arXiv:2108.09086v1 [math.NA] 20 Aug 2021

Spectral and norm estimates for matrix sequences arising from a finite difference approximation of elliptic operators

Armando Coco[1], Sven-Erik Ekström[2], Giovanni Russo[3], Stefano Serra-Capizzano[2,4], and Santina Chiara Stissi[3,5]

[1]School of Engineering, Computing and Mathematics, Oxford Brookes University - OX33 1HX, Oxford (UK)
[2]Department of Information Technology, Division of Scientific Computing, Uppsala University - ITC, Lägerhyddsv. 2, hus 2, P.O. Box 337, SE-751 05, Uppsala (SWEDEN)
[3]Department of Mathematics and Informatics, Catania University - Viale A. Doria 6, 95125 Catania (ITALY)
[4]Department of Humanities and Innovation, Insubria University - via Valleggio 11, 22100 Como (ITALY)
[5]Istituto Nazionale di Geofisica e Vulcanologia - Piazza Roma 2, 95125 Catania (ITALY)

Abstract

When approximating elliptic problems by using specialized approximation techniques, we obtain large structured matrices whose analysis provides information on the stability of the method. Here we provide spectral and norm estimates for matrix sequences arising from the approximation of the Laplacian via ad hoc finite differences. The analysis involves several tools from matrix theory and in particular from the setting of Toeplitz operators and Generalized Locally Toeplitz matrix sequences. Several numerical experiments are conducted, which confirm the correctness of the theoretical findings.

Keywords: Toeplitz matrix, generating function and spectral symbol, approximation of differential operators.

1 Introduction

In the numerical approximation of elliptic differential equations, by using specialized approximation techniques, we obtain large structured matrices whose analysis provides information on the stability of the method. Here we provide spectral and norm estimates for matrix sequences arising from the approximation of the Laplacian via ad hoc finite differences that is from the Coco–Russo method [5].

The analysis involves several tools from matrix theory and in particular from the setting of Toeplitz operators and Generalized Locally Toeplitz (GLT) matrix sequences. Several numerical experiments are conducted, which confirm the theoretical findings.

The paper is organized as follows. Subsection 1.1 contains a motivation and a description of the Coco–Russo method, together with a brief account on the related literature. Subsection 1.2 contains the necessary tools from the Toeplitz technology, while Section 2 contains the matrix formulation in $1\mathbb{D}$ in the language of Toeplitz structures, the analysis of the norm estimates in $1\mathbb{D}$, together with related numerical experiments and a preliminary discussion on the spectral features of the involved

matrix-sequences. Section 3 contains more details on the $2\mathbb{D}$ method, on its matrix formulation, on the spectral results in $1\mathbb{D}$ and in $2\mathbb{D}$, and the basic tools taken the GLT theory. A discussion on the more challenging case of the norm estimates in $2\mathbb{D}$ is also provided. A conclusion section ends the paper with a mention to a few open problems.

1.1 Method description and motivation

The design of numerical methods to solve Partial Differential Equations (PDE) on complex-shaped domains is obtaining an increasing interest in the scientific community. One of the bottlenecks of modern computer simulations is the modelling of physical processes around $3\mathbb{D}$ complex-shaped objects through PDE. Finite Element Methods (FEM) are well-established approaches to solve PDE and supported by rigorous theoretical analysis developed in the last decades to prove the convergence and accuracy order of the method when the grid size approaches zero.

However, some critical limitations are commonly associated in literature with FEM, especially when applied to curved boundaries. In particular, the generation of elements to conform highly varying curvatures of the boundary might become cumbersome, especially if the domain changes its shape over time. Also, the design of a balanced partition of the mesh for parallel FEM is unhandy. For these reasons, approaches based on Finite Difference Methods (FDM) where the domain is immersed into a fixed grid are increasing their popularity in literature, since they do not require any mesh generation effort and at the same time allow for a natural design of parallel solvers.

On the other hand, FDM are commonly based on heuristic approaches and convergence and stability analysis are not sufficiently developed in literature, especially for the case of curved boundaries.

The Immersed Boundary Method proposed by Peskin in [15] and further developed by LeVeque and Li in [12] is a pioneer approach based on FDM for general domains immersed on fixed grids.

A more recent approach is the Ghost-Fluid Method proposed by Fedkiw et al. in [6] and further extended to higher accuracy by Gibou et al. in [10, 9], where the values on grid nodes just outside the domain (ghost points) are obtained by accurate extrapolations of the boundary condition from inside values.

In [5], the authors present a highly efficient and accurate ghost-point method to solve a Poisson equation on a complex-shaped domain, modelled by a level-set function. Several numerical tests were presented to confirm the accuracy order and the efficiency of the multigrid solver. However, a theoretical analysis was missing. The method has been extended to several applications, such as compressible fluids in moving domains [3] or volcanology [4].

In this paper we present a technique to prove the stability of the Coco–Russo method [5] and the convergence to the predicted order of accuracy.

We start from the $1\mathbb{D}$ problem. Consider the elliptic boundary-value problem:

$$-\Delta u = f \text{ on } \Omega = (a, b) \tag{1}$$

$$u(a) = g_a, \quad u(b) = g_b \tag{2}$$

and a one-dimensional uniform grid $\mathcal{G}_h = \{x_0, x_1, \ldots, x_{n+1}\}$ with a constant spatial step $h = x_i - x_{i-1}$, for $i = 1, \ldots, n+1$. Then, $x_i = x_0 + i\,h$. Let $x_0 < a < x_1$ and $x_{n+1} = b$ (see Fig. 1).

The elliptic equation $-\Delta u = f$ is discretized by central differences on x_i for $i = 1, \ldots, n$ and the boundary condition on $x = b = x_{n+1}$ is included in the internal discretization (this is the so

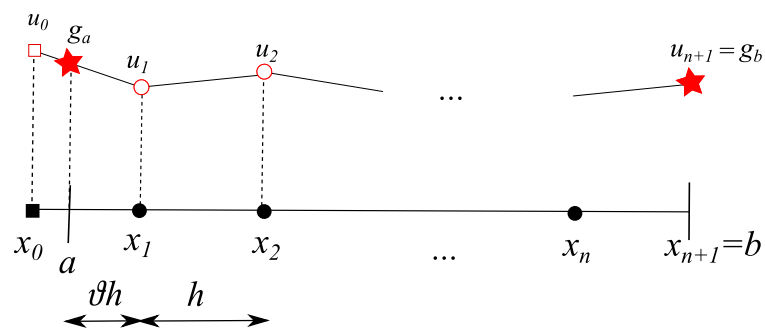

Figure 1: Discretization of the $1\mathbb{D}$ domain. Full black circles are the interior grid points, while the full square is the ghost point. Boundary values are indicated with red stars. Linear extrapolation is used to define the ghost value u_0 from u_1 and the left boundary value g_a.

called eliminated boundary condition approach):

$$\frac{-u_{i-1} + 2u_i - u_{i+1}}{h^2} = f_i \text{ for } i = 1, \ldots, n-1,$$

$$\frac{2u_n - u_{n-1}}{h^2} = f_n + \frac{g_b}{h^2}.$$

The boundary condition on $x = a$ is approximated by $q(a) = g_a$, where $q(x)$ is the polynomial of degree $s-1$ that interpolates u on the grid points $u_0, u_1, \ldots, u_{s-1}$. We call s the stencil size for the boundary condition on $x = a$.

The discretization of the boundary condition can be represented as:

$$\sum_{i=0}^{s-1} c_i u_i = g_\alpha. \tag{3}$$

For $s = 2$ we have

$$\vartheta u_0 + (1 - \vartheta) u_1 = g_\alpha, \tag{4}$$

where $\vartheta = (x_1 - a)/h$. The grid point x_0 is called ghost point and u_0 is the ghost value.

Although we can follow a similar technique for the boundary condition on $x = a$ to the one that we adopted for $x = b$ (i.e. we can solve (4) for u_0 and substitute its value into the internal equation for x_1), we keep a non-eliminated boundary condition approach in order to develop a theoretical analysis that can be straightforwardly extended to higher dimensional cases, where the eliminated approach is impractical.

The discretized problem is then a linear system $A_h \mathbf{u}_h = \mathbf{f}_h$ where $A_h \in \mathbb{R}^{(n+1)\times(n+1)}$:

$$A_h \mathbf{u}_h = \begin{bmatrix} \vartheta & 1-\vartheta & 0 & \dots & \dots & 0 \\ -\frac{1}{h^2} & \frac{2}{h^2} & -\frac{1}{h^2} & 0 & \dots & 0 \\ 0 & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & -\frac{1}{h^2} & \frac{2}{h^2} & -\frac{1}{h^2} \\ 0 & 0 & \dots & 0 & -\frac{1}{h^2} & \frac{2}{h^2} \end{bmatrix} \begin{bmatrix} u_0 \\ u_1 \\ \vdots \\ \vdots \\ u_{n-1} \\ u_n \end{bmatrix} = \begin{bmatrix} g_a \\ f_1 \\ f_2 \\ \vdots \\ f_{n-1} \\ f_n + \frac{g_b}{h^2} \end{bmatrix} = \mathbf{f}_h, \tag{5}$$

where $h = (b - x_0)/(n+1)$.

1.2 Toeplitz structures and related tools

Let T_n be a Toeplitz matrix of order n and let $\omega < n$ be a positive integer

$$T_n = \begin{bmatrix} a_0 & \cdots & a_{-\omega} & & & \\ \vdots & \ddots & & \ddots & & \\ & & \ddots & & \ddots & \\ a_\omega & & & \ddots & & a_{-\omega} \\ & \ddots & & & \ddots & \vdots \\ & & a_\omega & \cdots & & a_0 \end{bmatrix}, \tag{6}$$

where the coefficients a_k, $k = -\omega, \dots, \omega$, are complex numbers.

