldsymbol{m} = n+p-2$, is the standard linear interpolation and it should be a good choice, because from Lemma 3.2 we know that $\theta = 0$ is a zero of f_p of order 2 and $f_p(\theta) > 0$ for all $\theta \in (0, \pi]$. Hence, $q_1(\theta) = 1 + \cos\theta$ in (2.8) satisfies the conditions (2.11) and if $\frac{1}{n}A_n^{[p]}$ were exactly $\tau_{n+p-2}(f_p)$ the TGM optimality follows from Lemma 2.6. Moreover, $q_1(\theta)$ satisfies also the condition (2.14) that leads to the V-cycle optimality for the matrix $\tau_{n+p-2}(f_p)$ according to the results in [2].

Assuming $\beta = \gamma = 0$, the matrix $\frac{1}{n}A_n^{[p]} = K_n^{[p]}$ is SPD (see Lemma 3.1). Under such an assumption and under suitable conditions on the relaxation parameter $\omega^{[p]}$, we now show that for $p = 1, 2, 3$, our method 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 $n \in \mathcal{I}_p$. We start with elaborating Corollary 2.8 in our context for general $p \geq 1$.

COROLLARY 3.3. *Assume that for a certain $p \geq 1$ it holds that*

$$\exists\, \delta^{[p]} > 0 : \; K_n^{[p]} \geq \delta^{[p]}\tau_{n+p-2}(2 - 2\cos\theta), \quad \forall n \geq 2. \tag{3.8}$$

Then, for any $\omega^{[p]} \in (0, 2/\mu^{[p]})$ with $\mu^{[p]} := \sup_{n\in\mathcal{I}_p} \rho(K_n^{[p]})$, 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 3.1 we know that $\mu^{[p]} \leq 4p$ for any $p \geq 1$, and if $\omega^{[p]} \in (0, 2/\mu^{[p]})$ then $\rho(S_n^{[p]}) < 1$, $\forall n \in \mathcal{I}_p$. In particular, we have $\mu^{[1]} = 4$ and $\mu^{[2]} \leq \frac{3}{2} + \frac{1+\sqrt{2}}{6}$, see [23, Eq. (79)]. Moreover, from our numerical experiments it seems that $\mu^{[2]} = 3/2$ and $\mu^{[3]} \leq 1.80$.

In the next theorem we prove that the condition (3.8) holds for $p = 1, 2, 3$.

THEOREM 3.4. *For $1 \leq p \leq 3$, the condition (3.8) is satisfied with $\delta^{[1]} = 1$, $\delta^{[2]} = 1/3$ and $\delta^{[3]} = 28/465$. Hence, 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 $n \in \mathcal{I}_p$, for $p = 1, 2, 3$.*

Proof. Since $K_n^{[1]} = \tau_{n-1}(2-2\cos\theta)$ for any $n \geq 2$, it is obvious that (3.8) holds for $p = 1$ with $\delta^{[1]} = 1$.

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

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

and one can check that the matrix $K_n^{[2]} - \delta\,\tau_n(2-2\cos\theta)$ is nonnegative definite for $\delta = 1/3$ and for all $n \geq 5$, thanks to the Gershgorin theorems [7]. Since it can be directly verified that $K_n^{[2]} \geq (1/3)\tau_n(2-2\cos\theta)$ for $n = 2,\dots,4$, we conclude that (3.8) holds for $p = 2$ with $\delta^{[2]} = 1/3$.

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

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

We first note that $f_3(\theta) = (\cos^2\theta + 13\cos\theta + 16)(2-2\cos\theta)/30 \geq (2/15)(2-2\cos\theta)$, $\forall\theta \in [-\pi,\pi]$, implying $\tau_m(f_3) \geq (2/15)\tau_m(2-2\cos\theta)$, $\forall m \geq 1$. Then, by the Gershgorin theorems we find that $K_n^{[3]} \geq \varepsilon\,\tau_{n+1}(f_3)$ is met for all $n \geq 8$ with $\varepsilon = 14/31$, and so $K_n^{[3]} \geq (28/465)\tau_{n+1}(2-2\cos\theta)$ holds for all $n \geq 8$. A direct verification shows that it also holds for $n = 2,\dots,7$. □

REMARK 3.5. *There are at least two reasons why the condition (3.8) is likely to be satisfied for all* $p \geq 1$.

1. *The condition (3.8) would hold if we had* $\tau_{n+p-2}(f_p)$ *instead of* $K_n^{[p]}$. *Indeed, from Lemma 3.2 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.*
2. *The matrices* $K_n^{[p]}$ *and* $\tau_{n+p-2}(2-2\cos\theta)$ *are both associated with particular approximations of the elliptic problem (3.1) in the case* $\beta = \gamma = 0$.

When multiplying (3.8) by $(\tau_{n+p-2}(2-2\cos\theta))^{-1/2}$ *on the left and the right, and observing that*

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

we obtain that (3.8) is equivalent to

$$\exists\,\delta^{[p]} > 0 :\ \lambda_{\min}((\tau_{n+p-2}(2-2\cos\theta))^{-1}K_n^{[p]}) \geq \delta^{[p]}, \quad \forall n \geq 2. \tag{3.9}$$

The inequality (3.9) is certainly satisfied for $p = 1$ (with $\delta^{[1]} = 1$), and numerical experiments reveal that (3.9) *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 3.4 coincides with $\delta^{[p],*}$ not only for $p = 1$ but also for $p = 2$.*

4. Symbol-based (local Fourier) analysis of $TG(S_n^{[p]}, P_n^{[p]})$. In order to simplify the discussion, throughout this section, we assume $\beta = \gamma = 0$ so that $\frac{1}{n}A_n^{[p]} = K_n^{[p]}$. By analyzing the spectral symbols involved, we are able to predict the behavior of $TG(S_n^{[p]}, P_n^{[p]})$. The idea is to think about the matrix $K_n^{[p]}$ as if it were the τ-matrix $\tau_{n+p-2}(f_p)$, since $K_n^{[p]}$ and $\tau_{n+p-2}(f_p)$ have the same spectral symbol, and, in this perspective, a detailed analysis of $TG(S_n^{[p]}, P_n^{[p]})$ can be performed. This is equivalent to the classical Local Fourier Analysis (LFA) for multigrid methods as proved in [13]. Nevertheless, this approach is more general since it can be applied also to linear systems that do not arise from an approximation of a partial differential equation.

In order to avoid confusion, we slightly change the notation and we set

$$\widetilde{TG}(\widetilde{S}_n^{[p]}, P_n^{[p]}) := \widetilde{S}_n^{[p]}\left(I - (P_n^{[p]})^T\left(P_n^{[p]}\tau_{n+p-2}(f_p)(P_n^{[p]})^T\right)^{-1}P_n^{[p]}\tau_{n+p-2}(f_p)\right), \tag{4.1}$$

with the smoother $\widetilde{S}_n^{[p]} := I - \omega^{[p]}\tau_{n+p-2}(f_p)$ and the projector $P_n^{[p]}$ as in (3.7). We analyze the two-grid scheme (4.1) and provide sharp lower and upper bounds for its spectral radius.

4.1. Analysis of $\widetilde{TG}(\widetilde{S}_n^{[p]}, P_n^{[p]})$ in the 1D case. The two-grid scheme (4.1) fits into the framework described in [19] and so we can adopt the results given there. Let us define $q(\theta) := 1 + \cos\theta$, and

$$\begin{aligned} s_p &: [-\pi,\pi] \to \mathbb{R}, & s_p(\theta) &:= 1 - \omega^{[p]}f_p(\theta),\\ t_p &: \left[0, \frac{\pi}{2}\right] \to \mathbb{R}, & t_p(\theta) &:= \frac{q^2(\theta)f_p(\theta)s_p(\pi-\theta) + q^2(\pi-\theta)f_p(\pi-\theta)s_p(\theta)}{q^2(\theta)f_p(\theta) + q^2(\pi-\theta)f_p(\pi-\theta)}. \end{aligned}$$

The function t_p is well-defined and continuous over $(0, \pi/2]$; it is also well-defined for $\theta = 0$ by continuous extension, i.e.,

$$\lim_{\theta\to 0} t_p(\theta) = \lim_{\theta\to 0} \frac{\dfrac{q^2(\theta)}{f_p(\pi-\theta)}s_p(\pi-\theta) + \dfrac{q^2(\pi-\theta)}{f_p(\theta)}s_p(\theta)}{\dfrac{q^2(\theta)}{f_p(\pi-\theta)} + \dfrac{q^2(\pi-\theta)}{f_p(\theta)}} = 1 - \omega^{[p]}f_p(\pi) =: t_p(0), \tag{4.2}$$

where we used the fact that $\lim_{\theta\to 0}\frac{q^2(\pi-\theta)}{f_p(\theta)} = 0$. Hence, t_p is well-defined and continuous over $[0, \pi/2]$.

From [19] we know that the eigenvalues of $\widetilde{TG}(\widetilde{S}_n^{[p]}, P_n^{[p]})$ are given by

$$\lambda_j(\widetilde{TG}(\widetilde{S}_n^{[p]}, P_n^{[p]})) = \begin{cases} t_p\left(\frac{j\pi}{n+p-1}\right), & \text{for } j = 1, \ldots, \frac{n+p-1}{2},\\ 0, & \text{for } j = \frac{n+p+1}{2}, \ldots, n+p-2. \end{cases} \tag{4.3}$$

Let us set $\widetilde{\rho}_n^{[p]} := \rho(\widetilde{TG}(\widetilde{S}_n^{[p]}, P_n^{[p]}))$ and $\widetilde{\rho}_\infty^{[p]} := \lim_{n\to\infty} \widetilde{\rho}_n^{[p]}$, then

$$\widetilde{\rho}_n^{[p]} = \max\left\{\left|t_p\left(\frac{j\pi}{n+p-1}\right)\right| : 1 \le j \le \frac{n+p-1}{2}\right\} \le \|t_p\|_\infty, \tag{4.4}$$

$$\widetilde{\rho}_\infty^{[p]} = \|t_p\|_\infty := \max_{\theta\in[0,\pi/2]} |t_p(\theta)|. \tag{4.5}$$

The smallest asymptotic spectral radius with respect to $\omega^{[p]}$ is denoted by

$$\widetilde{\rho}_\infty^{[p],*} := \min_{\omega^{[p]}\in\mathbb{R}} \widetilde{\rho}_\infty^{[p]}, \tag{4.6}$$

and the corresponding best value for $\omega^{[p]}$ (if unique) is denoted by $\widetilde{\omega}^{[p],*}$. For those $\omega^{[p]}$ for which $\|t_p\|_\infty < 1$, the formulas (4.4)–(4.5) imply that the method with iteration matrix $\widetilde{TG}(\widetilde{S}_n^{[p]}, P_n^{[p]})$ is optimal, because $\widetilde{\rho}_n^{[p]} \le \widetilde{\rho}_\infty^{[p]} < 1$. However, in the following we will see that for large values of p (in fact, even for moderate values of p such as $p = 6$) the asymptotic spectral radius $\widetilde{\rho}_\infty^{[p]}$ is very close to 1, independently of the choice of $\omega^{[p]}$. Actually, we will prove that $\widetilde{\rho}_\infty^{[p],*}$ converges exponentially to 1 as $p \to \infty$. By looking carefully at the proof of Proposition 4.5, we may conclude that this exponential convergence to 1 of $\widetilde{\rho}_\infty^{[p],*}$ is related to the exponential convergence to 0 of $f_p(\pi)/M_{f_p}$, which is shown in Proposition 4.4 (see also Figure 3.1 and Table 3.1).

REMARK 4.1. *From* (4.3) *we deduce that, for every* $F \in C_c(\mathbb{C})$,

$$\lim_{n\to\infty} \frac{1}{n+p-2} \sum_{j=1}^{n+p-2} F(\lambda_j(\widetilde{TG}(\widetilde{S}_n^{[p]}, P_n^{[p]}))) = \frac{1}{2}F(0) + \frac{1}{2}\cdot\frac{2}{\pi}\int_0^{\pi/2} F(t_p(\theta))\mathrm{d}\theta.$$

Hence, t_p *is 'almost' the spectral symbol of the sequence* $\{\widetilde{TG}(\widetilde{S}_n^{[p]}, P_n^{[p]}) : n \in \mathcal{I}_p\}$*; see also the general theory in [34, Section 3.7]. Moreover, it is clear that* s_p *is the symbol of the sequence of smoothers* $\{\widetilde{S}_n^{[p]}\}_n$*, because* $\widetilde{S}_n^{[p]} = I - \omega^{[p]}\tau_{n+p-2}(f_p) = \tau_{n+p-2}(s_p)$.