Let $f \in L^1(-\pi, \pi)$ and let $T_n(f)$ be the Toeplitz matrix generated by f i.e. $(T_n(f))_{s,t} = a_{s-t}(f)$, $s, t = 1, \dots, n$, with f indicated as <u>generating function</u> of $\{T_n(f)\}$ and with $a_k(f)$ being the k-th Fourier coefficient of f that is

$$a_k(f) = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(\theta) \mathrm{e}^{-\mathbf{i}k\theta} \, \mathrm{d}\theta, \quad \mathbf{i}^2 = -1, \ k \in \mathbb{Z}. \tag{7}$$

With these notations the matrix reported in (6) can be written as $T_n = T_n(f)$, where the generating function is $f(\theta) = \sum_{j=-\omega}^{\omega} a_j \mathrm{e}^{\mathbf{i}j\theta}$. It is worth noticing that study of the generating function gives plenty of information on the spectrum of $T_n(f)$ for any fixed n, and also asymptotically as the matrix-size n diverges to infinity (see [7] and [8] for the multilevel setting). For instance,r if f is real-valued almost everywhere (a.e.), then $T_n(f)$ is Hermitian for all n. Furthermore, when f is real-valued and even a.e., the matrix $T_n(f)$ is (real) symmetric for all n, while f real-valued and nonnegative a.e., but not identically zero a.e., implies that $T_n(f)$ is Hermitian positive definite for all n: in such a setting the considered matrix-sequence could be ill-conditioned and indeed if f is nonnegative and bounded with essential supremum equal to $M > 0$ and a unique zero of order $\alpha > 0$, then the maximal eigenvalue converges monotonically from below to M, whereas the minimal eigenvalues converges to zero monotonically from above with a speed dictated by α, that is the minimal eigenvalue is asymptotical to $n^{-\alpha}$. In many practical applications we remind that it is required to solve numerically linear systems of Toeplitz kind and of (very) large dimensions and hence several specialized techniques of iterative type, such as preconditioned Krylov methods and ad hoc multigrid procedures have been designed; we refer the interested reader to the books [14, 2] and to the references therein. We recall that such types of large Toeplitz linear systems emerge from specific applications involving e.g. the numerical solution of (integro-) differential equations and of problems with Markov chains.

2 Matrix formulation and notation in $1\mathbb{D}$

The linear system to solve is (5), and we can decompose the matrix $A_h \in \mathbb{R}^{(n+1)\times(n+1)}$ as follows

$$A_h = \frac{1}{h^2}\begin{bmatrix} 2 & -1 & 0 & \dots & \dots & 0 \\ -1 & 2 & -1 & 0 & \dots & 0 \\ 0 & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & -1 & 2 & -1 \\ 0 & 0 & \dots & 0 & -1 & 2 \end{bmatrix} + \frac{1}{h^2}\begin{bmatrix} \vartheta h^2 - 2 & (1-\vartheta)h^2 + 1 & 0 & \dots & 0 \\ 0 & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ \vdots & \vdots & \vdots & & \vdots \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \dots & 0 \end{bmatrix}$$

$$= \frac{1}{h^2}T_{n+1}(2 - 2\cos(\theta)) + \frac{1}{h^2}\begin{bmatrix} \mathbf{v}_h^{\mathrm{T}} \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \end{bmatrix} \tag{8}$$

$$= S_{n+1} + \frac{1}{h^2}\mathbf{e}_1\mathbf{v}_h^{\mathrm{T}}, \tag{9}$$

where $T_{n+1}(f)$ in (8) is the Toeplitz matrix generated by f according to (7), with $f(\theta) = 2-2\cos(\theta)$ so that, in the matrix in (6), we have $\alpha = 1$, $a_0 = 2, a_1 = a_{-1} = -1$. Furthermore we have defined S_{n+1} in (9) as

$$S_{n+1} = \frac{1}{h^2}T_{n+1}(2 - 2\cos(\theta)).$$

For this matrix everything is known and in fact

$$T_{n+1}(2 - 2\cos(\theta)) = QDQ$$

with Q real symmetric and orthogonal and

$$Q = Q_{n+1} = \left(\sqrt{\frac{2}{n+2}}\sin\left(\frac{st\pi}{n+2}\right)\right)_{s,t=1}^{n+1}, \quad D = \mathrm{diag}\left(4\sin^2\left(\frac{s\pi}{2(n+2)}\right)\right).$$

Hence its conditioning $\kappa_2(\cdot)$ in spectral norm (the one induced by the Euclidean vector norm) is exactly known and it is equal to

$$\kappa_2(S_{n+1}) = \sin^2\left(\frac{(n+1)\pi}{2(n+2)}\right)\sin^{-2}\left(\frac{\pi}{2(n+2)}\right) \approx \frac{4}{\pi^2}n^2,$$

where $a_n \approx b_n$ means $a_n = b_n(1 + o(1))$ and where, in our setting, a even more precise relation can be derived, that is that is $\kappa_2(S_{n+1}) = \frac{4}{\pi^2}n^2 + O(1)$. Since everything is known regarding the term S_{n+1} our idea is to reduce the analysis as much as possible to information concerning the matrix S_{n+1} and its inverse and to this end the application of the Sherman–Morrison–Woodbury is appropriate.

The Sherman–Morrison–Woodbury formula states that for and invertible square matrix A, column vectors $\mathbf{u}$ and $\mathbf{v}$, and $1 + \mathbf{v}^{\mathrm{T}}A^{-1}\mathbf{u} \neq 0$

$$\left(A + \mathbf{u}\mathbf{v}^{\mathrm{T}}\right)^{-1} = A^{-1} - \frac{A^{-1}\mathbf{u}\mathbf{v}^{\mathrm{T}}A^{-1}}{1 + \mathbf{v}^{\mathrm{T}}A^{-1}\mathbf{u}}. \tag{10}$$

and thus we can obtain in our setting defined above in (10) with $A = S_{n+1}$ and $\mathbf{u} = \frac{\mathbf{e_1}}{h^2}$ and $\mathbf{v} = \mathbf{v}_h$.

$$\left(S_{n+1} + \frac{1}{h^2}\mathbf{e}_1\mathbf{v}_h^{\mathrm{T}}\right)^{-1} = S_{n+1}^{-1} - \frac{S_{n+1}^{-1}\frac{1}{h^2}\mathbf{e}_1\mathbf{v}_h^{\mathrm{T}}S_{n+1}^{-1}}{1 + \mathbf{v}_h^{\mathrm{T}}S_{n+1}^{-1}\frac{\mathbf{e}_1}{h^2}}.$$

or

$$\begin{aligned} A_h^{-1} &= S_{n+1}^{-1} - \frac{\frac{1}{h^2}S_{n+1}^{-1}\mathbf{e}_1\mathbf{v}_h^{\mathrm{T}}S_{n+1}^{-1}}{1 + \frac{1}{h^2}\mathbf{v}_h^{\mathrm{T}}S_{n+1}^{-1}\mathbf{e}_1} \\ &= S_{n+1}^{-1} - R_{n+1}. \end{aligned} \tag{11}$$

Our goal is to estimate quite accurately $\|A_h^{-1}\|_p$ with $p \in [1, \infty]$ and with $\|\cdot\|_p$ being the matrix norm induced by the vector norm $\|\mathbf{x}\|_p = [\sum |x_j|^p]^{1/p}$. We concentrate our efforts in the case where $p = 1, 2, \infty$, since the other estimates can be obtained via classical interpolation techniques.

We start by estimating $\|S_{n+1}^{-1}\|_1, \|S_{n+1}^{-1}\|_\infty, \|R_{n+1}\|_1, \|R_{n+1}\|_\infty$. The latter are used for giving quite precise bounds on $\|A_h^{-1}\|_1$ and $\|A_h^{-1}\|_\infty$. The estimate for $\|A_h^{-1}\|_2$ can be obtained by a direct check, but it essentially follows from the estimates on $\|A_h^{-1}\|_1$ and $\|A_h^{-1}\|_\infty$, by means of the inequality $\|A_h^{-1}\|_2 \leq \sqrt{\|A_h^{-1}\|_1\|A_h^{-1}\|_\infty}$.

2.1 Estimating $\|S_{n+1}^{-1}\|_p$ with $p = 1, \infty$

We have $S_{n+1}^{-1} = h^2 T_{n+1}^{-1}$ where $T_{n+1} = T_{n+1}(2 - 2\cos(\theta))$ and the inverse $\left(T_{n+1}^{-1}\right)_{r,c} = t_r^{(c)}$

$$T_{n+1}^{-1} = \begin{bmatrix} t_1^{(1)} & t_1^{(2)} & \dots & t_1^{(n+1)} \\ t_2^{(1)} & t_2^{(2)} & \dots & t_2^{(n+1)} \\ \vdots & \vdots & \ddots & \vdots \\ t_{n+1}^{(1)} & t_{n+1}^{(2)} & \dots & t_{n+1}^{(n+1)} \end{bmatrix} = \begin{bmatrix} \mathbf{t}^{(1)} & \mathbf{t}^{(2)} & \dots & \mathbf{t}^{(n+1)} \end{bmatrix}.$$

The components of the inverse T_{n+1}^{-1}, $t_r^{(c)}$, are defined by, for a fixed column c,

$$t_r^{(c)} = \frac{(n+2-c)r}{n+2}, \quad r = 1, \dots, c-1, \text{ for } c > 1, \tag{12}$$

$$t_r^{(c)} = \frac{(n+2-r)c}{n+2}, \quad r = c \dots, n+1, \tag{13}$$

and symmetrically for a fixed row r

$$t_r^{(c)} = \frac{(n+2-r)c}{n+2}, \quad c = 1, \dots, r-1, \text{ for } r > 1, \tag{14}$$

$$t_r^{(c)} = \frac{(n+2-c)r}{n+2}, \quad c = r, \dots, n+1. \tag{15}$$

All terms of S_{n+1}^{-1} (and T_{n+1}^{-1}) are positive and real, and they are symmetric. Hence by using

the explicit expressions of the considered norms, we find

$$\begin{aligned}\|S_{n+1}^{-1}\|_\infty &= \max_r \left\{ \sum_{c=1}^{n+1} \left(S_{n+1}^{-1}\right)_{r,c} \right\} = \max_r \left\{ h^2 \sum_{c=1}^{n+1} \left(T_{n+1}^{-1}\right)_{r,c} \right\} \\ &= \max_c \left\{ h^2 \sum_{r=1}^{n+1} \left(T_{n+1}^{-1}\right)_{r,c} \right\} = \max_c \left\{ \sum_{r=1}^{n+1} \left(S_{n+1}^{-1}\right)_{r,c} \right\} = \|S_{n+1}^{-1}\|_1. \end{aligned} \tag{16}$$

Numerically it is obvious that the highest row sum for matrices T_{n+1}^{-1} with $n+1$ even is for row index $r = (n+1)/2$ (or $r = (n+1)/2+1$, they are equal). For odd $n+1$, the highest row sum is for row index $r = (n+2)/2$.