REMARK 4.2. *If* $0 < \omega^{[p]} < 2/M_{f_p}$*, then* $\|t_p\|_\infty < 1$ *and the method with iteration matrix* $\widetilde{TG}(\widetilde{S}_n^{[p]}, P_n^{[p]})$ *is optimal. Indeed, assume* $0 < \omega^{[p]} < 2/M_{f_p}$*. Then* $s_p(\theta) = 1 - \omega^{[p]} f_p(\theta) \in (-1, 1)$ *for* $\theta \in (0, \pi]$ *and* $s_p(0) = 1$*. As a consequence,* $t_p(\theta) \in (-1, 1)$ *for* $\theta \in (0, \pi/2]$*, being a weighted mean of* $s_p(\theta)$ *and* $s_p(\pi - \theta)$ *both belonging to* $(-1, 1)$*, and also* $t_p(0) = s_p(\pi) \in (-1, 1)$*. Thus,* $\|t_p\|_\infty < 1$*. Note that for each value* $\omega^{[p]} \in (0, 2/M_{f_p})$ *all the smoothers* $\widetilde{S}_n^{[p]}$*,* $n \in \mathcal{I}_p$*, are convergent, i.e.* $\rho(\widetilde{S}_n^{[p]}) < 1$ *for every* $n \in \mathcal{I}_p$*. On the contrary, for each value* $\omega^{[p]} \notin [0, 2/M_{f_p}]$ *all the smoothers* $\widetilde{S}_n^{[p]}$ *for* n *large enough are not convergent.*

REMARK 4.3. *If we had* $f_p(\pi) = 0$ *(this is the catastrophic situation in which the symbol* f_p *vanishes in two mirror points,* $\widehat{\theta} = 0$ *and* $\pi - \widehat{\theta} = \pi$*), then the two-grid scheme with iteration matrix* $\widetilde{TG}(\widetilde{S}_n^{[p]}, P_n^{[p]})$ *would not be optimal because we would have* $\widetilde{\rho}_\infty^{[p]} = \|t_p\|_\infty = 1 = t_p(0)$*, independently of* $\omega^{[p]}$.

4.2. Lower and upper bounds for $\widetilde{\rho}_\infty^{[p],*}$. To improve the readability of the present section, the proofs of the propositions are collected in Appendix A. We first provide a relation between the values $f_p\left(\frac{\pi}{2}\right)$ and $f_p(\pi)$.

PROPOSITION 4.4. *For every* $p \ge 1$ *we have* $f_p\left(\frac{\pi}{2}\right) = 2^{p-2} f_p(\pi)$.

A first consequence of Proposition 4.4 is that, when $p \to \infty$, the ratio $f_p(\pi)/M_{f_p}$ converges to 0 exponentially, as observed numerically in Figure 3.1 and Table 3.1:

$$\frac{f_p(\pi)}{M_{f_p}} = \frac{f_p(\pi)}{f_p\left(\frac{\pi}{2}\right)} \frac{f_p\left(\frac{\pi}{2}\right)}{M_{f_p}} \le \frac{1}{2^{p-2}}.$$

Furthermore, in view of Lemma 3.2, it follows that

$$h_p\left(\frac{\pi}{2}\right) = 2^p h_p(\pi). \tag{4.7}$$

The next two propositions give a lower bound and upper bound for $\widetilde{\rho}_\infty^{[p],*}$.

Proposition 4.5. *Let $p \geq 1$. Then, independently of the choice of $\omega^{[p]} \in \mathbb{R}$,*

$$\widetilde{\rho}_\infty^{[p]} \geq \frac{2^{p-2}-1}{2^{p-2}+1} =: \sigma^{[p]}.$$

In particular, $\widetilde{\rho}_\infty^{[p],*} \geq \sigma^{[p]}$.

Proposition 4.6. *Let $p \geq 1$, then*

$$\widetilde{\rho}_\infty^{[p],*} \leq \frac{2^{p+1}+1}{2^{p+1}+3} =: \varsigma^{[p]}.$$

It can be checked that

$$\lim_{p\to\infty} \frac{1-\varsigma^{[p]}}{1-\sigma^{[p]}} = \frac{1}{8}, \tag{4.8}$$

so the lower bound $\sigma^{[p]}$ and the upper bound $\varsigma^{[p]}$ converge to 1 with the same (exponential) asymptotic speed, implying that $\widetilde{\rho}_\infty^{[p],*}$ converges exponentially to 1 as well.

4.3. Numerical experiments and some conjectures. Table 4.1 summarizes the results of some numerical experiments for $p = 1, \ldots, 9$ with respect to the spectral radius $\widetilde{\rho}_\infty^{[p]}$ in terms of the parameter $\omega^{[p]}$. The second column provides the optimal value $\widetilde{\omega}^{[p],*}$, while the best asymptotic spectral radius

$$\widetilde{\rho}_\infty^{[p],*} = \min_{\omega^{[p]}\in\mathbb{R}} \widetilde{\rho}_\infty^{[p]} = \widetilde{\rho}_\infty^{[p]}|_{\omega^{[p]}=\widetilde{\omega}^{[p],*}}$$

is shown in the fourth column. Referring to Remark 4.2, we know that the choice $\omega^{[p]} \in (0, 2/M_{f_p})$ leads to an optimal scheme $\widetilde{TG}(\widetilde{S}_n^{[p]}, P_n^{[p]})$. We see that $\widetilde{\omega}^{[p],*}$ is not necessarily in the range $(0, 2/M_{f_p})$. This means that the method with iteration matrix $\widetilde{TG}(\widetilde{S}_n^{[p]}, P_n^{[p]})$ may be optimal (and even reach its better asymptotic convergence rate) with a value $\widetilde{\omega}^{[p],*}$ for which the smoothers $S_n^{[p]}$, $n \in \mathcal{I}_p$, are not convergent at all (see Remark 4.2). Finally, the last column illustrates the lower bound $\sigma^{[p]}$ given in Proposition 4.5.

The case $p = 2$ is somewhat peculiar and can be interpreted as a 'case of resonance', since $f_2(\pi) = f_2\left(\frac{\pi}{2}\right)$, see Proposition 4.4. As a consequence, $t_2(0) = t_2\left(\frac{\pi}{2}\right)$, the derivative $t_2'(\theta)$ vanishes at $\theta \in \left\{0,\ \alpha := \arccos\sqrt{-2+\sqrt{6}},\ \frac{\pi}{2}\right\}$, and

$$\widetilde{\rho}_\infty^{[2]} = \max\left(|t_2(0)| = \left|t_2\left(\frac{\pi}{2}\right)\right|, |t_2(\alpha)|\right) = \max\left(|1-\omega^{[2]}f_2(\pi)|, |t_2(\alpha)|\right). \tag{4.9}$$

Note that the best asymptotic spectral radius $\widetilde{\rho}_\infty^{[p],*}$ attains its smallest value $\frac{5-2\sqrt{6}}{9-2\sqrt{6}}$ precisely in the resonance case $p = 2$ (and not in the case $p = 1$).

Figure 4.1 shows $\widetilde{\rho}_\infty^{[p],*}$ together with the lower bound $\sigma^{[p]}$ and the upper bound $\varsigma^{[p]}$ varying $p = 1, \ldots, 9$. Note that the bounds become extremely accurate for increasing p and all quantities approach quickly one.

TABLE 4.1
Some numerical results related to $\widetilde{\rho}_{\infty}^{[p]}$ for $p = 1, \ldots, 9$.

p	$\widetilde{\omega}^{[p],*}$	$2/M_{f_p}$	$\widetilde{\rho}_{\infty}^{[p],*}$	$\sigma^{[p]}$
1	$\frac{1}{3} \approx 0.33$	0.5000	$\frac{1}{3} \approx 0.333$	$-\frac{1}{3}$
2	$\frac{3}{9-2\sqrt{6}} \approx 0.73$	1.3333	$\frac{5-2\sqrt{6}}{9-2\sqrt{6}} \approx 0.025$	0
3	$\frac{5}{4} = 1.25$	1.8531	$\frac{1}{3} \approx 0.333$	$\frac{1}{3}$
4	$\frac{63}{34} \approx 1.85$	2.3106	$\frac{3}{5} = 0.600$	$\frac{3}{5}$
5	$\frac{315}{124} \approx 2.54$	2.7639	$\frac{7}{9} \approx 0.778$	$\frac{7}{9}$
6	$\frac{155925}{46988} \approx 3.32$	3.2169	$\frac{15}{17} \approx 0.882$	$\frac{15}{17}$
7	$\frac{184275}{43688} \approx 4.22$	3.6699	$\frac{31}{33} \approx 0.939$	$\frac{31}{33}$
8	$\frac{9823275}{1859138} \approx 5.28$	4.1229	$\frac{63}{65} \approx 0.969$	$\frac{63}{65}$
9	$\frac{3618239625}{550794052} \approx 6.57$	4.5760	$\frac{127}{129} \approx 0.984$	$\frac{127}{129}$

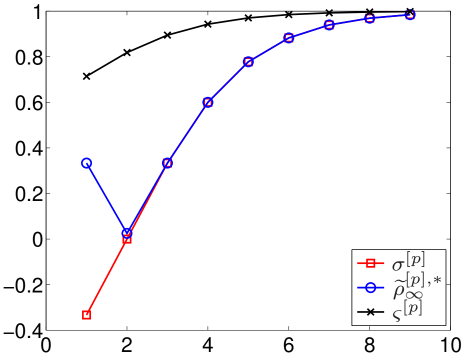

FIGURE 4.1. $\widetilde{\rho}_{\infty}^{[p],*}$ *with the lower bound* $\sigma^{[p]}$ *and the upper bound* $\varsigma^{[p]}$ *varying* $p = 1, \ldots, 9$.

From these results we can formulate the following conjecture:

$$\sigma^{[p]} = \widetilde{\rho}_{\infty}^{[p],*}, \qquad \forall p \geq 3. \tag{4.10}$$

This conjecture is verified in Table 4.1 for $p = 3, \ldots, 9$. Actually, supported by additional numerical experiments, we can formulate a stronger conjecture than (4.10), which has been deferred to Appendix B.

5. Multi-iterative method: multigrid with PCG. Despite its optimality, the basic method $TG(S_n^{[p]}, P_n^{[p]})$ suffers from a 'pathology': its convergence rate rapidly worsens when p increases. This phenomenon can be explained as follows. Due to the projector $P_n^{[p]}$, the method $TG(S_n^{[p]}, P_n^{[p]})$ uses a reduction strategy with reduction factor 2, meaning that the system at the coarse level is of size one half of the system at the fine level. In this way, the mirror point of $\theta = 0$ (i.e. the zero of f_p) is $\theta = \pi$. When p is large, $\theta = \pi$ is a numerical zero of the normalized symbol f_p/M_{f_p} (see Figure 3.1 and Proposition 4.4), and so f_p/M_{f_p} essentially possesses two zeros: one in $\theta = 0$ and the other in the corresponding mirror point $\theta = \pi$. This leads to a slow convergence of $TG(S_n^{[p]}, P_n^{[p]})$.

To overcome this problem, following the multi-iterative idea [30], we propose to use PCG as smoother, whose preconditioner takes care of dampening the 'high frequencies' corresponding to values of θ near $\theta = \pi$. A similar strategy has been employed in [8] to deal with a rank deficient projector.

Let $\beta = 0$. Under this assumption, $\frac{1}{n} A_n^{[p]} = K_n^{[p]} + \frac{\gamma}{n^2} M_n^{[p]}$ is SPD (see Lemma 3.1) and the PCG method can be applied to it. For the case $\beta \neq 0$ (not considered here), we simply suggest to replace the PCG method with the Preconditioned GMRES (P-GMRES) method, where we use for P-GMRES the same preconditioner that we are going to devise for PCG, see [14].

5.1. Two-grid with PCG. As mentioned above, the idea for improving the convergence rate of $TG(S_n^{[p]}, P_n^{[p]})$ for large p is the following: we substitute, in Algorithm 2.3, the single smoothing iteration by $S_n^{[p]}$, with a few smoothing iterations (say $s^{[p]}$ iterations) by the PCG method, using the SPD preconditioner $T_{n+p-2}(h_{p-1})$. Due to the presence of the PCG smoother, the resulting method is no longer a stationary iterative method and hence it is not a two-grid, in the classical sense. However, using an expressive notation, we denote this method by $TG((PCG)^{s^{[p]}}, P_n^{[p]})$, where the exponent $s^{[p]}$ simply indicates that we apply $s^{[p]}$ steps of the PCG method and it is assumed that the preconditioner is $T_{n+p-2}(h_{p-1})$.

We now motivate our choice of the preconditioner $T_{n+p-2}(h_{p-1})$. First, we recall from Lemma 3.2 that the function h_{p-1} appears in the factorization of the symbol $f_p(\theta) = (2 - 2\cos\theta)h_{p-1}(\theta)$. Since $2 - 2\cos\theta$ is monotone increasing over $[0, \pi]$, the factor h_{p-1} is responsible for the exponential convergence to 0 of $f_p(\pi)/M_{f_p}$ (which is proved in Section 4.2). The idea of using the preconditioner $T_{n+p-2}(h_{p-1})$ as smoother is then a way to 'erase' the numerical zero $\theta = \pi$ of the normalized symbol f_p/M_{f_p}. Moreover, we notice that such PCG works exactly as a smoother reducing the error in the high frequencies, as requested by an efficient multigrid algorithm for differential operators. In other words, the eigenvalues of the preconditioned matrices $T_{n+p-2}^{-1}(h_{p-1})K_n^{[p]}$ and $T_{n+p-2}^{-1}(h_{p-1})\frac{1}{n}A_n^{[p]}$ behave like a uniform sampling of the standard Finite Difference Laplacian symbol $2-2\cos\theta$ for n large enough, as formally stated in the next theorem.

THEOREM 5.1. *The symbol of the sequences of preconditioned matrices*

$$\{T_{n+p-2}^{-1}(h_{p-1})K_n^{[p]}\}_n \qquad \textit{and} \qquad \{T_{n+p-2}^{-1}(h_{p-1})\frac{1}{n}A_n^{[p]}\}_n$$

is $2 - 2\cos\theta$, *i.e.,*

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

with the matrix $X_n^{[p]}$ *being either* $T_{n+p-2}^{-1}(h_{p-1})K_n^{[p]}$ *or* $T_{n+p-2}^{-1}(h_{p-1})\frac{1}{n}A_n^{[p]}$.

Proof. We heavily rely on the results on Generalized Locally Toeplitz (GLT) sequences [33, 34] and the seminal paper by Tilli [36]:

1. each GLT sequence of Hermitian matrices has a symbol, and the eigenvalues of the sequence are spectrally distributed as the symbol;
2. each Toeplitz sequence with Lebesgue integrable symbol is a GLT sequence with the same symbol;
3. the product of GLT sequences is a GLT sequence whose symbol is the product of the symbols, the inverse of a GLT sequence is a GLT sequence as long as the symbol has, at most, a set of zeros of zero Lebesgue measure;

4. if $\{G_n\}_n$ is a GLT sequence with symbol f, $\{E_n\}_n$ is a sequence of infinitesimal spectral norm with respect to the size n of the matrix, and $\{R_n\}_n$ is such that $\mathrm{rank}(R_n)/n$ is infinitesimal as n tends to infinity, then $\{G_n+E_n+R_n\}_n$ is a GLT sequence with symbol f.

We first deal with the pure Hermitian case, i.e. $\beta=0$. From the second item, we know that $\{T_{n+p-2}(h_{p-1})\}_n$ is a GLT sequence with symbol h_{p-1}. Moreover, due to the fourth item, we infer that $\{K_n^{[p]}\}_n$ is a GLT sequence with symbol $f_p(\theta) = (2-2\cos\theta)h_{p-1}(\theta)$, because the rank of $K_n^{[p]} - T_{n+p-2}(f_p)$ is bounded by a constant depending on p but independent of n, see [23]. By using again item 4 and the fact that $\{K_n^{[p]}\}_n$ is a GLT sequence with symbol f_p, we obtain that $\{\frac{1}{n}A_n^{[p]}\}_n$ is a GLT sequence with the same symbol f_p, because the norm of $\frac{1}{n}A_n^{[p]} - K_n^{[p]}$ is infinitesimal as n tends to infinity. Finally, by invoking items 1 and 3, and by employing a symmetrization trick ($X_n^{[p]}$ is not Hermitian, but it is similar to a Hermitian matrix in both cases), the desired results follow.

When β is nonzero, due to the small norm of the non-symmetric term, it is enough to invoke the perturbation arguments in [24]. □

We are now ready to give an LFA interpretation:

- the preconditioned matrix is spectrally equivalent to the standard Finite Difference Laplacian with symbol $2-2\cos\theta$;
- the projector $P_n^{[p]}$ takes care of the zero $\theta=0$, by reconstructing the error $\mathbf{e}$ at the coarse level in the subspace generated by the low frequencies, corresponding to values of θ near 0 (note that the polynomial $1+\cos\theta$ associated with $P_n^{[p]}$ attains its maximum at $\theta=0$).

Finally, we remark that the proposed preconditioner $T_{n+p-2}(h_{p-1})$ can be interpreted as a small rank perturbation of the (normalized) B-spline mass matrix $M_{n+1}^{[p-1]}$ related to the fineness parameter $n+1$ and the spline degree $p-1$, see [23].

5.2. V-cycle with PCG. Inspired by the specialized two-grid method designed in Section 5.1, we now develop an effective V-cycle multigrid method. For the sake of simplicity, we just focus on the case where $\beta=\gamma=0$, so $\frac{1}{n}A_n^{[p]} = K_n^{[p]}$.

The finest level is indicated by index 0 and the coarsest level by $\ell_n^{[p]} := \log_2(n+p-1)-1$, assuming that $n+p-1$ is a power of 2. Let $K_{n,i}^{[p]}$ be the matrix at level i and let $m_{n,i}^{[p]}$ denote its dimension, $0\le i\le \ell_n^{[p]}$. In this notation, we have $K_{n,0}^{[p]} := K_n^{[p]}$,

$$K_{n,i+1}^{[p]} := P_{n,i}^{[p]}\,K_{n,i}^{[p]}\,(P_{n,i}^{[p]})^T, \quad i=0,\ldots,\ell_n^{[p]}-1,$$

and $K_{n,\ell_n^{[p]}}^{[p]}$ has dimension 1. In the above expression,

$$P_{n,i}^{[p]} := P_{m_{n,i}^{[p]}}, \quad i=0,\ldots,\ell_n^{[p]}-1,$$

is the projector employed at level i, defined by (2.8) for $d=1$ and $\boldsymbol{m} = m_{n,i}^{[p]}$. Given the structure of $P_{m_{n,i}^{[p]}}$, one can show by induction on i that $m_{n,i+1}^{[p]} = (m_{n,i}^{[p]}-1)/2$, $i=0,\ldots,\ell_n^{[p]}-1$, and $m_{n,i}^{[p]} = \frac{n+p-1}{2^i}-1$, $i=0,\ldots,\ell_n^{[p]}$. Regarding the smoother, at each coarser level $i\ge 1$ we choose the simple Gauss-Seidel smoother. On the other hand, at the finest level $i=0$ we apply $s^{[p]}$ smoothing iterations by the PCG method with preconditioner $T_{n+p-2}(h_{p-1})$. At each level i, we first perform a coarse-grid correction, with one recursive call and then we apply one Gauss-Seidel smoothing iteration (if $i \geq 1$) or $s^{[p]}$ smoothing iterations by the proposed PCG (if $i = 0$).

We point out that the choice of the projector $P_{n,i}^{[p]}$ at each level i has the same motivation as the projector $P_n^{[p]}$ for $K_n^{[p]}$. Indeed, referring to [32, Proposition 2.2] or [2, Proposition 2.5], if $K_n^{[p]}$ were exactly $\tau_{m_{n,0}^{[p]}}(f_{p,0}) := \tau_{n+p-2}(f_p)$, then $K_{n,i}^{[p]}$ would be exactly $\tau_{m_{n,i}^{[p]}}(f_{p,i})$, with the symbol $f_{p,i}$ at level i sharing the same properties of the symbol $f_{p,0} := f_p$ at level 0: $f_{p,i}(0) = 0$, with $\theta = 0$ a zero of order two, and $f_{p,i}(\theta) > 0$ for all $\theta \in [-\pi, \pi]\backslash\{0\}$. These properties coincide with those of f_p used in Section 3.2 for devising the appropriate projector $P_n^{[p]}$ for $K_n^{[p]}$. Furthermore, Lemma 3.2 ensures that a certain modification of our V-cycle, considered and analyzed in [2], would be optimal when applied to $\tau_{n+p-2}(f_p)$ instead of $K_n^{[p]}$. Finally, we want to motivate why the $s^{[p]}$ PCG smoothing steps are only used at the finest level. Let $M_{f_{p,i}} := \max_\theta f_{p,i}(\theta)$. Referring again to [32, Proposition 2.2 (item 2)], and taking into account also some additional numerical experiments that we performed, it seems that the numerical zero $\theta = \pi$ of $f_{p,0}/M_{f_{p,0}}$ disappears for $i \geq 1$, and each $f_{p,i}/M_{f_{p,i}}$, $i \geq 1$, only possesses the actual zero $\theta = 0$. Hence, a single smoothing iteration by the simple Gauss-Seidel method is all we need at the coarser levels $i \geq 1$.

6. The 2D setting. In this section we focus on our model problem (1.1) in the case $d = 2$ with $\Omega = (0,1)^2$, and we perform the same study as in Section 3. Although the argumentation in the 2D case follows more or less the same pattern as in the 1D case, we will briefly describe it, both for the sake of completeness and for illustrating the strict analogies between the 1D and 2D setting. Given any two functions $f, g : [a,b] \to \mathbb{R}$, we denote by $f \otimes g$ the tensor-product function

$$f \otimes g : [a,b]^2 \to \mathbb{R}, \quad (f \otimes g)(x,y) := f(x)g(y). \tag{6.1}$$

We now approximate the weak solution u of (1.1) by means of the approximation space $\mathcal{W}$ chosen as a space spanned by tensor-product B-splines. More precisely, we choose $\mathcal{W} = \mathcal{W}_{n_1,n_2}^{[p_1,p_2]}$, for some $p_1, p_2 \geq 1$, $n_1, n_2 \geq 2$, where

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

and $N_{j,[p]}$ are the basis functions used in Section 3 and defined in [23, Section 4]. We choose the tensor-product B-spline basis $\{N_{j_1,[p_1]} \otimes N_{j_2,[p_2]} : j_1 = 2, \ldots, n_1 + p_1 - 1, \ j_2 = 2, \ldots, n_2 + p_2 - 1\}$ ordered in the same way as considered in [23, Eq. (85)], namely

$$\Big[\big[N_{j_1,[p_1]} \otimes N_{j_2,[p_2]} \big]_{j_1=2,\ldots,n_1+p_1-1} \Big]_{j_2=2,\ldots,n_2+p_2-1}.$$

Then, we obtain in (2.3) the following stiffness matrix, see [23, Section 5.1]:

$$A_{n_1,n_2}^{[p_1,p_2]} := K_{n_1,n_2}^{[p_1,p_2]} + \frac{\beta_1}{n_2} M_{n_2}^{[p_2]} \otimes H_{n_1}^{[p_1]} + \frac{\beta_2}{n_1} H_{n_2}^{[p_2]} \otimes M_{n_1}^{[p_1]} + \frac{\gamma}{n_1 n_2} M_{n_2}^{[p_2]} \otimes M_{n_1}^{[p_1]},$$

where

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

and the matrices $K_n^{[p]}$, $H_n^{[p]}$, $M_n^{[p]}$ are defined for all $p \geq 1$ and $n \geq 2$ in Section 3.

REMARK 6.1. *By Lemma 3.1 and by the fact that $X \otimes Y$ is SPD whenever X, Y are SPD, we know that $A_{n_1,n_2}^{[p_1,p_2]}$ is SPD for all $p_1, p_2 \geq 1$ and $n_1, n_2 \geq 2$, provided that $\beta_1 = \beta_2 = 0$.*

6.1. The symbol of the sequence $\{A^{[p_1,p_2]}_{\nu_1 n,\nu_2 n}\}_n$. For $p_1, p_2 \ge 1$ and $\nu_1, \nu_2 \in \mathbb{Q}_+ := \{r \in \mathbb{Q} : \ r > 0\}$, we define the function

$$f^{(\nu_1,\nu_2)}_{p_1,p_2} : [-\pi,\pi]^2 \to \mathbb{R}, \quad f^{(\nu_1,\nu_2)}_{p_1,p_2} := \frac{\nu_1}{\nu_2} h_{p_2} \otimes f_{p_1} + \frac{\nu_2}{\nu_1} f_{p_2} \otimes h_{p_1}, \tag{6.2}$$

see (3.2)–(3.3) for the definition of h_p and f_p. From now on we always assume that $n \in \mathbb{N}$ is chosen such that $\nu_1 n, \nu_2 n \in \mathbb{N}$. Consider the sequence of matrices

$$A^{[p_1,p_2]}_{\nu_1 n,\nu_2 n} = K^{[p_1,p_2]}_{\nu_1 n,\nu_2 n} + \frac{\beta_1}{\nu_2 n} M^{[p_2]}_{\nu_2 n} \otimes H^{[p_1]}_{\nu_1 n} + \frac{\beta_2}{\nu_1 n} H^{[p_2]}_{\nu_2 n} \otimes M^{[p_1]}_{\nu_1 n} + \frac{\gamma}{\nu_1 \nu_2 n^2} M^{[p_2]}_{\nu_2 n} \otimes M^{[p_1]}_{\nu_1 n},$$

with n varying in the set of indices where $\nu_1 n \ge 2$ and $\nu_2 n \ge 2$. It was proved in [23, Section 5.2] that, $\forall F \in C_c(\mathbb{C})$,

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

with $N := (\nu_1 n + p_1 - 2)(\nu_2 n + p_2 - 2)$, and so $f^{(\nu_1,\nu_2)}_{p_1,p_2}$ is the symbol of the sequence $\{A^{[p_1,p_2]}_{\nu_1 n,\nu_2 n}\}_n$. The symbol $f^{(\nu_1,\nu_2)}_{p_1,p_2}$ is independent of $\boldsymbol{\beta} := (\beta_1, \beta_2)$ and γ. Moreover, it possesses the following properties (consequences of Lemma 3.2).