Thus for $n+1$ even

$$\begin{aligned}\|T_{n+1}^{-1}\|_\infty &= \sum_{c=1}^{n+1} t_{(n+1)/2}^{(c)} = \sum_{c=1}^{(n-1)/2} \frac{(n+2-(n+1)/2)c}{n+2} + \sum_{c=(n+1)/2}^{n+1} \frac{(n+2-c)(n+1)/2}{n+2} \\ &= \frac{(n+1)^2 + 2(n+1)}{8} = \frac{1+2h}{8h^2}\end{aligned}$$

and for $n+1$ odd

$$\begin{aligned}\|T_{n+1}^{-1}\|_\infty &= \sum_{c=1}^{n+1} t_{(n+2)/2}^{(c)} = \sum_{c=1}^{n/2} \frac{(n+2-(n+2)/2)c}{n+2} + \sum_{c=(n+2)/2}^{n+1} \frac{(n+2-c)(n+2)/2}{n+2} \\ &= \frac{(n+2)^2}{8} = \frac{(1+h)^2}{8h^2}\end{aligned}$$

Consequently, for $n+1$ even, we deduce

$$\|S_{n+1}^{-1}\|_\infty = h^2 \|T_{n+1}^{-1}\|_\infty = \frac{1+2h}{8}, \tag{17}$$

and for $n+1$ odd, we have

$$\|S_{n+1}^{-1}\|_\infty = h^2 \|T_{n+1}^{-1}\|_\infty = \frac{1+2h+h^2}{8}. \tag{18}$$

As a conclusion, for all $n+1$ and using the symmetry and (17) and (18), we obtain that

$$\|S_{n+1}^{-1}\|_\infty = \|S_{n+1}^{-1}\|_1 \leq \frac{1+2h+h^2}{8}, \tag{19}$$

and the limit as the matrix size tends to infinity, that is $h \to 0$, is $\|S_{n+1}^{-1}\|_\infty = \|S_{n+1}^{-1}\|_1 \to \frac{1}{8}$.

2.2 Estimating $\|R_{n+1}\|_p$ for $p = 1, \infty$

Since $S_{n+1}^{-1} = h^2 T_{n+1}^{-1}$ and $T_{n+1}^{-1}\mathbf{e}_1 = \mathbf{t}^{(1)}$, we find that

$$\begin{aligned} R_{n+1} &= \frac{\frac{1}{h^2} S_{n+1}^{-1}\mathbf{e}_1 \mathbf{v}_h^{\mathrm{T}} S_{n+1}^{-1}}{1 + \frac{1}{h^2}\mathbf{v}_h^{\mathrm{T}} S_{n+1}^{-1}\mathbf{e}_1} \\ &= \frac{T_{n+1}^{-1}\mathbf{e}_1 \mathbf{v}_h^{\mathrm{T}} S_{n+1}^{-1}}{1 + \mathbf{v}_h^{\mathrm{T}} T_{n+1}^{-1}\mathbf{e}_1} \\ &= \frac{\mathbf{t}^{(1)} \mathbf{v}_h^{\mathrm{T}} S_{n+1}^{-1}}{1 + \mathbf{v}_h^{\mathrm{T}} \mathbf{t}^{(1)}}. \end{aligned} \tag{20}$$

Moreover we have from (13) that the components of $\mathbf{t}^{(1)}$ are

$$t_r^{(1)} = \frac{n+2-r}{n+2} = 1 - \frac{hr}{1+h} = \frac{1-h(r-1)}{1+h}, \quad r = 1, \ldots, n+1. \tag{21}$$

and we have from (8)

$$\mathbf{v}_h^{\mathrm{T}} = \begin{bmatrix} \vartheta h^2 - 2 & (1-\vartheta)h^2 + 1 & 0 & \ldots & 0 \end{bmatrix} = \begin{bmatrix} v_1 & v_2 & 0 & \ldots & 0 \end{bmatrix}.$$

Thus

$$\begin{aligned} \mathbf{v}_h^{\mathrm{T}} \mathbf{t}^{(1)} &= v_1 t_1^{(1)} + v_2 t_2^{(1)} = (\vartheta h^2 - 2)\frac{1}{1+h} + ((1-\vartheta)h^2 + 1)\frac{1-h}{1+h} \\ &= \frac{(\vartheta - 1)h^3 + h^2 - h - 1}{1+h}, \end{aligned}$$

and

$$1 + \mathbf{v}_h^{\mathrm{T}} \mathbf{t}^{(1)} = \frac{h^2\left((\vartheta - 1)h + 1\right)}{1+h}. \tag{22}$$

Also

$$\begin{aligned} \mathbf{v}_h^{\mathrm{T}} S_{n+1}^{-1} &= h^2 \mathbf{v}_h^{\mathrm{T}} T_{n+1}^{-1} \\ &= h^2 \begin{bmatrix} v_1 t_1^{(1)} + v_2 t_2^{(1)} & v_1 t_1^{(2)} + v_2 t_2^{(2)} & \ldots & v_1 t_1^{(c)} + v_2 t_2^{(c)} & \ldots & v_1 t_1^{(n+1)} + v_2 t_2^{(n+1)} \end{bmatrix}. \end{aligned}$$

and thus the components of the row vector $\left(\mathbf{v}_h^{\mathrm{T}} S_{n+1}^{-1}\right)_c$ are

$$\left(\mathbf{v}_h^{\mathrm{T}} S_{n+1}^{-1}\right)_c = h^2 \left(v_1 t_1^{(c)} + v_2 t_2^{(c)}\right),$$

and the components of the matrix $\left(\mathbf{t}^{(1)} \mathbf{v}_h^{\mathrm{T}} S_{n+1}^{-1}\right)_{r,c}$ are

$$\left(\mathbf{t}^{(1)} \mathbf{v}_h^{\mathrm{T}} S_{n+1}^{-1}\right)_{r,c} = t_r^{(1)} h^2 \left(v_1 t_1^{(c)} + v_2 t_2^{(c)}\right), \tag{23}$$

where $t_r^{(1)}$ is defined in (21), and

$$t_1^{(c)} = \frac{1-h(c-1)}{1+h}, \quad c = 1, \ldots, n+1, \tag{24}$$

$$t_2^{(c)} = \begin{cases} \frac{1-h}{1+h}, & c = 1, \\ \frac{2-2h(c-1)}{1+h}, & c > 1, \end{cases} \tag{25}$$

are defined in (14) and (15). Therefore for $c = 1$

$$\begin{aligned} \left(\mathbf{t}^{(1)}\mathbf{v}_h^{\mathrm{T}} S_{n+1}^{-1}\right)_{r,1} &= t_r^{(1)} h^2 \left(v_1 t_1^{(1)} + v_2 t_2^{(1)}\right) = t_r^{(1)} h^2 \mathbf{v}_h^{\mathrm{T}} \mathbf{t}^{(1)} \\ &= \frac{1-h(r-1)}{1+h} h^2 \frac{(\vartheta-1)h^3 + h^2 - h - 1}{1+h}, \\ &= \frac{h^2(1-h(r-1))((\vartheta-1)h^3 + h^2 - h - 1)}{(h+1)^2}, \end{aligned} \tag{26}$$

and for $c > 1$

$$\begin{aligned} \left(\mathbf{t}^{(1)}\mathbf{v}_h^{\mathrm{T}} S_{n+1}^{-1}\right)_{r,c} &= t_r^{(1)} h^2 \left(v_1 t_1^{(c)} + v_2 t_2^{(c)}\right) = t_r^{(1)} h^2 \mathbf{v}_h^{\mathrm{T}} \mathbf{t}^{(c)} \\ &= \frac{1-h(r-1)}{1+h} h^2 \left(\left(\vartheta h^2 - 2\right) \frac{1-h(c-1)}{1+h} + \left((1-\vartheta)h^2 + 1\right) \frac{2-2h(c-1)}{1+h}\right) \\ &= \frac{h^4(2-\vartheta)(1-h(c-1))(1-h(r-1))}{(h+1)^2} \end{aligned} \tag{27}$$

Thus we can now define the components of $(R_{n+1})_{r,c} = \left(\frac{\mathbf{t}^{(1)}\mathbf{v}_h^{\mathrm{T}} S_{n+1}^{-1}}{1+\mathbf{v}_h^{\mathrm{T}}\mathbf{t}^{(1)}}\right)_{r,c}$, defined in (20), since we have (22), (26), and (27). For $c = 1$ we have

$$\begin{aligned} (R_{n+1})_{r,1} &= \frac{h^2(1-h(r-1))((\vartheta-1)h^3 + h^2 - h - 1)}{(h+1)^2} \Big/ \frac{h^2\left((\vartheta-1)h+1\right)}{1+h} \\ &= \frac{h(r-1)-1}{h(\vartheta-1)+1} - \frac{h^2(h(r-1)-1)}{h+1} \end{aligned} \tag{28}$$

and for $c > 1$

$$\begin{aligned} (R_{n+1})_{r,c} &= \\ &= \frac{h^4(2-\vartheta)(1-h(c-1))(1-h(r-1))}{(h+1)^2} \Big/ \frac{h^2\left((\vartheta-1)h+1\right)}{1+h} \\ &= \frac{h^2(2-\vartheta)(1-(c-1)h)(1-h(r-1))}{(h+1)(h(\vartheta-1)+1)} \end{aligned} \tag{29}$$

Numerically it is obvious that $\|R_{n+1}\|_1$ and $\|R_{n+1}\|_\infty$ is always for the first column and first row.