Lemma 6.2. *Let $p_1, p_2 \ge 1$ and $\nu_1, \nu_2 \in \mathbb{Q}_+$. Then, $\forall(\theta_1,\theta_2) \in [-\pi,\pi]^2$,*

$$f^{(\nu_1,\nu_2)}_{p_1,p_2}(\theta_1,\theta_2) \ge \left(\frac{4}{\pi^2}\right)^{p_1+p_2+1} \min\left(\frac{\nu_2}{\nu_1}, \frac{\nu_1}{\nu_2}\right) (4 - 2\cos\theta_1 - 2\cos\theta_2),$$
$$f^{(\nu_1,\nu_2)}_{p_1,p_2}(\theta_1,\theta_2) \le \max\left(\frac{\nu_2}{\nu_1}, \frac{\nu_1}{\nu_2}\right) (4 - 2\cos\theta_1 - 2\cos\theta_2).$$

Let $M_{f^{(\nu_1,\nu_2)}_{p_1,p_2}} := \max_{\boldsymbol{\theta}\in[0,\pi]^2} f^{(\nu_1,\nu_2)}_{p_1,p_2}(\boldsymbol{\theta})$. By Lemma 6.2, the normalized symbol $f^{(\nu_1,\nu_2)}_{p_1,p_2}/M_{f^{(\nu_1,\nu_2)}_{p_1,p_2}}$ has only one (theoretical) zero at $\boldsymbol{\theta} = \mathbf{0}$. However, when p_1, p_2 are large, it also has infinitely many 'numerical zeros' over $[0,\pi]^2$, located at the edge points

$$\{(\theta_1,\pi) : \ 0 \le \theta_1 \le \pi\} \cup \{(\pi,\theta_2) : \ 0 \le \theta_2 \le \pi\}. \tag{6.3}$$

Indeed, by (6.2) and (6.1), and recalling equation (4.7) and Proposition 4.4, we have

$$\begin{aligned} f^{(\nu_1,\nu_2)}_{p_1,p_2}(\theta_1,\pi) &= \frac{\nu_1}{\nu_2} h_{p_2}(\theta_1) f_{p_1}(\pi) + \frac{\nu_2}{\nu_1} f_{p_2}(\theta_1) h_{p_1}(\pi) \\ &= \frac{\nu_1}{\nu_2} h_{p_2}(\theta_1) \frac{f_{p_1}(\pi/2)}{2^{p_1-2}} + \frac{\nu_2}{\nu_1} f_{p_2}(\theta_1) \frac{h_{p_1}(\pi/2)}{2^{p_1}} \le \frac{1}{2^{p_1-2}} f^{(\nu_1,\nu_2)}_{p_1,p_2}(\theta_1,\pi/2) \\ &\le \frac{1}{2^{p_1-2}} M_{f^{(\nu_1,\nu_2)}_{p_1,p_2}}, \end{aligned}$$

and similarly $f^{(\nu_1,\nu_2)}_{p_1,p_2}(\pi,\theta_2) \le \frac{1}{2^{p_2-2}} M_{f^{(\nu_1,\nu_2)}_{p_1,p_2}}$. Because of this unpleasant property, the two-grid schemes that we are going to devise for the matrix $A^{[p_1,p_2]}_{\nu_1 n,\nu_2 n}$ are expected to show a bad (though optimal) convergence rate when p_1, p_2 are large. A possible way to overcome this problem is discussed in Section 6.3 and consists in adopting a multi-iterative strategy involving a specialized PCG smoother, just as we have seen in Section 5 for the 1D case.

6.2. Symbol-based construction and optimality of the TGM for $A^{[p_1,p_2]}_{\nu_1 n,\nu_2 n}$. We now develop a specialized two-grid method for linear systems having $A^{[p_1,p_2]}_{\nu_1 n,\nu_2 n}$ as matrix. To this end, we are going to follow a recipe analogous to the one in Section 3.2. In particular, we will exploit specific properties of the symbol $f^{(\nu_1,\nu_2)}_{p_1,p_2}$ in order to choose an appropriate projector. The underlying idea is again to treat $A^{[p_1,p_2]}_{\nu_1 n,\nu_2 n}$ as if it were the two-level τ-matrix $\tau_{\nu_2 n+p_2-2,\,\nu_1 n+p_1-2}(f^{(\nu_1,\nu_2)}_{p_1,p_2})$ or the two-level Toeplitz matrix $T_{\nu_2 n+p_2-2,\,\nu_1 n+p_1-2}(f^{(\nu_1,\nu_2)}_{p_1,p_2})$, and to design a two-grid (multigrid) method that has been proved to be optimal for both the sequences of two-level τ-matrices and Toeplitz matrices in the two-grid case [32], while the proof of optimality of the V-cycle is known only for multilevel τ-matrices [1].

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 $\mathcal{I}^{(\nu_1,\nu_2)}_{p_1,p_2} \subseteq \{n : \nu_1 n \geq 2, \nu_2 n \geq 2, \nu_1 n + p_1 - 2 \geq 3 \text{ odd}, \nu_2 n + p_2 - 2 \geq 3 \text{ odd}\}$, $\#\mathcal{I}^{(\nu_1,\nu_2)}_{p_1,p_2} = \infty$. Note that we require $\nu_1 n + p_1 - 2$ and $\nu_2 n + p_2 - 2$ to be odd and ≥ 3 in view of the definition of the projector, see (6.5). We are looking for solving $A^{[p_1,p_2]}_{\nu_1 n,\nu_2 n}\mathbf{u} = \mathbf{b}$, with $n \in \mathcal{I}^{(\nu_1,\nu_2)}_{p_1,p_2}$ and $\mathbf{b} \in \mathbb{R}^{(\nu_1 n+p_1-2)(\nu_2 n+p_2-2)}$.

Like in the 1D setting, we take the relaxed Richardson iteration as smoother,

$$S^{[p_1,p_2]}_{\nu_1 n,\nu_2 n} := I - \omega^{[p_1,p_2,\nu_1,\nu_2]} A^{[p_1,p_2]}_{\nu_1 n,\nu_2 n}, \tag{6.4}$$

where $\omega^{[p_1,p_2,\nu_1,\nu_2]}$ is the relaxation parameter (independent of n). For the choice of the projector, we exploit certain properties of $f^{(\nu_1,\nu_2)}_{p_1,p_2}$, together with the suggestions coming from [20]. From Lemma 6.2 we know that $f^{(\nu_1,\nu_2)}_{p_1,p_2}(0,0) = 0$ and $f^{(\nu_1,\nu_2)}_{p_1,p_2}(\theta_1,\theta_2) > 0$ for all $(\theta_1,\theta_2) \in [0,\pi]^2\backslash\{(0,0)\}$. Therefore, we look for a bivariate cosine trigonometric polynomial (possibly depending on p_1, p_2, ν) that vanishes at the mirror points of $(0,0)$, i.e. at $\{(\pi,0),(0,\pi),(\pi,\pi)\}$, and satisfies the conditions (2.11). The simple choice $q_2(\theta_1,\theta_2) = (1+\cos\theta_1)(1+\cos\theta_2)$ satisfies all these requirements for all p_1 and p_2. Hence, we choose the projector

$$P^{[p_1,p_2]}_{\nu_1 n,\nu_2 n} := U_{\nu_2 n+p_2-2,\nu_1 n+p_1-2}\,\tau_{\nu_2 n+p_2-2,\nu_1 n+p_1-2}((1+\cos\theta_1)(1+\cos\theta_2)), \tag{6.5}$$

as defined in (2.8) for $d = 2$ and $\boldsymbol{m} = (\nu_2 n + p_2 - 2, \nu_1 n + p_1 - 2)$. Note that if $A^{[p_1,p_2]}_{\nu_1 n,\nu_2 n}$ were $\tau_{\nu_2 n+p_2-2,\nu_1 n+p_1-2}(f^{(\nu_1,\nu_2)}_{p_1,p_2})$ the projector (6.5) ensures the optimality of the two-grid method [32] and of the V-cycle [1].

Assuming $\beta_1 = \beta_2 = \gamma = 0$, the matrix $A^{[p_1,p_2]}_{\nu_1 n,\nu_2 n} = K^{[p_1,p_2]}_{\nu_1 n,\nu_2 n}$ is SPD, see Remark 6.1. In the bilinear case $p_1 = p_2 = 1$, it can be shown that $K^{[1,1]}_{\nu_1 n,\nu_2 n} = \tau_{\nu_2 n-1,\nu_1 n-1}(f^{(\nu_1,\nu_2)}_{1,1})$. So, the eigenvalues of $K^{[1,1]}_{\nu_1 n,\nu_2 n}$ are given by $f^{(\nu_1,\nu_2)}_{1,1}\left(\frac{j_2\pi}{\nu_2 n}, \frac{j_1\pi}{\nu_1 n}\right)$, $j_2 = 1,\ldots,\nu_2 n - 1$, $j_1 = 1,\ldots,\nu_1 n - 1$, and

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

Therefore, for any choice of $\omega^{[1,1,\nu_1,\nu_2]} \in \left(0, 2/\mu^{[1,1,\nu_1,\nu_2]}\right)$ and $\nu_1, \nu_2 \in \mathbb{Q}_+$, the optimality of the method 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 [32] and that of the V-cycle in [1]. More generally, for $1 \leq p_1, p_2 \leq 3$ and $\nu_1, \nu_2 \in \mathbb{Q}_+$, we will show that the two-grid scheme with iteration matrix $TG(S^{[p_1,p_2]}_{\nu_1 n,\nu_2 n}, P^{[p_1,p_2]}_{\nu_1 n,\nu_2 n})$ and

$\beta_1 = \beta_2 = \gamma = 0$ 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 $\mu^{[p_1,p_2,\nu_1,\nu_2]} := \sup_{n\in\mathcal{I}^{(\nu_1,\nu_2)}_{p_1,p_2}} \rho(K^{[p_1,p_2]}_{\nu_1 n,\nu_2 n})$. Note that from Lemma 3.1 we know that $\forall n \geq 2$,

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

where we used the fact that, whenever X, Y are normal matrices, $\|X \otimes Y\|_2 = \|X\|_2\|Y\|_2$ and $\|X\|_2 = \rho(X) \leq \|X\|_\infty$.

In our 2D context, the condition (2.13) reads as

(6.6)
$$\exists\, \delta^{[p_1,p_2,\nu_1,\nu_2]} > 0 : \; K^{[p_1,p_2]}_{\nu_1 n,\nu_2 n} \geq \delta^{[p_1,p_2,\nu_1,\nu_2]}\, \tau_{\nu_2 n+p_2-2,\nu_1 n+p_1-2}(4 - 2\cos\theta_1 - 2\cos\theta_2).$$

Since the two-grid optimality follows from Corollary 2.8, in the next theorem we show that the condition (6.6) holds for $1 \leq p_1, p_2 \leq 3$.

Theorem 6.3. *Let $1 \leq p_1, p_2 \leq 3$. Then, (6.6) 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 $C^{[p]}$, $p \geq 1$, is given in Lemma 3.1 and $\delta^{[p]}$, $1 \leq p \leq 3$, is specified in Theorem 3.4. Hence, the scheme 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 $1 \leq p_1, p_2 \leq 3$ 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 $X \geq X'$ and $Y \geq Y'$, then $X \otimes Y$ and $X' \otimes Y'$ are SPD with $X \otimes Y \geq X' \otimes Y'$. Hence, for every $\nu_1 n, \nu_2 n \geq 2$ integer, from Theorem 3.4 we deduce

$$\begin{aligned}K^{[p_1,p_2]}_{\nu_1 n,\nu_2 n} &= \frac{\nu_1}{\nu_2}M^{[p_2]}_{\nu_2 n} \otimes K^{[p_1]}_{\nu_1 n} + \frac{\nu_2}{\nu_1}K^{[p_2]}_{\nu_2 n} \otimes M^{[p_1]}_{\nu_1 n}\\ &\geq \frac{\nu_1}{\nu_2}C^{[p_2]}I_{\nu_2 n+p_2-2} \otimes \delta^{[p_1]}\tau_{\nu_1 n+p_1-2}(2 - 2\cos\theta_1)\\ &\qquad\qquad + \frac{\nu_2}{\nu_1}\,\delta^{[p_2]}\tau_{\nu_2 n+p_2-2}(2 - 2\cos\theta_2) \otimes C^{[p_1]}I_{\nu_1 n+p_1-2}\\ &\geq \delta^{[p_1,p_2,\nu_1,\nu_2]}\,\tau_{\nu_2 n+p_2-2,\nu_1 n+p_1-2}(4 - 2\cos\theta_1 - 2\cos\theta_2).\end{aligned}$$

□

6.3. Multigrid with PCG. With the aim of accelerating the convergence rate of $TG(S^{[p_1,p_2]}_{\nu_1 n,\nu_2 n}, P^{[p_1,p_2]}_{\nu_1 n,\nu_2 n})$ for large p_1, p_2, we propose to substitute in Algorithm 2.3 the single relaxation iteration (step 6) with a few smoothing iterations (say $s^{[p_1,p_2]}$ iterations) by the PCG method using as preconditioner

$$T_{\nu_2 n+p_2-2,\nu_1 n+p_1-2}(h_{p_2-1} \otimes h_{p_1-1}) = T_{\nu_2 n+p_2-2}(h_{p_2-1}) \otimes T_{\nu_1 n+p_1-2}(h_{p_1-1}). \tag{6.7}$$

We denote the resulting multi-iterative method by $TG((PCG)^{s^{[p_1,p_2]}}, P^{[p_1,p_2]}_{\nu_1 n,\nu_2 n})$, where the exponent $s^{[p_1,p_2]}$ means that we apply $s^{[p_1,p_2]}$ smoothing steps by the PCG method with preconditioner (6.7). In this section we assume $\beta_1 = \beta_2 = 0$ to ensure that the matrix $A^{[p_1,p_2]}_{\nu_1 n,\nu_2 n}$ is SPD. If this is not the case, then we could simply replace PCG with P-GMRES and a similar reasoning holds as well.