Now we compute $\|R_{n+1}\|_1$,

$$\begin{aligned}
\|R_{n+1}\|_1 &= \sum_{r=1}^{n+1} \left|(R_{n+1})_{r,1}\right| \\
&= \sum_{r=1}^{n+1} \left| \frac{h(r-1)-1}{h(\vartheta-1)+1} - \frac{h^2(h(r-1)-1)}{h+1} \right| \\
&= \sum_{r=1}^{n+1} \left| \frac{(1-h(r-1))(h^3(\vartheta-1)+h^2-h-1)}{(h+1)(h(\vartheta-1)+1)} \right| \\
&= \frac{-h^3(\vartheta-1)-h^2+h+1}{(h+1)(h(\vartheta-1)+1)} \sum_{r=1}^{n+1} (1+h-hr)) \\
&= \frac{-h^3(\vartheta-1)-h^2+h+1}{(h+1)(h(\vartheta-1)+1)} \left((n+1)(1+h) - h\frac{(n+1)(n+2)}{2} \right) \\
&= \frac{-h^3(\vartheta-1)-h^2+h+1}{(h+1)(h(\vartheta-1)+1)} \left(\frac{1+h}{h} - \frac{1+h}{2h} \right) \\
&= \frac{h^3(1-\vartheta)-h^2+h+1}{2h(h(\vartheta-1)+1)}, \qquad (30)
\end{aligned}$$

since $0 < \vartheta < 1$.

We now compute $\|R_{n+1}\|_\infty$, by taking into account that all coefficients are positive except the first in the first column. We have from (28) and (29)

$$(R_{n+1})_{1,1} = \frac{h^2}{h+1} - \frac{1}{h(\vartheta-1)+1} \qquad (31)$$

$$(R_{n+1})_{1,c} = \frac{h^2(2-\vartheta)(1-(c-1)h)}{(h+1)(h(\vartheta-1)+1)}, \quad c = 2, \ldots, n+1 \qquad (32)$$

Thus

$$
\begin{aligned}
\|R_{n+1}\|_\infty &= -\left(R_{n+1}\right)_{1,1} + \sum_{c=2}^{n+1} \left(R_{n+1}\right)_{1,c} \\
&= -\left(\frac{h^2}{h+1} - \frac{1}{h(\vartheta-1)+1}\right) + \sum_{c=2}^{n+1} \frac{h^2(2-\vartheta)(1-(c-1)h)}{(h+1)(h(\vartheta-1)+1)} \\
&= -\left(\frac{h^2}{h+1} - \frac{1}{h(\vartheta-1)+1}\right) + \frac{h^2(2-\vartheta)}{(h+1)(h(\vartheta-1)+1)} \sum_{c=2}^{n+1} (1+h-ch) \\
&= -\left(\frac{h^2}{h+1} - \frac{1}{h(\vartheta-1)+1}\right) + \frac{h^2(2-\vartheta)}{(h+1)(h(\vartheta-1)+1)} \left((n+1-1)(1+h) - h\left(\frac{(n+1)(n+2)}{2} - 1\right)\right) \\
&= -\left(\frac{h^2}{h+1} - \frac{1}{h(\vartheta-1)+1}\right) + \frac{h^2(2-\vartheta)}{(h+1)(h(\vartheta-1)+1)} \left(\frac{1-h^2}{h} - \frac{1+h-2h^2}{2h}\right) \\
&= -\left(\frac{h^2}{h+1} - \frac{1}{h(\vartheta-1)+1}\right) + \frac{1}{2}\frac{h(2-\vartheta)(1-h)}{(h+1)(h(\vartheta-1)+1)} \\
&= \frac{1}{2}\frac{h(2-\vartheta)(1-h) - 2h^2(h(\vartheta-1)+1) + 2(h+1)}{(h+1)(h(\vartheta-1)+1)} \\
&\to 1 \text{ as } h \to 0.
\end{aligned}
$$

2.3 Estimating $\|A_h^{-1}\|_p$ for $p = 1, \infty$

Numerically it is obvious that $\|A_h^{-1}\|_1$ is computed on the first column, thus since $A_h^{-1} = S_{n+1}^{-1} - R_{n+1}$, S_{n+1}^{-1} and R_{n+1} positive and negative respectively, we can just compute the norm directly for A_h^{-1}. The sum of the positive elements of the first column of T_{n+1}^{-1} is equal to $\frac{n+1}{2}$, and thus the sum for the first column of S_{n+1}^{-1} is $\frac{h^2(n+1)}{2} = \frac{h}{2}$, and the sum of components $-(R_{n+1})_{r,1}$ is given in (30), that is

$$
\begin{aligned}
\|A_h^{-1}\|_1 &= \frac{h}{2} + \frac{h^3(1-\vartheta) - h^2 + h + 1}{2h(h(\vartheta-1)+1)} \\
&= \frac{h^3(\vartheta-1) + h^2 + h^3(1-\vartheta) - h^2 + h + 1}{2h(h(\vartheta-1)+1)} \\
&= \frac{h+1}{2h(h(\vartheta-1)+1)}
\end{aligned} \tag{33}
$$

Now we compute $\|A_h^{-1}\|_\infty$. $A_h^{-1} = S_{n+1}^{-1} - R_{n+1}$. We have from (15)

$$
\left(S_{n+1}^{-1}\right)_{1,c} = h^2 t_1^{(c)} = h^2 \frac{1+h(1-c)}{1+h}, \quad c = 1, \ldots, n+1
$$

and by using (31) we get

$$\begin{aligned}\left(A_h^{-1}\right)_{1,1} &= (S_{n+1}^{-1})_{1,1} - (R_{n+1})_{1,1} \\ &= \frac{h^2}{1+h} - \left(\frac{h^2}{h+1} - \frac{1}{h(\vartheta-1)+1}\right) \\ &= \frac{1}{h(\vartheta-1)+1} \end{aligned} \tag{34}$$

and by (32)

$$\begin{aligned}\left(A_h^{-1}\right)_{1,c} &= (S_{n+1}^{-1})_{1,c} - (R_{n+1})_{1,c} \\ &= h^2\frac{1+h(1-c)}{1+h} - \frac{h^2(2-\vartheta)(1-(c-1)h)}{(h+1)(h(\vartheta-1)+1)} \\ &= \frac{h^2(1+h(1-c))}{1+h}\left(1 - \frac{2-\vartheta}{h(\vartheta-1)+1}\right) \\ &= \frac{h^2(1+h(1-c))(\vartheta-1)}{h(\vartheta-1)+1}. \end{aligned} \tag{35}$$

Since $\left(A_h^{-1}\right)_{1,1}$ in (34) is always positive and $\left(A_h^{-1}\right)_{1,c}$ of (35) is always negative we have

$$\begin{aligned}\|A_h^{-1}\|_\infty &= \left(A_h^{-1}\right)_{1,1} - \sum_{c=2}^{n+1}\left(A_h^{-1}\right)_{1,c} \\ &= \frac{1}{h(\vartheta-1)+1} - \sum_{c=2}^{n+1}\frac{h^2(1+h(1-c))(\vartheta-1)}{h(\vartheta-1)+1} \\ &= \frac{1}{h(\vartheta-1)+1} - \frac{h^2(\vartheta-1)}{h(\vartheta-1)+1}\sum_{c=2}^{n+1}(1+h-hc) \\ &= \frac{1}{h(\vartheta-1)+1} - \frac{h^2(\vartheta-1)}{h(\vartheta-1)+1}\left((n+1-1)(1+h) - h\left(\frac{(n+1)(n+2)}{2} - 1\right)\right) \\ &= \frac{1}{h(\vartheta-1)+1} - \frac{h^2(\vartheta-1)}{h(\vartheta-1)+1}\left(\frac{1-h^2}{h} - \frac{1+h-2h^2}{2h}\right) \\ &= \frac{2-h(\vartheta-1)(1-h)}{2(h(\vartheta-1)+1)} \end{aligned} \tag{36}$$

As a conclusion we deduce from (33) and (36)

$$\begin{aligned}
\|A_h^{-1}\|_2 &\le \sqrt{\|A_h^{-1}\|_1 \|A_h^{-1}\|_\infty} \\
&= \sqrt{\frac{h+1}{2h(h(\vartheta-1)+1)} \frac{2-h(\vartheta-1)(1-h)}{2(h(\vartheta-1)+1)}} \\
&= \frac{1}{2(h(\vartheta-1)+1)} \sqrt{\frac{h+1}{h}(2-h(\vartheta-1)(1-h))} \\
&= \frac{1}{2(h(\vartheta-1)+1)} \sqrt{\frac{2(h+1)+(h^2-1)h(\vartheta-1)}{h}} \\
&= \frac{1}{2(h(\vartheta-1)+1)} \sqrt{\frac{2}{h}+2+(h^2-1)(\vartheta-1)}
\end{aligned}$$

In order to make a comparison, we recall that we know the exact asymptotical behavior of $\|S_{n+1}^{-1}\|_2$, with S_{n+1} being the pure Toeplitz counterpart of A_h, as reported below

$$\|S_{n+1}^{-1}\|_2 = \frac{1}{\lambda_{\min}(S_{n+1})} = \frac{h^2}{4\sin^2\left(\frac{\pi}{2(n+2)}\right)} = \left(\frac{h}{2\sin\left(\frac{\pi h}{2(1+h)}\right)}\right)^2 \overset{h\to 0}{\to} \frac{1}{\pi^2}. \tag{37}$$

2.4 Spectral results: comments

Here we give a short discussion on few items that, for some aspects, will be considered in more detail in Section 3 and for other aspects will be listed as open problems in the conclusion Section 4.