We now sketch a motivation why (6.7) should be a suitable smoothing preconditioner. Thanks to Lemma 3.2, the symbol $f_{p_1,p_2}^{(\nu_1,\nu_2)}(\theta_1,\theta_2)$ can be factored as follows:

$$\begin{aligned} &f_{p_1,p_2}^{(\nu_1,\nu_2)}(\theta_1,\theta_2) = h_{p_2-1}(\theta_1)h_{p_1-1}(\theta_2)r_{p_1,p_2}^{(\nu_1,\nu_2)}(\theta_1,\theta_2),\\ &r_{p_1,p_2}^{(\nu_1,\nu_2)}(\theta_1,\theta_2) := \left[\frac{\nu_2}{\nu_1}w_{p_1}(\theta_2)(2-2\cos\theta_1)+\frac{\nu_1}{\nu_2}w_{p_2}(\theta_1)(2-2\cos\theta_2)\right], \end{aligned} \tag{6.8}$$

where $w_p(\theta) := \dfrac{h_p(\theta)}{h_{p-1}(\theta)}$ is a function 'well-separated' from zero for all $\theta \in [0,\pi]$ and all $p \geq 1$. This means that $r_{p_1,p_2}^{(\nu_1,\nu_2)}(\theta_1,\theta_2)$ does not have numerical zeros and only presents a zero at $\boldsymbol{\theta} = \mathbf{0}$, which, however, does not create problems to our two-grid schemes, because the projector $P_{\nu_1 n,\nu_2 n}^{[p_1,p_2]}$ takes care of it.

Therefore, the function $h_{p_2-1}(\theta_1)h_{p_1-1}(\theta_2)$ is responsible for the existence of numerical zeros at the edge points (6.3) when p_1, p_2 are large. Hence, the same function is also responsible for the poor behavior of our two-grid scheme $TG(S_{\nu_1 n,\nu_2 n}^{[p_1,p_2]}, P_{\nu_1 n,\nu_2 n}^{[p_1,p_2]})$ when p_1, p_2 are large. The choice of using the PCG method with preconditioner (6.7) as a smoother is made in order to 'erase' the numerical zeros at the edge points (6.3) as summarized in the following result.

Theorem 6.4. *The symbol of the sequences of preconditioned matrices*

$$\{T_{\nu_2 n+p_2-2,\nu_1 n+p_1-2}^{-1}(h_{p_2-1}\otimes h_{p_1-1})K_{\nu_1 n,\nu_2 n}^{[p_1,p_2]}\}_n$$

and

$$\{T_{\nu_2 n+p_2-2,\nu_1 n+p_1-2}^{-1}(h_{p_2-1}\otimes h_{p_1-1})A_{\nu_1 n,\nu_2 n}^{[p_1,p_2]}\}_n$$

is $r_{p_1,p_2}^{(\nu_1,\nu_2)}(\theta_1,\theta_2)$ defined in (6.8), i.e.,

$$\lim_{n\to\infty}\frac{1}{N}\sum_{j=1}^{N}F\left(\lambda_j\left(X_{\nu_1 n,\nu_2 n}^{[p_1,p_2]}\right)\right) = \frac{1}{2\pi}\int_{-\pi}^{\pi}F(r_{p_1,p_2}^{(\nu_1,\nu_2)}(\theta_1,\theta_2))\,\mathrm{d}\theta, \quad \forall F \in C_c(\mathbb{C}),$$

with the matrix $X_{\nu_1 n,\nu_2 n}^{[p_1,p_2]}$ being either $T_{\nu_2 n+p_2-2,\nu_1 n+p_1-2}^{-1}(h_{p_2-1}\otimes h_{p_1-1})K_{\nu_1 n,\nu_2 n}^{[p_1,p_2]}$ or $T_{\nu_2 n+p_2-2,\nu_1 n+p_1-2}^{-1}(h_{p_2-1}\otimes h_{p_1-1})A_{\nu_1 n,\nu_2 n}^{[p_1,p_2]}$, and $N := (\nu_1 n+p_1-2)(\nu_2 n+p_2-2)$.

Theorem 6.4 can be shown by following verbatim the proof of Theorem 5.1. Indeed, no changes have to be considered in the two-dimensional setting (and actually in any d-dimensional setting, see [14] for the formulation in d dimensions).

The proposed preconditioner (6.7) can be interpreted as a small rank perturbation of the (normalized) B-spline mass matrix $M_{\nu_2 n+1}^{[p_2-1]} \otimes M_{\nu_1 n+1}^{[p_1-1]}$ related to the fineness parameters $(\nu_1 n+1, \nu_2 n+1)$ and the spline degrees (p_1-1, p_2-1). Last but not the least, the preconditioner (6.7) is effectively solvable: due to the tensor-product structure, the computational cost for solving a linear system with matrix (6.7) is linear in the matrix size $(\nu_1 n+p_1-2)(\nu_2 n+p_2-2)$.

The V-cycle can be defined in a similar way as in the 1D case, see Section 5.2. For the sake of simplicity, we focus on a linear system with coefficient matrix $K_{n,n,0}^{[p,p]} := K_{n,n}^{[p,p]}$. The finest level is again indicated by index 0 and the coarsest level by index $\ell_n^{[p]} := \log_2(n+p-1)-1$. Let $K_{n,n,i}^{[p,p]}$ be the matrix at level i, whose dimension is $(m_{n,i}^{[p]})^2$, $0 \leq i \leq \ell_n^{[p]}$. We have $K_{n,n,i+1}^{[p,p]} := P_{n,n,i}^{[p,p]}K_{n,n,i}^{[p,p]}{P_{n,n,i}^{[p,p]}}^T$, where $P_{n,n,i}^{[p,p]} :=$

Table 7.1
Values of $\rho_n^{[p]} := \rho(TG(S_n^{[p]}, P_n^{[p]}))$ in the case $\beta = \gamma = 0$, for the specified parameter $\omega^{[p]}$.

n	$\rho_n^{[1]}$ $[\omega^{[1]} = 1/3]$	$\rho_n^{[3]}$ $[\omega^{[3]} = 1.0368]$	$\rho_n^{[5]}$ $[\omega^{[5]} = 1.2576]$
320	0.333333	0.447201	0.892595
640	0.333333	0.447073	0.892595
1280	0.333333	0.447037	0.892595
2560	0.333333	0.447039	0.892595
n	$\rho_n^{[2]}$ $[\omega^{[2]} = 0.7311]$	$\rho_n^{[4]}$ $[\omega^{[4]} = 1.2229]$	$\rho_n^{[6]}$ $[\omega^{[6]} = 1.2235]$
321	0.025287	0.737126	0.959435
641	0.025215	0.737102	0.959399
1281	0.025200	0.737102	0.959399
2561	0.025200	0.737102	0.959399

$P_{m_{n,i}^{[p]}, m_{n,i}^{[p]}}$, for $i = 0, \ldots, \ell_n^{[p]} - 1$, is the projector employed at level i, defined by (2.8) for $d = 2$ and $\boldsymbol{m} = (m_{n,i}^{[p]}, m_{n,i}^{[p]})$. Regarding the smoother, at each coarser level $i \geq 1$ the simple Gauss-Seidel smoother is used, whereas for the finest level $i = 0$ we propose to use $s^{[p,p]}$ smoothing iterations by the PCG method with preconditioner (6.7).

7. Numerical experiments.

In the numerical experiments we use Matlab 7.0 in double precision; the stopping criterion is the scaled residual with 10^{-8} tolerance; and the initial guess is the zero vector. We start with addressing the pure Laplacian, in order to show the adherence of the theoretical findings with the numerical results. Then, we consider a 2D problem with non-zero values of $\boldsymbol{\beta}$ and γ, for demonstrating the effectiveness of the proposed multi-iterative technique illustrated in Section 5 in a more general setting.

7.1. 1D Examples.

We fix $\beta = \gamma = 0$, so that $\frac{1}{n} A_n^{[p]} = K_n^{[p]}$. Table 7.1 shows the results of some numerical experiments for $TG(S_n^{[p]}, P_n^{[p]})$. For $p = 1, \ldots, 6$ we determined experimentally the best parameter $\omega^{[p]}$, in the sense that $\omega^{[p]}$ minimizes $\rho_n^{[p]} := \rho(TG(S_n^{[p]}, P_n^{[p]}))$ with $n = 2560$ (if p is odd) and $n = 2561$ (if p is even) among all $\omega \in \mathbb{R}$ with at most four nonzero decimal digits after the comma. We note that the choice $\omega^{[1]} = 1/3$ has a theoretical motivation, see Section 4.3. Finally, we computed the spectral radii $\rho_n^{[p]}$ for increasing values of n.

In all the considered experiments, the proposed two-grid scheme is optimal. Moreover, when $n \to \infty$, $\rho_n^{[p]}$ converges to a limit $\rho_\infty^{[p]}$, which is minimal not for $p = 1$ but 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. Finally, from some numerical experiments we notice that $\rho(K_n^{[4]}) \approx 1.8372$, $\forall n \geq 15$. Therefore, for the set of indices $\mathcal{I}_4 = \{81, 161, \ldots, 2561\}$ considered in Table 7.1, the best parameter $\omega^{[4]} = 1.2229$ produces a non-convergent smoother $S_n^{[4]} = I - 1.2229\, K_n^{[4]}$ having $\rho(S_n^{[p]}) \approx 1.2467$. This shows that the two-grid scheme can be convergent even when the smoother $S_n^{[p]}$ is not and, moreover, $\rho_n^{[p]}$ can attain its minimum at a value of $\omega^{[p]}$ for which $\rho(S_n^{[p]}) > 1$, according to the multi-iterative idea [30].

Thanks to the results given in Section 4, we can now interpret some observations about $TG(S_n^{[p]}, P_n^{[p]})$.

Observation 7.1. *The existence of an asymptotic spectral radius $\rho_\infty^{[p]}$, observed in Table 7.1 for $p = 1, \ldots, 6$, is not surprising: if $K_n^{[p]}$ were replaced by $\tau_{n+p-2}(f_p)$, the* *existence of an asymptotic spectral radius would follow from the analysis in Section 4, see in particular the equation (4.5).*

Observation 7.2. *Section 4.3 provides the key to understand why the asymptotic spectral radius $\rho_\infty^{[p]}$ worsens for increasing p: if $K_n^{[p]}$ were replaced by $\tau_{n+p-2}(f_p)$, this behavior would be a direct consequence from Proposition 4.5. Moreover, by combining Propositions 4.5 and 4.6, we expect that $\rho_\infty^{[p]}$ exponentially converges to 1 when $p \to \infty$. This 'exponentially poor' behavior is related to the fact that $f_p(\pi)/M_{f_p}$ exponentially approaches 0 when p increases (see Figure 3.1 and Table 3.1).*

Observation 7.3. *In Table 7.1 we have chosen for $\omega^{[p]}$ the best value among all $\omega \in \mathbb{R}$ with at most four nonzero decimal digits after the comma. The corresponding asymptotic spectral radius $\rho_\infty^{[p]}$ is then (almost) the best one. The fact that the best $\rho_\infty^{[p]}$ is minimal for $p = 2$ and not for $p = 1$ can be explained by means of Table 4.1. Indeed, if $K_n^{[p]}$ were replaced by $\tau_{n+p-2}(f_p)$, the best $\rho_\infty^{[p]}$ would be nothing else than $\widetilde{\rho}_\infty^{[p],*}$, which is minimal precisely for $p = 2$. In this regard, recall that $p = 2$ is the 'case of resonance' in which $f_2(\pi) = f_2\left(\frac{\pi}{2}\right)$, see the discussion in Section 4.3.*

Observation 7.4. *For $p = 1, 2$,*

- *$\widetilde{\rho}_\infty^{[p],*}$ is obtained with a value $\widetilde{\omega}^{[p],*}$ (shown in Table 4.1) which is very close to the value $\omega^{[p]}$ (shown in Table 7.1) for which the best $\rho_\infty^{[p]}$ is obtained;*
- *$\widetilde{\rho}_\infty^{[p],*}$ has a value very close to the best $\rho_\infty^{[p]}$.*

For $p = 3, 4, 5, 6$,

- *$\widetilde{\rho}_\infty^{[p],*}$ is obtained with a value $\widetilde{\omega}^{[p],*}$ (shown in Table 4.1) which is greater than the value $\omega^{[p]}$ (shown in Table 7.1) for which the best $\rho_\infty^{[p]}$ is obtained;*
- *$\widetilde{\rho}_\infty^{[p],*}$ has a value smaller than the best $\rho_\infty^{[p]}$.*

In view of this discussion, we stress that $K_n^{[p]}$ is 'more similar' to $\tau_{n+p-2}(f_p)$ when p is small. Indeed, $\operatorname{rank}(K_n^{[p]} - \tau_{n+p-2}(f_p))$ grows with p, although it is bounded by a constant independent of n.