- The estimates for $\|A_h^{-1}\|_p$ are tight and the growth is like $n^{1/p}$: however the numerical growth of the error seems to be bounded by a constant independently of p. The reason relies on the vectors for which the norm is attained. Such vectors should be concentrated on the first component and this is quite unphysical and it is not observed in practice.
- Even if A_h and its inverse are not symmetric we can prove the spectrum of the related matrix-sequence is clustered along a real positive interval, using the results of the GLT technology reported in Subsection 3.2 (see also [1, 11]): we refer to Subsection 3.4 where the analysis is performed both in $1\mathbb{D}$ and $2\mathbb{D}$.
- Regarding the estimates of $\|A_h^{-1}\|_p$, the $2\mathbb{D}$ case (and generically the $d\mathbb{D}$ case) is more difficult, but we can take advantage of the one dimensional case and from a clever tensor structure of the problem when the domain is rectangular (hyper-rectangular in the $d\mathbb{D}$ case).
- When the domain is generic a possibility is given by embedding techniques already exploited in the distributional setting via the GLT approach (see [16, 17]).

2.5 Numerical tests in $1\mathbb{D}$

We consider the $1\mathbb{D}$ problem (1) with $a = 0$ and $b = \pi$. We choose $f = -\sin(x)$, $g_a = 0$, and $g_b = 0$ so that $\mathbf{u} = -\sin(x_i)$ is the exact solution in points x_i.

We perform several tests varying the value of $\vartheta \in [0,1]$, in order to establish whether the convergence of the method depends on the choice of ϑ. In practice, we choose ϑ and n and we compute h and x_0 accordingly:

$$h = \frac{b-a}{n+\vartheta}, \qquad x_0 = b - (n+1)h.$$

The numerical error $\mathbf{e}_h = \mathbf{u} - \mathbf{u}_h$ satisfies the following equation:

$$A_h \mathbf{e}_h = \tau_h,$$

where τ_h is the consistency error:

$$\tau_h = \mathbf{f}_h - A_h \mathbf{u}.$$

Consider the $p-$norm:

$$\|\tau_h\|_{L^p} \approx \left(h \sum_{i=0}^{n} |\tau_h(x_i)|^p \right)^{\frac{1}{p}}, \tag{38}$$

$$\|\mathbf{e}_h\|_{L^p} \approx \left(h \sum_{i=0}^{n} |\mathbf{e}_h(x_i)|^p \right)^{\frac{1}{p}}, \tag{39}$$

In Fig. 2 we show that:

$$\|\tau_h\|_{L^p}, \|\mathbf{e}_h\|_{L^p} \approx O(h^2), \text{ for } p = 1, 2, \infty, \tag{40}$$

confirming that the method is second-order consistent and accurate.

We complete the analysis showing the behaviour of the spectral radius of the matrix A_h^{-1}. Fig. 2.5 shows how the smallest eigenvalue (in absolute value) of the matrix A_h changes in relation to n (left panel) and in relation to ϑ (right panel).

Fig. 2.5 shows that the smallest eigenvalue (in absolute value) of the matrix A_h essentially does not depend on the value of ϑ and it approaches a constant values when n goes to infinity.

Since $\|\mathbf{e}_h\|_{L^p} \leq \left\|A_h^{-1}\right\|_p \|\tau_h\|_{L^p}$, we can conclude that $\|\mathbf{e}_h\|_{L^p} \approx O(h^2)$ and $\left\|A_h^{-1}\right\|_p \|\tau_h\|_{L^p} \approx O(h^{2-\frac{1}{p}})$, as predicted in the first item of Subsection 2.4.

3 Problem formulation in $2\mathbb{D}$ and related analysis

The section is organized into three parts: first we introduce the d-level notation and the d-level Toeplitz matrices in Subsection 3.1, secondly we define the notion of spectral and singular value distribution and the $*$-algebra of Generalized Locally Toeplitz matrix-sequences in Subsection 3.2, then we describe the matrices arising in the approximation of a Dirichlet problem by the Coco–Russo method in Subsection 3.3, and finally we give a spectral analysis of the resulting matrix-sequences in Subsection 3.4.

3.1 Multilevel notation: the case of multilevel Toeplitz and diagonal sampling matrices

We start by introducing the multi-index notation, which is useful in our context. A multi-index $\mathbf{i} \in \mathbb{Z}^d$, also called a d-index, is simply a (row) vector in $\mathbb{Z}^d$; its components are denoted by $i_1, \ldots, i_d$.

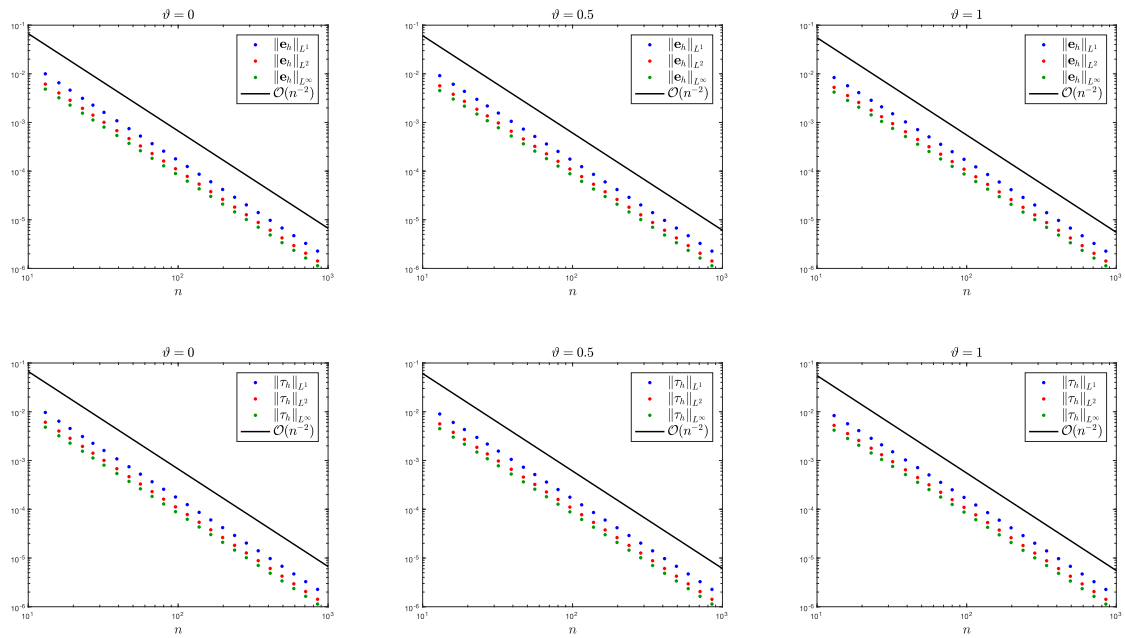

Figure 2: The dots represent the p - norm of the numerical error $\mathbf{e}_h$ (top) and consistency error τ_h (bottom) for different values of n (horizontal axis) and ϑ: $\vartheta = 0$ (left), $\vartheta = 0.5$ (middle), $\vartheta = 1$ (right). The solid line is a reference for second-order decay.

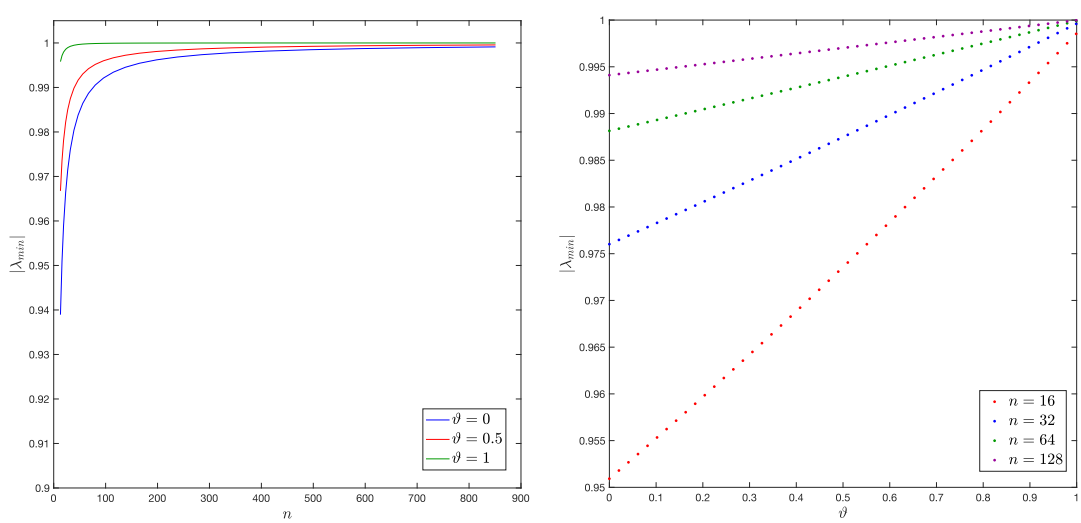

Figure 3: Smallest eigenvalue in absolute value (vertical axis) for different values of n (horizontal axis, left plot) or for different values of ϑ (horizontal axis, right plot).

- $\mathbf{0}$, $\mathbf{1}$, $\mathbf{2}$, ... are the vectors of all zeros, all ones, all twos, ... (their size will be clear from the context).
- For any d-index $\mathbf{m}$, we set $N(\mathbf{m}) = \prod_{j=1}^{d} m_j$ and we write $\mathbf{m} \to \infty$ to indicate that $\min(\mathbf{m}) \to \infty$.
- If $\mathbf{h}, \mathbf{k}$ are d-indices, $\mathbf{h} \le \mathbf{k}$ means that $h_r \le k_r$ for all $r = 1, \ldots, d$.
- The standard lexicographic ordering is assumed uniformly

$$\Big[\ldots \Big[[\ (j_1, \ldots, j_d)\ ]_{j_d = h_d, \ldots, k_d} \Big]_{j_{d-1} = h_{d-1}, \ldots, k_{d-1}} \ldots \Big]_{j_1 = h_1, \ldots, k_1}. \tag{41}$$

For instance, in the case $d = 2$ the ordering is the following: (h_1, h_2), $(h_1, h_2 + 1)$, ..., (h_1, k_2), $(h_1 + 1, h_2)$,

Multilevel Toeplitz Matrices.

We now briefly summarize the definition and few relevant properties of multilevel Toeplitz matrices, that we will employ in the analysis of the $2\mathbb{D}$ setting. Given $\mathbf{n} \in \mathbb{N}^d$, a matrix of the form

$$[a_{\mathbf{i}-\mathbf{j}}]_{\mathbf{i},\mathbf{j}=\mathbf{e}}^{\mathbf{n}} \in \mathbb{C}^{N(\mathbf{n}) \times N(\mathbf{n})}$$

with $\mathbf{e}$ vector of all ones, with entries $a_{\mathbf{k}} \in \mathbb{C}$, $\mathbf{k} = -(\mathbf{n}-\mathbf{e}), \ldots, \mathbf{n}-\mathbf{e}$, is called a multilevel Toeplitz matrix, or, more precisely, a d-level Toeplitz matrix. Let $\phi : [-\pi, \pi]^d \to \mathbb{C}^{r \times r}$ a matrix-valued function in which each entry belongs to $L^1([-\pi, \pi]^d)$. We denote the Fourier coefficients of the generating function ϕ as

$$\hat{\phi}_{\mathbf{k}} = \frac{1}{(2\pi)^d} \int_{[-\pi,\pi]^d} \phi(\boldsymbol{\theta}) e^{-\hat{\imath}(\mathbf{k},\boldsymbol{\theta})}\, d\boldsymbol{\theta} \in \mathbb{C}, \quad \mathbf{k} \in \mathbb{Z}^d,$$

where the integrals are computed component-wise and $(\mathbf{k}, \boldsymbol{\theta}) = k_1\theta_1 + \ldots + k_d\theta_d$. For every $\mathbf{n} \in \mathbb{N}^d$, the $\mathbf{n}$-th Toeplitz matrix associated with ϕ is defined as

$$T_{\mathbf{n}}(\phi) := [\hat{\phi}_{\mathbf{i}-\mathbf{j}}]_{\mathbf{i},\mathbf{j}=\mathbf{e}}^{\mathbf{n}}$$

or, equivalently, as

$$T_{\mathbf{n}}(\phi) = \sum_{|j_1| < n_1} \ldots \sum_{|j_d| < n_d} \hat{\phi}_{(j_1,\ldots,j_d)} [J_{n_1}^{(j_1)} \otimes \ldots \otimes J_{n_d}^{(j_d)}], \tag{42}$$

where $\otimes$ denotes the (Kronecker) tensor product of matrices, while $J_m^{(l)}$ is the matrix of order m whose (i, j) entry equals 1 if $i - j = l$ and zero otherwise. We call $\{T_{\mathbf{n}}(\phi)\}_{\mathbf{n} \in \mathbb{N}^d}$ the family of (multilevel block) Toeplitz matrices associated with ϕ, which, in turn, is called the generating function of $\{T_{\mathbf{n}}(\phi)\}_{\mathbf{n} \in \mathbb{N}^d}$.

Multilevel Diagonal Sampling Matrices. For $n \in \mathbb{N}$ and $a : [0, 1] \to \mathbb{C}$, we define the diagonal sampling matrix $D_n(a)$ as the diagonal matrix

$$D_n(a) = \operatorname*{diag}_{i=1,\ldots,n} a\Big(\frac{i}{n}\Big) = \begin{bmatrix} a(\frac{1}{n}) & & & \\ & a(\frac{2}{n}) & & \\ & & \ddots & \\ & & & a(1) \end{bmatrix} \in \mathbb{C}^{n \times n}.$$

For $\mathbf{n} \in \mathbb{N}^d$ and $a : [0,1]^d \to \mathbb{C}$, we define the multilevel diagonal sampling matrix $D_{\mathbf{n}}(a)$ as the diagonal matrix

$$D_{\mathbf{n}}(a) = \underset{\mathbf{i}=\mathbf{1},\dots,\mathbf{n}}{\mathrm{diag}}\ a\Big(\frac{\mathbf{i}}{\mathbf{n}}\Big) \in \mathbb{C}^{N(\mathbf{n})\times N(\mathbf{n})},$$

with the lexicographical ordering (41) as discussed at the beginning of the subsection.

3.2 GLT matrix-sequences: operative features

We start with the definition of distribution in the sense of the eigenvalues (spectral distribution) and in the sense of the singular values (singular value distribution) for a given matrix-sequence. Then we give the operative feature of the $*$-algebra of matrix-sequences.

Definition 1. *Let $\{A_n\}_n$ be a sequence of matrices, with A_n of size d_n, and let $f : D \subset \mathbb{R}^t \to \mathbb{C}$ be a measurable function defined on a set D with $0 < \mu_t(D) < \infty$.*

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

$$\lim_{n\to\infty} \frac{1}{d_n} \sum_{i=1}^{d_n} F(\sigma_i(A_n)) = \frac{1}{\mu_t(D)} \int_D F(|f(\mathbf{x})|)\ d\mathbf{x}, \quad \forall\, F \in C_c(\mathbb{R}). \tag{43}$$

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

$$\lim_{n\to\infty} \frac{1}{d_n} \sum_{i=1}^{d_n} F(\lambda_i(A_n)) = \frac{1}{\mu_t(D)} \int_D F(f(\mathbf{x}))\ d\mathbf{x}, \quad \forall\, F \in C_c(\mathbb{C}). \tag{44}$$

If $\{A_n\}_n$ has both a singular value and an eigenvalue distribution described by f, then we write $\{A_n\}_n \sim_{\sigma,\lambda} f$.

The symbol f contains spectral/singular value information briefly described informally as follows. With reference to (44), assuming that d_n is large enough and f is at least Riemann integrable, except possibly for a small number of outliers, the eigenvalues of A_n are approximately formed by the samples of f over a uniform grid in D, so that the range of f is a (weak) cluster for the eigenvalues of $\{A_n\}_n$. It is then clear that the symbol f provides a 'compact' and a quite accurate description of the spectrum of the matrices A_n for n large enough. Relation (43) has the same meaning when talking of the singular values of A_n and by replacing f with $|f|$.

A d-level ($d \geq 1$ integer) GLT matrix-sequence $\{A_n\}_n$ is nothing more than a matrix–sequence endowed with a measurable function $\kappa : [0,1]^d \times [-\pi,\pi]^d \to \mathbb{C}$ called <u>symbol</u> characterizing the distributional properties of its singular values, and, under certain hypothesis, of its spectrum. For a complete overview of the theory we refer to the books [7, 8], while here we recall only the operative features we need for our restricted setting. Since we have already introduced the multilevel Toeplitz and diagonal matrix-sequences, the only other class we need is that of zero–distributed matrix-sequences, whose definition depends on Definition 1.

Definition 2. *[Zero–distributed sequence] A matrix-sequence $\{Z_n\}_n$ such that $\{Z_n\}_n \sim_\sigma 0$ is referred to as a zero-distributed sequence. In other words, $\{Z_n\}_n$ is zero-distributed if and only if*

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

In a different language, more common in the context of preconditioning and of the convergence analysis of (preconditioned) Krylov methods, a zero–distributed matrix-sequence is a sequence of matrices showing a (weak) clustering at zero in the sense of the singular values (see e.g.[7, 19] and references therein).

With the notaion indicating by $\|\cdot\|$ the spectral norm (i.e. the maximal singular value or equivalently the induced Euclidean norm) and by $\|\cdot\|_1$ the trace norm (i.e. the sum of all singular values), the following result holds true [7].

Theorem 3. GLT 1. *If $\{A_n\}_n \sim_{\mathrm{GLT}} \kappa$ then $\{A_n\}_n \sim_\sigma \kappa$. If $\{A_n\}_n \sim_{\mathrm{GLT}} \kappa$ and the matrices A_n are Hermitian then $\{A_n\}_n \sim_\lambda \kappa$.*

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

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

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

GLT 3. *We have*

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

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

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

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

A more general and more advanced result regarding item **GLT2** can be found in [1, 11], even if for our purposes item **GLT2** is sufficient in our setting.

3.3 Coco–Russo method in $2\mathbb{D}$: Dirichlet problem in a square domain

We consider the following Dirichlet problem:

$$\begin{cases} -u_{xx} - u_{yy} = f & \text{in } \Omega \\ u = g & \text{in } \partial\Omega, \end{cases} \tag{45}$$

where $\Omega = [0,1] \times [a,1] \subset [0,1]^2$, $f, g : \Omega \to \mathbb{R}$ are assigned functions and $u : \Omega \to \mathbb{R}$ is the unknown function.