Observation 7.5. *Recall that in Table 7.1 we experimentally determined the best values $\omega^{[p]}$ for $p = 1, \dots, 6$ when $n = 2560$ (if p is odd) and $n = 2561$ (if p is even). We now substitute these values in the expression of $\widetilde{\rho}_\infty^{[p]}$ (see (4.9) and (B.2) based on our conjecture), which would be the exact expression of $\rho_\infty^{[p]}$ if $K_n^{[p]}$ were replaced by $\tau_{n+p-2}(f_p)$. In this way we obtain*

$$\begin{array}{ll}
\widetilde{\rho}_\infty^{[1]}|_{\omega_1=1/3} = \dfrac{1}{3} \approx 0.333333, & \widetilde{\rho}_\infty^{[2]}|_{\omega_2=0.7311} = \dfrac{63}{2500} = 0.0252, \\
\widetilde{\rho}_\infty^{[3]}|_{\omega_3=1.0368} = \dfrac{1397}{3125} = 0.44704, & \widetilde{\rho}_\infty^{[4]}|_{\omega_4=1.2229} = \dfrac{82801}{112500} \approx 0.736009, \\
\widetilde{\rho}_\infty^{[5]}|_{\omega_5=1.2576} = \dfrac{126157}{141750} \approx 0.889996, & \widetilde{\rho}_\infty^{[6]}|_{\omega_6=1.2235} = \dfrac{37290373}{38981250} \approx 0.956623.
\end{array}$$

These values are very close to the exact values of $\rho_\infty^{[p]}$ for $p = 1, \dots, 6$, given in Table 7.1. In fact, for $p = 1, 2$ the value is exact; for $p = 3, \dots, 6$ the difference with the exact value is at most of the order of 10^{-3}. Unfortunately, the converse to this observation does not hold: if we use the best $\widetilde{\omega}^{[p],*}$ in Table 4.1 as relaxation parameter for $S_n^{[p]}$ in the two-grid algorithm $TG(S_n^{[p]}, P_n^{[p]})$, the resulting $\rho_\infty^{[p]}$ is not at all close to $\rho_\infty^{[p],*}$ for $p \geq 3$, and, for $p \geq 4$, it is also greater than 1.

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 the relaxed Gauss-Seidel iteration

Table 7.2
Values of $\widehat{\rho}_n^{[p]} := \rho(TG(\widehat{S}_n^{[p]}, P_n^{[p]}))$ in the case $\beta = \gamma = 0$, for the specified parameter $\omega^{[p]}$.

n	$\widehat{\rho}_n^{[1]}$ $[\omega^{[1]} = 0.9065]$	$\widehat{\rho}_n^{[3]}$ $[\omega^{[3]} = 0.9483]$	$\widehat{\rho}_n^{[5]}$ $[\omega^{[5]} = 1.1999]$
320	0.195630	0.156779	0.462856
640	0.222806	0.158920	0.471018
1280	0.235823	0.160239	0.475829
2560	0.241693	0.160975	0.478694

n	$\widehat{\rho}_n^{[2]}$ $[\omega^{[2]} = 0.9109]$	$\widehat{\rho}_n^{[4]}$ $[\omega^{[4]} = 1.0602]$	$\widehat{\rho}_n^{[6]}$ $[\omega^{[6]} = 1.3292]$
321	0.064874	0.320103	0.600236
641	0.064874	0.325533	0.610415
1281	0.064874	0.328651	0.616444
2561	0.064966	0.330459	0.619784

as smoother, which will be denoted by $\widehat{S}_n^{[p]}$. Table 7.2 illustrates the behavior of $\rho(TG(\widehat{S}_n^{[p]}, P_n^{[p]}))$. Like in Table 7.1, the relaxation parameter $\omega^{[p]}$ was chosen so as to minimize $\widehat{\rho}_n^{[p]} := \rho(TG(\widehat{S}_n^{[p]}, P_n^{[p]}))$ with $n = 2560$ (if p is odd) and $n = 2561$ (if p is even) among all $\omega \in \mathbb{R}$ with four nonzero decimal digits after the comma. It follows from Table 7.2 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. It is likely that such a worsening is an intrinsic feature of the problem and is related to the fact that $f_p(\pi)/M_{f_p}$ converges exponentially to 0 as p increases.

We also investigated the behavior of the two-grid scheme in the case $\beta = 100$ and $\gamma = 1$, for $p = 1, \ldots, 6$ and with $\omega^{[p]}$ chosen as in Table 7.1. Due to the presence of the dominating convection term, it turns out that the scheme is not convergent in this case for small values of n. However, we verified that the method with iteration matrix $TG(S_n^{[p]}, P_n^{[p]})$ is optimal if the set of indices $\mathcal{I}_p$ for which we solve (3.5) does not contain small values of n. In addition, for large n the value of $\rho_n^{[p]}$ is almost the same as the corresponding value in Table 7.1. This is not at all surprising because $\frac{1}{n} A_n^{[p]} = K_n^{[p]} + \frac{\beta}{n} H_n^{[p]} + \frac{\gamma}{n^2} M_n^{[p]}$ equals $K_n^{[p]}$ plus a matrix whose infinity norm tends to zero as $n \to \infty$, see Lemma 3.1.

7.2. 2D examples. Similar to the 1D case, the convergence rate of the two-grid and multigrid schemes rapidly worsens for increasing p_1, p_2 and for classical stationary smoothers like Gauss-Seidel (see [14]). Without providing a full theoretical justification, we limit ourselves to say that this is due to the presence of infinitely many 'numerical zeros' in the symbol $f_{p_1,p_2}^{(\nu_1,\nu_2)}$ when p_1, p_2 are large, see Section 6.1.

Table 7.3 collects the results of the V-cycle multigrid method when solving the system $K_{n,n}^{[p,p]}\mathbf{u} = \mathbf{b}_p \otimes \mathbf{b}_p$, with

$$\mathbf{b}_p = \frac{1}{n}\left[\frac{2}{p+1} \quad \frac{3}{p+1} \quad \cdots \quad \frac{p}{p+1} \quad 1 \quad \cdots \quad 1 \quad \frac{p}{p+1} \quad \cdots \quad \frac{3}{p+1} \quad \frac{2}{p+1}\right]^T,$$

for $p = 1, \ldots, 6$ and for increasing n. The V-cycle with the PCG smoother at the finest level (see Section 6.3) is compared to the same V-cycle but with relaxed Gauss-Seidel at the finest level (the relaxation parameter is chosen such that it minimizes the spectral radius of the two-grid method, see [14]). We can conclude that, when

Table 7.3

Number of V-cycle multigrid iterations $\widetilde{c}_n^{[p]}$ (resp. $\widehat{c}_n^{[p]}$) needed for solving $K_{n,n}^{[p,p]}\mathbf{u} = \mathbf{b}_p \otimes \mathbf{b}_p$, up to a precision of 10^{-8}, when using the multigrid cycle with $s^{[p,p]}$ smoothing steps by the PCG algorithm (resp. by the relaxed Gauss-Seidel smoother) at the finest level, and one smoothing step by the simple Gauss-Seidel smoother at the coarser levels. The parameters $s^{[p,p]}$ and $\omega^{[p,p]}$ are specified between brackets $[\cdot]$ near the labels $\widetilde{c}_n^{[p]}$ and $\widehat{c}_n^{[p]}$, respectively.

n	$\widetilde{c}_n^{[1]}$ [2]	$\widehat{c}_n^{[1]}$ [1.0035]	n	$\widetilde{c}_n^{[3]}$ [2]	$\widehat{c}_n^{[3]}$ [1.3143]	n	$\widetilde{c}_n^{[5]}$ [4]	$\widehat{c}_n^{[5]}$ [1.3990]
16	10	9	14	7	16	12	7	85
32	11	10	30	9	15	28	8	59
64	12	11	62	9	14	60	10	49
128	13	12	126	10	13	124	11	42
n	$\widetilde{c}_n^{[2]}$ [2]	$\widehat{c}_n^{[2]}$ [1.1695]	n	$\widetilde{c}_n^{[4]}$ [3]	$\widehat{c}_n^{[4]}$ [1.3248]	n	$\widetilde{c}_n^{[6]}$ [6]	$\widehat{c}_n^{[6]}$ [1.4914]
15	8	8	13	7	37	11	7	204
31	9	8	29	8	30	27	8	129
63	10	9	61	10	27	59	10	105
127	11	10	125	11	25	123	11	86

Table 7.4

Number of iterations $\widetilde{c}_n^{[p]}$ needed by $TG((P\text{-}GMRES)^{s^{[p,p]}}, P_{n,n}^{[p,p]})$ for solving $A_{n,n}^{[p,p]}\mathbf{u} = \mathbf{b}$ with $\boldsymbol{\beta} = (5, -5)$ and $\gamma = 1$, up to a precision of 10^{-8}. The parameter $s^{[p,p]}$ is specified between brackets $[\cdot]$.

n	$\widetilde{c}_{n,n}^{[1,1]}$ [2]	$\widetilde{c}_{n,n}^{[3,3]}$ [2]	$\widetilde{c}_{n,n}^{[5,5]}$ [4]	n	$\widetilde{c}_{n,n}^{[2,2]}$ [2]	$\widetilde{c}_{n,n}^{[4,4]}$ [3]	$\widetilde{c}_{n,n}^{[6,6]}$ [6]
20	7	6	7	21	6	6	6
40	6	6	6	41	6	6	6
60	6	6	6	61	6	6	6
80	6	6	6	81	6	6	6
100	6	6	6	101	6	6	5
120	7	6	6	121	6	6	6

using a few PCG smoothing steps at the finest level, the resulting V-cycle shows a convergence rate that is independent not only of n but also of p, and apparently also of the dimension d of the elliptic problem. This means that the proposed method is robust and optimal with respect to all the parameters n, p and d, i.e., the fineness parameter, the approximation parameter and the dimensionality of the elliptic problem, respectively.

In the final example, we consider a 2D problem with non-zero convection and reaction terms. The main message is that the replacement of the PCG smoother by a P-GMRES smoother, with the very same tensor preconditioner (6.7), does not change the effectiveness of the proposal. We take the linear system $A_{n,n}^{[p,p]}\mathbf{u} = \mathbf{b}$ coming from the B-spline IgA approximation of the model problem (1.1) in the case $d = 2$ with $\boldsymbol{\beta} = (5, -5)$, $\gamma = 1$ and $\mathrm{f} = 1$. The matrix $A_{n,n}^{[p,p]}$ is no longer symmetric, and so we replace the PCG smoother in the two-grid method $TG((PCG)^{s^{[p,p]}}, P_{n,n}^{[p,p]})$ with a GMRES smoother preconditioned by $T_{n+p-2}(h_{p-1}) \otimes T_{n+p-2}(h_{p-1})$. The results of the numerical experiment are given in Table 7.4.

An extensive numerical testing and comparison of the presented different solvers can be found in the twin paper [14]. In particular, additional experiments are provided for the W-cycle multigrid scheme with a PCG smoother, and also the three-dimensional setting is considered. Moreover, an outlook is given on how to extend the machinery towards more general elliptic problems with variable coefficients and defined on more complicated physical domains (using a geometry map).

8. Conclusion and perspectives.

In this paper we have proposed two-grid and multigrid methods for the solution of linear systems associated with certain stiffness matrices arising from the Galerkin B-spline IgA approximation of 1D and 2D elliptic problems. Through numerical experiments, we have provided evidence of their optimality and we have formally proved their optimality in certain cases. In particular, we have observed that, when using a few PCG smoothing steps at the finest level and adopting a properly chosen Toeplitz preconditioner, the resulting two-grid and V-cycle methods present an optimal convergence rate with respect to both the fineness parameter $h = \frac{1}{n}$ and the spline degree p. It is important to point out that

- the proposals and the analysis of the methods are based on the spectral symbol and on the corresponding techniques for τ-matrices and Toeplitz matrices (see [19, 20, 32]);
- the optimality proofs are based on classical tools (see [29, 32]);
- the spectral properties of the considered matrices, as well as the properties of the associated symbol, were analyzed in a previous work [23].