The square $[0,1]^2$ is discretized through a uniform Cartesian grid with $(n+2)^2$ grid points $(x_i, y_j) = (i\,h, j\,h)$, for $i, j = 0, \ldots, n+1$, where $h = 1/(n+1)$. As in the $1\mathbb{D}$ case, let $0 < a < h$ and call $\vartheta_{\mathrm{S}} = (x_1 - a)/h$ (see Fig. 4). The subscript S stands for south, since the boundary $y = a$ is the bottom side of the domain. A similar approach can be followed in the other cases.

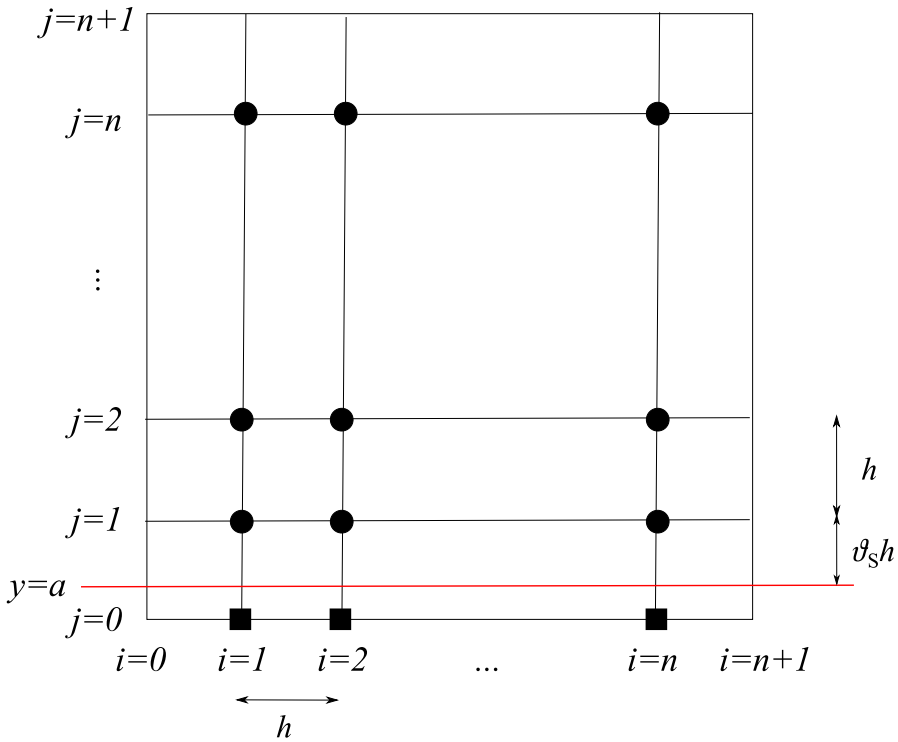

Figure 4: Discretization of the $2\mathbb{D}$ domain. Full circles are the n^2 inside grid points, while full squares are the n ghost points. Linear extrapolation is used to define the ghost values $u_{i,0}$ from $u_{i,1}$ and the boundary values $g(x_i, 0)$, for $i = 1, \ldots, n$.

The elliptic equation $-\Delta u = f$ of problem (45) is discretized by central finite difference on internal grid points, with eliminated boundary conditions on the boundaries $x = 0$, $x = 1$ and $y = 1$. Then, for $2 \le i \le n-1$ and $1 \le j \le n-1$ we have:

$$\frac{4u_{ij} - (u_{i-1\,j} + u_{i+1\,j} + u_{i\,j-1} + u_{i\,j+1})}{h^2} = f_{ij},$$

while for for $i = 1$ and $j = 1, \ldots, n-1$ we eliminate the boundary condition on $x = 0$:

$$\frac{4u_{1\,j} - (u_{2\,j} + u_{1\,j-1} + u_{1\,j+1})}{h^2} = f_{ij} + \frac{g(0, y_j)}{h^2}.$$

Similarly, we elimiante the boundary conditions on $x = 1$ and $y = 1$. The boundary condition on $y = a$ is discretized by linear interpolation:

$$\vartheta_{\mathrm{s}} u_{i,0} + (1 - \vartheta_{\mathrm{s}}) u_{i,1} = g(x_i, a), \qquad \text{for } i = 1, \dots, n.$$

Overall, there are n^2 inside grid points (x_i, y_j) for $i, j = 1, \dots, n$ and n ghost points for $(x_i, 0)$ for $i = 1 \dots, n$.

Using a total lexicographical order, the matrix of coefficients that we obtain is a 2-level matrix with the following structure:

$$A_h = \left(\begin{array}{c|c|c|c|c|c}
\vartheta_{\mathrm{s}}\mathbb{I}_n & (1-\vartheta_{\mathrm{s}})\mathbb{I}_n & & & & \\ \hline
B & G & B & & & \\ \hline
 & B & G & B & & \\ \hline
 & & \ddots & \ddots & \ddots & \\ \hline
 & & & B & G & B \\ \hline
 & & & & B & G
\end{array}\right),$$

where

$$\vartheta_{\mathrm{s}}\mathbb{I}_n \in \mathbb{R}^{n\times n},$$
$$(1-\vartheta_{\mathrm{s}})\mathbb{I}_n \in \mathbb{R}^{n\times n},$$
$$B = -\frac{1}{h^2}\mathbb{I}_n \in \mathbb{R}^{n\times n},$$

and

$$G = \frac{1}{h^2}\begin{pmatrix} 4 & -1 & & & \\ -1 & 4 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 4 & -1 \\ & & & -1 & 4 \end{pmatrix} \in \mathbb{R}^{n\times n},$$

A_h has n blocks of G, so $A_h \in \mathbb{R}^{n(n+1)\times n(n+1)}$.

3.4 Spectral analysis in $1\mathbb{D}$ and in $2\mathbb{D}$

Having in mind the notations of Subsection 3.1, the matrix A_h can be decomposed in the following way

$$A_h = \frac{1}{h^2}\left[T_{\mathbf{n}}(f) + X_{\mathbf{n}}\right] \tag{46}$$

where $\mathbf{n} = (n+1, n)$, the size of A_h is $N(\mathbf{n}) = n(n+1)$,

$$T_{\mathbf{n}}(f) = T_{n+1}(2 - 2\cos(\theta)) \otimes \mathbb{I}_n + \mathbb{I}_{n+1} \otimes T_n(2 - 2\cos(\theta)), \tag{47}$$

$T_k(2 - 2\cos(\theta))$ is a Toeplitz matrix, already used in the $1\mathbb{D}$ case in Section 2, and

$$X_{\mathbf{n}} = \left[\begin{array}{c|c|c} T_n\left(h^2\vartheta_S - 4 + 2\cos(\theta)\right) & T_n\left(h^2(1-\vartheta_S) + 1\right) & \mathbf{0}_{n\times n(n-1)} \\ \hline \mathbf{0}_{n^2\times n} & \mathbf{0}_{n^2\times n} & \mathbf{0}_{n^2\times n(n-1)} \end{array}\right]. \tag{48}$$

Of course, taking into account relation (42) with $d = 2$ and (47), the function f is bivariate and can be written as

$$f(\theta_1, \theta_2) = 4 - 2\cos(\theta_1) - 2\cos(\theta_2).$$

Therefore by using item **GLT1** in Theorem 3 we have

$$\{T_{\mathbf{n}}(f)\} \sim_{\mathrm{GLT}} f$$

in the sense of of Subsection 3.2, so that

$$\{T_{\mathbf{n}}(f)\} \sim_{\sigma} f,$$

according to Definition 1. Furthermore, since $T_{\mathbf{n}}(f)$ is Hermitian (in fact real symmetric) for any choice of the partial sizes, thanks to item **GLT1**, we deduce $\{T_{\mathbf{n}}(f)\} \sim_{\lambda} f$ as well.

Now, taking into account Definition 2, it is easy to see that $\{X_{\mathbf{n}}\}$ ia a zero–distributed matrix-sequence, sinply because its rank is bounded by n and hence the number of nonzero singular values is at most $n = o(n(n+1))$ with $N(\mathbf{n}) = n(n+1)$ being the sinze of $X_{\mathbf{n}}$. Therefore by item **GLT3**

$$\{X_{\mathbf{n}}\} \sim_{\mathrm{GLT}} 0,$$

so that $\{h^2 A_h\} \sim_{\mathrm{GLT}} f$ by item **GLT4**, since both $\{T_{\mathbf{n}}(f)\}, \{X_{\mathbf{n}}\}$ are GLT matrix-sequences and $h^2 A_h = T_{\mathbf{n}}(f) + X_{\mathbf{n}}$ for any choice of the partial sizes. Then, again by item **GLT1** we deduce

$$\{h^2 A_h\} \sim_{\sigma} f.$$

However, $X_{\mathbf{n}}$ is non-Hermitian and therefore we cannot apply item **GLT1** for concluding $\{h^2 A_h\} \sim_{\lambda} f$. However, this can be done by using item **GLT2**, as proven in the following lines both in $1\mathbb{D}$ and in $2\mathbb{D}$.