From a theoretical viewpoint, 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 (3.8). If (3.8) were true for all $p \geq 1$, then it would be easy to give a formal proof of optimality for all $p \geq 1$ (in the 1D setting) and for all $p_1, p_2 \geq 1$ (in the 2D setting). Indeed, the former would be a direct consequence of Corollary 3.3, whereas the latter would follow by replicating the argument used in Theorem 6.3. Furthermore, the matrix inequality (3.8) would be also of interest in the context of preconditioning connected with the CG method and the GMRES method. Indeed, in the light of the Axelsson-Lindskog theorems [3], it can be shown that (3.8), which is equivalent to (3.9) by Remark 3.5, ensures that $\tau_{n+p-2}(2-2\cos\theta) = T_{n+p-2}(2-2\cos\theta)$ is an optimal CG preconditioner for $K_n^{[p]}$. Hence, for $p = 1, 2, 3$, Theorem 3.4 ensures $\tau_{n+p-2}(2-2\cos\theta)$ to be an optimal CG preconditioner for $K_n^{[p]}$.

Summarizing, a plan for a next future research should include the following topics.

- Proving (3.8) for all $p \geq 1$ would give at once the optimality proof of the two-grid and – with little more efforts – the optimality proof of the W-cycle.
- We need more results on the asymptotic behavior with respect to p of the extreme eigenvalues of the preconditioned matrix reported in Theorem 5.1 and Theorem 6.4, for studying in more detail the $\boldsymbol{p}$ independence of the smoother, used in the multi-iterative scheme of Section 5.
- A complete theoretical proof of optimality for the V-cycle in all the different proposals given so far.
- We would like to extend the presented machinery towards a more general problem setting involving more complicated geometries, variable coefficient operators, etc. Some promising numerical experiments are provided in the twin paper [14] for testing the effectiveness of the multi-iterative technique illustrated in Section 5 beyond the formulation of problem (1.1).

With regard to the last item, we expect that the global symbol of the associated matrix sequences can be formed, in analogy with the FD/FE/IgA collocation contexts [4, 33, 15, 34], by exploiting the information from the main operator (the principal symbol in the Hörmander theory [25]), the used approximation techniques, and the involved physical domain via a geometric map.

Acknowledgments. This work was partially supported by INdAM-GNCS Gruppo Nazionale per il Calcolo Scientifico, by the Research Foundation Flanders, and by the MIUR 'Futuro in Ricerca' Programme through the project DREAMS.

REFERENCES

[1] A. Aricò, M. Donatelli, *A V-cycle multigrid for multilevel matrix algebras: proof of optimality*, Numer. Math., 105 (2007), pp. 511–547.

[2] A. Aricò, M. Donatelli, S. Serra-Capizzano, *V-cycle optimal convergence for certain (multilevel) structured linear systems*, SIAM J. Matrix Anal. Appl., 26 (2004), pp. 186–214.

[3] O. Axelsson, G. Lindskog, *On the rate of convergence of the preconditioned conjugate gradient method*, Numer. Math., 48 (1986), pp. 499–523.

[4] B. Beckermann, S. Serra-Capizzano, *On the asymptotic spectrum of Finite Elements matrices*, SIAM J. Numer. Anal., 45 (2007), pp. 746–769.

[5] L. Beirão da Veiga, D. Cho, L.F. Pavarino, S. Scacchi, *BDDC preconditioners for isogeometric analysis*, Math. Models Methods Appl. Sci., 23 (2013), pp. 1099–1142.

[6] ———, *Isogeometric Schwarz preconditioners for linear elasticity systems*, Comput. Methods Appl. Mech. Engrg., 253 (2013), pp. 439–454.

[7] R. Bhatia, *Matrix analysis*, Springer-Verlag, New York, 1997.

[8] M. Bolten, M. Donatelli, T. Huckle, *Generalized grid transfer operators for multigrid methods applied on Toeplitz matrices*, BIT Numer. Math., accepted.

[9] C. de Boor, *A practical guide to splines*, Springer-Verlag, New York, 2001.

[10] H. Brezis, *Functional analysis, Sobolev spaces and partial differential equations*, Springer, 2011.

[11] A. Buffa, H. Harbrecht, A. Kunoth, G. Sangalli, *BPX-preconditioning for isogeometric analysis*, Comput. Methods Appl. Mech. Engrg., 265 (2013), pp. 63–70.

[12] J.A. Cottrell, T.J.R. Hughes, Y. Bazilevs, *Isogeometric analysis: toward integration of CAD and FEA*, John Wiley & Sons, 2009.

[13] M. Donatelli, *An algebraic generalization of local Fourier analysis for grid transfer operators in multigrid based on Toeplitz matrices*, Numer. Linear Algebra Appl., 17 (2010), pp. 179–197.

[14] M. Donatelli, C. Garoni, C. Manni, S. Serra-Capizzano, H. Speleers, *Robust and optimal multi-iterative techniques for IgA Galerkin linear systems*, Comput. Methods Appl. Mech. Engrg., (2014), DOI: 10.1016/j.cma.2014.06.001.

[15] ———*Spectral analysis of matrices in isogeometric collocation methods*, Tech. Report TW648, Dept. Computer Science, KU Leuven, 2014.

[16] M. Donatelli, M. Molteni, V. Pennati, S. Serra-Capizzano, *Multigrid methods for cubic spline solution of two points (and 2D) boundary value problems*, Appl. Numer. Math., (2014), DOI: 10.1016/j.apnum.2014.04.004.

[17] M. Donatelli, S. Serra-Capizzano, D. Sesana, *Multigrid methods for Toeplitz linear systems with different size reduction*, BIT Numer. Math., 52 (2012), pp. 305–327.

[18] H. Engl, M. Hanke, and A. Neubauer, *Regularization of Inverse Problems*, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1996.

[19] G. Fiorentino, S. Serra, *Multigrid methods for Toeplitz matrices*, Calcolo, 28 (1991), pp. 283–305.

[20] ———, *Multigrid methods for symmetric positive definite block Toeplitz matrices with nonnegative generating functions*, SIAM J. Sci. Comput., 17 (1996), pp. 1068–1081.

[21] K.P.S. Gahalaut, J.K. Kraus, S.K. Tomar, *Multigrid methods for isogeometric discretization*, Comput. Methods Appl. Mech. Engrg., 253 (2013), pp. 413–425.

[22] K.P.S. Gahalaut, S.K. Tomar, J.K. Kraus, *Algebraic multilevel preconditioning in isogeometric analysis: construction and numerical studies*, Comput. Methods Appl. Mech. Engrg., 266 (2013), pp. 40–56.

[23] C. Garoni, C. Manni, F. Pelosi, S. Serra-Capizzano, H. Speleers, *On the spectrum of stiffness matrices arising from isogeometric analysis*, Numer. Math., (2013), DOI: 10.1007/s00211-013-0600-2.

[24] L. Golinskii, S. Serra-Capizzano, *The asymptotic properties of the spectrum of nonsymmetrically perturbed Jacobi matrix sequences*, J. Approx. Theory, 144 (2007), pp. 84–102.

[25] L. Hörmander, *Pseudo-differential operators and non-elliptic boundary problems*, Ann. of Math., 2 (1966), pp. 129–209.

[26] T.J.R. Hughes, J.A. Cottrell, Y. Bazilevs, *Isogeometric analysis: CAD, finite elements, NURBS, exact geometry and mesh refinement*, Comput. Methods Appl. Mech. Engrg., 194 (2005), pp. 4135–4195.

[27] X.Q. Jin, *Developments and Applications of Block Toeplitz Iterative Solvers*, Kluwer Academic Publishers, Dordrecht, 2002.

[28] D. Noutsos, S. Serra-Capizzano, P. Vassalos, *Matrix algebra preconditioners for multilevel Toeplitz systems do not insure optimal convergence rate*, Theoret. Comput. Sci., 315 (2004), pp. 557–579.

[29] J.W. Ruge, K. Stüben, *Algebraic multigrid*, Chapter 4 of the book *Multigrid methods* by S. McCormick, SIAM publications (1987), pp. 73–130.

[30] S. Serra, *Multi-iterative methods*, Comput. Math. Appl., 26 (1993), pp. 65–87.

[31] S. Serra-Capizzano, *Matrix algebra preconditioners for multilevel Toeplitz matrices are not superlinear*, Linear Algebra Appl., 343–344 (2002), pp. 303–319.

[32] ———, *Convergence analysis of two-grid methods for elliptic Toeplitz and PDEs matrix-sequences*, Numer. Math., 92 (2002), pp. 433–465.

[33] ———, *Generalized Locally Toeplitz sequences: spectral analysis and applications to discretized partial differential equations*, Linear Algebra Appl., 366 (2003), pp. 371–402.

[34] ———, *The GLT class as a generalized Fourier analysis and applications*, Linear Algebra Appl., 419 (2006), pp. 180–233.

[35] P. Tilli, *A note on the spectral distribution of Toeplitz matrices*, Linear Multilinear Algebra, 45 (1998), pp. 147–159.

[36] ———, *Locally Toeplitz sequences: spectral properties and applications*, Linear Algebra Appl., 278 (1998), pp. 91–120.

Appendix A. This appendix collects the proofs of Propositions 4.4, 4.5 and 4.6 in Section 4.2.

Proposition 4.4. *For every $p \geq 1$ we have $f_p\left(\frac{\pi}{2}\right) = 2^{p-2} f_p(\pi)$.*

Proof. From Lemma 3.2 we know that

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

Hence,

$$f_p\left(\frac{\pi}{2}\right) = 2 \sum_{k\in\mathbb{Z}} \left(\frac{2}{(\frac{\pi}{2} + 2k\pi)^2}\right)^p = \frac{2^{3p+1}}{\pi^{2p}} \sum_{k\in\mathbb{Z}} \frac{1}{(4k+1)^{2p}}, \tag{A.1}$$

$$f_p(\pi) = 4 \sum_{k\in\mathbb{Z}} \left(\frac{4}{(\pi + 2k\pi)^2}\right)^p = \frac{2^{2p+2}}{\pi^{2p}} \sum_{k\in\mathbb{Z}} \frac{1}{(2k+1)^{2p}}. \tag{A.2}$$

By splitting the latter sum into a sum over the even integers and a sum over the odd integers, we get

$$\begin{aligned} \sum_{k\in\mathbb{Z}} \frac{1}{(2k+1)^{2p}} &= \sum_{l\in\mathbb{Z}} \frac{1}{(4l+1)^{2p}} + \frac{1}{(4l+3)^{2p}} = \sum_{l\in\mathbb{Z}} \frac{1}{(4l+1)^{2p}} + \sum_{m\in\mathbb{Z}} \frac{1}{(-4m-1)^{2p}} \\ &= \sum_{l\in\mathbb{Z}} \frac{1}{(4l+1)^{2p}} + \sum_{m\in\mathbb{Z}} \frac{1}{(4m+1)^{2p}} = 2 \sum_{k\in\mathbb{Z}} \frac{1}{(4k+1)^{2p}}. \end{aligned} \tag{A.3}$$

Therefore, by combining (A.3) with (A.1) and (A.2), we obtain

$$\frac{f_p\left(\frac{\pi}{2}\right)}{f_p(\pi)} = \frac{2^{3p+1}}{2^{2p+2}} \frac{\sum_{k\in\mathbb{Z}} \frac{1}{(4k+1)^{2p}}}{2\sum_{k\in\mathbb{Z}} \frac{1}{(4k+1)^{2p}}} = 2^{p-2}.$$

□

Proposition 4.5. *Let $p \geq 1$. Then, independently of the choice of $\omega^{[p]} \in \mathbb{R}$,*

$$\widetilde{\rho}_{\infty}^{[p]} \geq \frac{2^{p-2} - 1}{2^{p-2} + 1} =: \sigma^{[p]}.$$

In particular, $\widetilde{\rho}_{\infty}^{[p],*} \geq \sigma^{[p]}$.

Proof. We start with showing that for each fixed threshold $\sigma \geq 0$ and for any $\omega^{[p]} \in \mathbb{R}$, if $\widetilde{\rho}_{\infty}^{[p]} \leq \sigma$ then it holds

$$(1 - \sigma)2^{p-2} - 1 \leq \widetilde{\rho}_{\infty}^{[p]} \leq \sigma. \tag{A.4}$$

Fix $\sigma \geq 0$ and choose arbitrarily $\omega^{[p]} \in \mathbb{R}$. By (4.2) we have

$$\widetilde{\rho}_{\infty}^{[p]} = \|t_p\|_\infty \geq t_p(0) = 1 - \omega^{[p]} f_p(\pi),$$

and if $\widetilde{\rho}_{\infty}^{[p]} \leq \sigma$ then

$$\omega^{[p]} \geq \frac{1-\sigma}{f_p(\pi)}. \tag{A.5}$$

$$\widetilde{\rho}_{\infty}^{[p]} \geq |t_p(\pi/2)| = |s_p(\pi/2)| = \left|1 - \omega^{[p]} f_p(\pi/2)\right| \geq \omega^{[p]} f_p(\pi/2) - 1. \tag{A.6}$$

Hence, by combining (A.6) with (A.5) and by Proposition 4.4, we find that

$$\widetilde{\rho}_{\infty}^{[p]} \geq \omega^{[p]} f_p(\pi/2) - 1 \geq \frac{1-\sigma}{f_p(\pi)} f_p(\pi/2) - 1 = (1-\sigma)2^{p-2} - 1,$$

resulting in (A.4). Note that both the lower and upper bound in (A.4) are independent of the choice of $\omega^{[p]}$.