Theorem 4. *With the notations used so far in $1\mathbb{D}$ we have*

$$\{h^2 A_h\} \sim_{\lambda} 2 - 2\cos(\theta), \tag{49}$$

while in $2\mathbb{D}$ we have

$$\{h^2 A_h\} \sim_{\lambda} 4 - 2\cos(\theta_1) - 2\cos(\theta_2). \tag{50}$$

Proof In $1\mathbb{D}$ we recall the identity

$$h^2 A_h = T_n(2 - 2\cos(\theta)) + \mathbf{e}_1 \mathbf{v}_h^{\mathrm{T}}.$$

Since $\mathbf{e}_1 \mathbf{v}_h^{\mathrm{T}}$ is a rank one matrix, it has a unique nozero singular value so that

$$\|\mathbf{e}_1 \mathbf{v}_h^{\mathrm{T}}\|_1 = \|\mathbf{e}_1 \mathbf{v}_h^{\mathrm{T}}\| = \|\mathbf{v}_h^{\mathrm{T}}\|_2$$

and hence a trivial computation shows that

$$\lim_{n\to\infty} \frac{\|\mathbf{e}_1 \mathbf{v}_h^{\mathrm{T}}\|_1}{n} = 0.$$

Therefore, by item **GLT2**, we infer that both the GLT matrix sequences $\{h^2 A_h\}, \{T_n(2 - 2\cos(\theta))\}$ share the same eigenvalue distribution function $2 - 2\cos(\theta)$, which is the GLT symbol, so that (49) is proven.

In $2\mathbb{D}$, according to the 2-level notation, we remind that

$$h^2 A_h = T_{\mathbf{n}}(4 - 2\cos(\theta_1) - 2\cos(\theta_2)) + X_{\mathbf{n}}, \quad \mathbf{n} = (n+1, n).$$

Now in the light of (48) we deduce that

$$\|X_{\mathbf{n}}\|_1 \le \|T_n\left(h^2\vartheta_S - 4 + 2\cos(\theta)\right)\|_1 + \|T_n\left(h^2(1-\vartheta_S) + 1\right)\|_1.$$

Now, using the fact that $2\pi\|T_n(g)\|_1 \le n\int_{[-\pi,\pi]}|g(\theta)|\,d\theta$ (see [18]), we obtain

$$\|X_{\mathbf{n}}\|_1 \le n2\pi\int_{[-\pi,\pi]}|h^2\vartheta_S - 4 + 2\cos(\theta)|\,d\theta + n(h^2(1-\vartheta_S)+1)$$

and, as in the $1\mathbb{D}$ setting, if we divide by the size of $X_{\mathbf{n}}$ i.e. $n(n+1)$ we find

$$\lim_{\mathbf{n}\to\infty}\frac{\|X_{\mathbf{n}}\|_1}{n(n+1)} = 0.$$

Consequently, again by item **GLT2**, we deduce that both the GLT matrix sequences $\{h^2A_h\}$, $\{T_{\mathbf{n}}(4-2\cos(\theta_1)-2\cos(\theta_2))\}$ share the same eigenvalue distribution function $4-2\cos(\theta_1)-2\cos(\theta_2)$, which is the GLT symbol, and hence (50) is proven. •

The previous result shows a spectral distribution as nonnegative functions both in $1\mathbb{D}$ and $2\mathbb{D}$. More precisely, looking at the range of the spectral symbols, we deduce that $[0,4]$ is a cluster for the eigenvalues of $\{h^2A_h\}$ in $1\mathbb{D}$, while $[0,8]$ is a cluster for the eigenvalues of $\{h^2A_h\}$ in $2\mathbb{D}$.

This is nontrivial (and somehow unexpected), given the fact that the related corrections are non-Hermitian and possess only strictly negative eigenvalues and zero eigenvalues.

4 Conclusions

We have provided spectral and norm estimates for matrix sequences arising from the approximation of the Laplacian via the Coco–Russo method and we have validated them with a few numerical experiments. The analysis has involved several tools from matrix theory and in particular from the setting of Toeplitz operators and Generalized Locally Toeplitz matrix sequences. Open problems remain involving variable coefficients and non square domains: both cases can be handled form a spectral view point using the GLT machinery. In particular when considering variable coefficients, the use of the diagonal sampling matrix-sequences allows to remain in GLT $*$-algebra, while the case of non square domains can be treated using the reduced GLT theory (see page 398-399 in [16] and Subsection 3.1.4 in [17]).

More involved is the case of the norm estimates of the inverse even in the case of a square in $2\mathbb{D}$. Below we present an idea in this direction.

Actually the decomposition (48) suggests, as in the $1\mathbb{D}$ setting, the use of the Sherman–Morrison–Woodbury formula: we can set $A = T_{\mathbf{n}}(f)$, $X_{\mathbf{n}} = UCV$, $\mathbf{n} = (n+1, n)$, so that

$$\begin{aligned}
U &= \left[\begin{array}{c}\mathbb{I}_n\\ \hline \mathbf{0}_{n^2\times n}\end{array}\right] \in \mathbb{R}^{n(n+1)\times n}\\
C &= \mathbb{I}_n \in \mathbb{R}^{n\times n}\\
V &= \left[T_n\left(h^2\vartheta_S - 4 + 2\cos(\theta)\right)|T_n\left(h^2(1-\vartheta_S)+1\right)|\mathbf{0}_{n\times n(n-1)}\right] \in \mathbb{R}^{n\times n(n+1)}\\
&= [V_1|V_2|\mathbf{0}_{n\times n(n-1)}].
\end{aligned}$$

Hence

$$(A+UCV)^{-1} = A^{-1} - A^{-1}U(C^{-1}+VA^{-1}U)^{-1}VA^{-1}$$

and thus $A_h^{-1} = h^2(A+UCV)^{-1} = h^2\left(A^{-1} - A^{-1}U(C^{-1}+VA^{-1}U)^{-1}VA^{-1}\right)$, with $C^{-1} = C = \mathbb{I}_n$

The previous reasoning can be useful and promising, since the entries of the inverse of $A = T_{\mathbf{n}}(f)$, $f(\theta_1,\theta_2) = 4 - 2\cos(\theta_1) - 2\cos(\theta_2)$, are explicitly known (see [13]). However technical difficulties remain due to the complicate expression of the entries of $T_{\mathbf{n}}^{-1}(f)$: this task will be the subject of future investigations.

Acknowledgments

Giovanni Russo and Stefano Serra-Capizzano are grateful to GNCS-INdAM for the support in the present research. Giovanni Russo acknowledges support from the Italian Ministry of Instruction, University and Research (MIUR), PRIN Project 2017 (No. 2017KKJP4X entitled Innovative numerical methods for evolutionary partial differential equations and applications).

References

[1] G. Barbarino and S. Serra-Capizzano. Non-Hermitian perturbations of Hermitian matrix-sequences and applications to the spectral analysis of the numerical approximation of partial differential equations. Numerical Linear Algebra with Applications, 27(3):e2286, 2020.

[2] R. H.-F. Chan and X.-Q. Jin. An introduction to iterative Toeplitz solvers. SIAM, 2007.

[3] A. Chertock, A. Coco, A. Kurganov, and G. Russo. A second-order finite-difference method for compressible fluids in domains with moving boundaries. Communication in Computational Physics, 23:230–263, 2018.

[4] A. Coco, G. Currenti, C. Del Negro, and G. Russo. A second order finite-difference ghost-point method for elasticity problems on unbounded domains with applications to volcanology. Communications in Computational Physics, 16(4):983–1009, 2014.

[5] A. Coco and G. Russo. Finite-difference ghost-point multigrid methods on Cartesian grids for elliptic problems in arbitrary domains. Journal of Computational Physics, 241:464–501, 2013.

[6] R. P. Fedkiw, T. Aslam, B. Merriman, and S. Osher. A non-oscillatory Eulerian approach to interfaces in multimaterial flows (the ghost fluid method). Journal of computational physics, 152(2):457–492, 1999.

[7] C. Garoni and S. Serra-Capizzano. Generalized locally Toeplitz sequences: Theory and applications, volume 1. Springer, 2017.

[8] C. Garoni and S. Serra-Capizzano. Generalized locally Toeplitz sequences: Theory and applications, volume 2. Springer, 2018.

[9] F. Gibou and R. Fedkiw. A fourth order accurate discretization for the Laplace and heat equations on arbitrary domains, with applications to the Stefan problem. Journal of Computational Physics, 202(2):577–601, 2005.

[10] F. Gibou, R. P. Fedkiw, L.-T. Cheng, and M. Kang. A second-order-accurate symmetric discretization of the Poisson equation on irregular domains. Journal of Computational Physics, 176(1):205–227, 2002.

[11] L. Golinskii and S. Serra-Capizzano. The asymptotic properties of the spectrum of nonsymmetrically perturbed Jacobi matrix sequences. Journal of Approximation Theory, 144(1):84–102, 2007.

[12] R. J. LeVeque and Z. Li. The immersed interface method for elliptic equations with discontinuous coefficients and singular sources. SIAM Journal on Numerical Analysis, 31(4):1019–1044, 1994.

[13] G. Meurant. A review on the inverse of symmetric tridiagonal and block tridiagonal matrices. SIAM Journal on Matrix Analysis and Applications, 13(3):707–728, 1992.

[14] M. K. Ng. Iterative methods for Toeplitz systems. Numerical Mathematics and Scie, 2004.

[15] C. S. Peskin. Numerical analysis of blood flow in the heart. Journal of computational physics, 25(3):220–252, 1977.

[16] S. Serra-Capizzano. Generalized locally Toeplitz sequences: spectral analysis and applications to discretized partial differential equations. Linear Algebra and its Applications, 366:371–402, 2003.

[17] S. Serra-Capizzano. The GLT class as a generalized Fourier analysis and applications. Linear Algebra and its Applications, 419(1):180–233, 2006.

[18] S. Serra-Capizzano and P. Tilli. On unitarily invariant norms of matrix-valued linear positive operators. Journal of Inequalities and Applications, 2002(3):176593, 2002.

[19] E. Tyrtyshnikov and N. Zamarashkin. Spectra of multilevel Toeplitz matrices: advanced theory via simple matrix relationships. Linear algebra and its applications, 270(1-3):15–27, 1998.