Since the interval $[(1-\sigma)2^{p-2}-1, \sigma]$ is empty for $\sigma < \sigma^{[p]}$, it follows from (A.4) that $\widetilde{\rho}_{\infty}^{[p]} \geq \sigma^{[p]}$, and the proposition is proved. □

In order to prove Proposition 4.6, we introduce the definition of some specific functions and three auxiliary lemmas. By defining

$$g_p(\theta) := \frac{q^2(\pi-\theta)}{f_p(\theta)} = \frac{(1-\cos\theta)^2}{f_p(\theta)}, \qquad g_p(0) := \lim_{\theta\to 0} g_p(\theta) = 0, \tag{A.7}$$

and

$$v_p(\theta) := \frac{q^2(\theta) + q^2(\pi-\theta)}{g_p(\theta) + g_p(\pi-\theta)} = \frac{2 + 2\cos^2(\theta)}{g_p(\theta) + g_p(\pi-\theta)}, \tag{A.8}$$

it follows that

$$t_p(\theta) = 1 - \omega^{[p]}\, v_p(\theta). \tag{A.9}$$

Note that $v_p(\theta) > 0$ for all θ, because $f_p(\theta) \geq 0$ for all θ, see Lemma 3.2.

Lemma A.1. *Let $p \geq 1$ and assume that*

$$0 \leq L^{[p]} \leq v_p(\theta) \leq U^{[p]}, \qquad \theta \in \left[0, \frac{\pi}{2}\right]. \tag{A.10}$$

Then, for the optimal value $\widetilde{\rho}_{\infty}^{[p],*}$ defined in (4.6) we have the following upper bound

$$\widetilde{\rho}_{\infty}^{[p],*} \leq 1 - \frac{2L^{[p]}}{L^{[p]} + U^{[p]}} = \frac{U^{[p]} - L^{[p]}}{U^{[p]} + L^{[p]}}.$$

Proof. Using the bounds (A.10) we obtain

$$|t_p(\theta)| = |1 - \omega^{[p]} v_p(\theta)| \leq \max\left\{|1 - \omega^{[p]} L^{[p]}|, |1 - \omega^{[p]} U^{[p]}|\right\}.$$

Hence,

$$\widetilde{\rho}_{\infty}^{[p],*} = \min_{\omega^{[p]}\in\mathbb{R}} \|t_p\|_\infty \leq \min_{\omega^{[p]}\in\mathbb{R}} \max\left\{|1 - \omega^{[p]} L^{[p]}|, |1 - \omega^{[p]} U^{[p]}|\right\} = 1 - \frac{2L^{[p]}}{L^{[p]} + U^{[p]}}.$$

□

To provide suitable bounds for v_p (in Lemma A.3), we make use of the following property of h_p defined in (3.2).

LEMMA A.2. *For $p \geq 1$, h_p decreases monotonically in* $[0, \pi]$.

Proof. Clearly, $h_1(\theta) = \frac{2}{3} + \frac{1}{3}\cos\theta$ is monotone decreasing in $[0, \pi]$. We now show that $h_p'(\theta) < 0$ for $p \geq 2$ and $\theta \in (0, \pi)$. From (3.4) we get

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

Let $\omega := \theta/2 \in (0, \pi/2)$. We consider the series of derivatives[3]

$$\begin{aligned} h_p'(\theta) &= (p+1) \sum_{k\in\mathbb{Z}} \left(\frac{\sin\omega}{\omega + k\pi}\right)^{2p+1} \left[\frac{\cos\omega}{\omega + k\pi} - \frac{\sin\omega}{(\omega + k\pi)^2}\right] \\ &= (p+1)(\sin\omega)^{2p+1} \cos\omega \sum_{k\in\mathbb{Z}} \left[\frac{1}{(\omega + k\pi)^{2p+2}} - \frac{\tan\omega}{(\omega + k\pi)^{2p+3}}\right] \\ &= (p+1)(\sin\omega)^{2p+1} \cos\omega \left[\frac{1}{\omega^{2p+2}} - \frac{\tan\omega}{\omega^{2p+3}} + x_p(\omega)\right], \end{aligned}$$

where

$$x_p(\omega) := \sum_{k=1}^{\infty} x_{p,k}^{(1)}(\omega) + \tan\omega \sum_{k=1}^{\infty} x_{p,k}^{(2)}(\omega).$$

and

$$x_{p,k}^{(1)}(\omega) := \frac{1}{(k\pi + \omega)^{2p+2}} + \frac{1}{(k\pi - \omega)^{2p+2}}, \quad x_{p,k}^{(2)}(\omega) := \frac{1}{(k\pi - \omega)^{2p+3}} - \frac{1}{(k\pi + \omega)^{2p+3}}.$$

For $p \geq 2$ and $k \geq 1$, one can check that the functions $x_{p,k}^{(1)}(\omega)$ and $x_{p,k}^{(2)}(\omega)$ are monotone increasing, and

$$\begin{aligned} x_{p,k}^{(1)}(\omega) &\leq \frac{1}{(k\pi + \pi/2)^{2p+2}} + \frac{1}{(k\pi - \pi/2)^{2p+2}} \leq \left(\frac{2}{\pi}\right)^{2p+2} \left[\frac{1}{(2k+1)^6} + \frac{1}{(2k-1)^6}\right], \\ x_{p,k}^{(2)}(\omega) &\leq \frac{1}{(k\pi - \pi/2)^{2p+3}} - \frac{1}{(k\pi + \pi/2)^{2p+3}} = \left(\frac{2}{\pi}\right)^{2p+3} \left[\frac{1}{(2k-1)^{2p+3}} - \frac{1}{(2k+1)^{2p+3}}\right]. \end{aligned}$$

Hence,

$$x_p(\omega) \leq \left(\frac{2}{\pi}\right)^{2p+2} \left(\frac{\pi^6}{480} - 1 + \frac{2}{\pi}\tan\omega\right),$$

and the derivative of h_p is bounded above by

$$h_p'(\theta) \leq (p+1) \left(\frac{\sin\omega}{\omega}\right)^{2p+1} \frac{\cos\omega}{\omega}\, y_p(\omega), \tag{A.11}$$

where

$$y_p(\omega) := 1 - \frac{\tan\omega}{\omega} + \left(\frac{2\omega}{\pi}\right)^{2p+2} \left(\frac{\pi^6}{480} - 1 + \frac{2}{\pi}\tan\omega\right).$$

[3] The equality holds due to the uniform convergence in $[-\pi, \pi]$.

Moreover, for $\omega \in (0, \pi/2)$ and $p \geq 2$,

$$y_p(\omega) \leq 1 - \frac{\tan\omega}{\omega} + \left(\frac{2\omega}{\pi}\right)^6 \left(\frac{\pi^6}{480} - 1 + \frac{2}{\pi}\tan\omega\right) < 0. \tag{A.12}$$

From (A.11)–(A.12) we conclude that h_p decreases monotonically in $[0, \pi]$ for $p \geq 2$. □

We now propose a suitable upper and lower bound for v_p.

Lemma A.3. *For every $p \geq 1$, let*

$$L^{[p]} := \frac{2}{(2^p + 1)g_p(\pi/2)} = \frac{2f_p(\pi/2)}{2^p + 1} \quad \textit{and} \quad U^{[p]} := \frac{4}{g_p(\pi/2)} = 4f_p(\pi/2).$$

Then, we have

$$L^{[p]} \leq v_p(\theta) \leq U^{[p]}, \quad \theta \in \left[0, \frac{\pi}{2}\right]. \tag{A.13}$$

Proof. From (A.7) and Lemma 3.2 we have

$$g_p(\theta) = \frac{1 - \cos\theta}{2h_{p-1}(\theta)}.$$

It follows that g_p is monotone increasing in $[0, \pi]$, because h_{p-1} is monotone decreasing in the same interval if $p \geq 2$ (Lemma A.2) and is constant if $p = 1$. Then,

$$g_p(0) + g_p(\pi/2) \leq g_p(\theta) + g_p(\pi - \theta) \leq g_p(\pi/2) + g_p(\pi), \quad \theta \in \left[0, \frac{\pi}{2}\right]. \tag{A.14}$$

Moreover, we have $g_p(0) = 0$ and, in view of Proposition 4.4,

$$\frac{g_p(\pi)}{g_p(\pi/2)} = \frac{4}{f_p(\pi)} f_p(\pi/2) = 2^p. \tag{A.15}$$

By combining (A.14)–(A.15) and $2 \leq 2 + 2\cos^2(\theta) \leq 4$ with (A.8), we obtain (A.13). □

Proposition 4.6. *Let $p \geq 1$, then*

$$\widetilde{\rho}_\infty^{[p],*} \leq \frac{2^{p+1} + 1}{2^{p+1} + 3} =: \varsigma^{[p]}.$$

Proof. The upper bound for $\widetilde{\rho}_\infty^{[p],*}$ follows immediately from Lemmas A.1 and A.3. □

Appendix B. In this appendix we formulate a stronger conjecture than (4.10), namely

$$v_p'(\theta) \geq 0, \quad \theta \in \left[0, \frac{\pi}{2}\right], \quad p \geq 3. \tag{B.1}$$

Lemma B.1. *If the conjecture (B.1) is true, then (4.10) is also true.*

Proof. By assuming (B.1) and recalling (A.9), we deduce that t_p is monotone over $\left[0, \frac{\pi}{2}\right]$ for every $p \geq 3$ and every $\omega^{[p]} \in \mathbb{R}$. Hence,

$$\widetilde{\rho}_\infty^{[p]} = \max\left(|t_p(0)|, \left|t_p\left(\frac{\pi}{2}\right)\right|\right) = \max\left(\left|1 - \omega^{[p]} f_p(\pi)\right|, \left|1 - \omega^{[p]} f_p\left(\frac{\pi}{2}\right)\right|\right),$$

Table B.1
Values of $f_p(\pi)$ for $p = 1, \ldots, 9$.

p	1	2	3	4	5	6	7	8	9
$f_p(\pi)$	4	$\frac{4}{3}$	$\frac{8}{15}$	$\frac{68}{315}$	$\frac{248}{2835}$	$\frac{5528}{155925}$	$\frac{87376}{6081075}$	$\frac{3718276}{638512875}$	$\frac{25618328}{10854718875}$

and by using Proposition 4.4,

$$\widetilde{\rho}_{\infty}^{[p]} = \max\left(\left|1 - \omega^{[p]} f_p(\pi)\right|, \left|1 - \omega^{[p]} 2^{p-2} f_p\left(\pi\right)\right|\right). \tag{B.2}$$

In particular, the best value $\widetilde{\omega}^{[p],*}$ that minimizes $\widetilde{\rho}_{\infty}^{[p]}$ is the solution of the equation $1 - \widetilde{\omega}^{[p],*}\, f_p(\pi) = -\left(1 - \widetilde{\omega}^{[p],*}\, 2^{p-2} f_p(\pi)\right)$, i.e.,

$$\widetilde{\omega}^{[p],*} = \frac{2}{f_p(\pi) \cdot (2^{p-2} + 1)} = \frac{2}{f_p(\pi) + f_p\left(\frac{\pi}{2}\right)}, \tag{B.3}$$

and the best asymptotic spectral radius is

$$\widetilde{\rho}_{\infty}^{[p],*} = \widetilde{\rho}_{\infty}^{[p]}|_{\omega^{[p]} = \widetilde{\omega}^{[p],*}} = \left|\frac{2^{p-2} - 1}{2^{p-2} + 1}\right| = |\sigma^{[p]}|, \tag{B.4}$$

which is equal to $\sigma^{[p]}$ for $p \geq 2$. □

The validity of (B.1) has been observed experimentally for $p = 3, \ldots, 9$. Consequently, the formulas (B.2)–(B.4), as well as (4.10), hold for these values of p, see Table 4.1 and also Table B.1 for the values of $f_p(\pi)$. For $p \geq 10$, the validity of (B.1) has not been proved and so we cannot assert that (4.10) certainly holds for all $p \geq 10$. However, we know from Propositions 4.5 and 4.6 that

$$\sigma^{[p]} \leq \widetilde{\rho}_{\infty}^{[p],*} \leq \varsigma^{[p]}, \quad \forall p \geq 1,$$

and the gap between $\sigma^{[p]}$ and $\varsigma^{[p]}$ is quite small. Recall that $\sigma^{[p]}$ and $\varsigma^{[p]}$ converges to 1 with the same asymptotic speed, see (4.8). For instance, if $p = 10$ we have $\sigma^{[10]} \approx 0.9922$ and $\varsigma^{[10]} \approx 0.9990$.

It is worth noting that the formulas (B.2)–(B.4) hold even for $p = 1$, see Tables 4.1 and B.1. In fact, the above derivation of (B.2)–(B.4) only requires the monotonicity of t_p, which is verified for $p = 1$ too. On the other hand, in the case $p = 2$ the formulas (B.2)–(B.4) do not hold. The case $p = 2$ is actually somewhat peculiar and can be interpreted as a 'case of resonance', see Section 4.3.