Received: 4 June 2021 | Revised: 2 December 2021 | Accepted: 19 January 2022

DOI: 10.1002/nla.2433

RESEARCH ARTICLE

WILEY

Spectral analysis of the finite element matrices approximating 2D linearly elastic structures and multigrid proposals

Quoc Khanh Nguyen[1] | **Stefano Serra-Capizzano**[2,3] | **Cristina Tablino-Possio**[4] | **Eddie Wadbro**[1,5]

[1]Department of Computing Science, Umeå University, Umeå, Sweden

[2]Department of Humanities and Innovation, University of Insubria, INDAM Unit, Como, Italy

[3]Department of Information Technology, Uppsala University, Uppsala, Sweden

[4]Dipartimento di Matematica e Applicazioni, Università di Milano Bicocca, Milano, Italy

[5]Department of Mathematics and Computer Science, Karlstad University, Karlstad, Sweden

Correspondence
Quoc Khanh Nguyen, Department of Computing Science, Umeå University, SE-901 87 Umeå, Sweden.
Email: qukh0001@student.umu.se

Funding information
The Italian Institution for High Mathematics (INdAM); The Swedish strategic research programme eSSENCE

Abstract

Topology optimization aims to find the best material layout subject to given constraints. The so-called material distribution methods cast the governing equation as an extended or fictitious domain problem, in which a coefficient field represents the design. When solving the governing equation using the finite element method, a large number of elements are used to discretize the design domain, and an element-wise constant function approximates the coefficient field in the considered design domain. This article presents a spectral analysis of the (large) coefficient matrices associated with the linear systems stemming from the finite element discretization of a linearly elastic problem for an arbitrary coefficient field. Based on the spectral information, we design a multigrid method which turns out to be optimal, in the sense that the (arithmetic) cost for solving the related linear systems, up to a fixed desired accuracy, is proportional to the matrix-vector cost, which is linear in the corresponding matrix size. The method is tested, and the numerical results are very satisfactory in terms of linear cost and number of iterations, which is bounded by a constant independent of the matrix size.

KEYWORDS

finite element approximations, matrix sequences, spectral analysis

1 | INTRODUCTION

Topology optimization[1] is a computational method for finding an optimal distribution material within a given design domain $\Omega \subset \mathbb{R}^d$, satisfying some given performance measure, subject to assigned constraints in d-dimension(s). The most typical problems in topology optimization are binary design variable problems, in which we consider two material types, typically denoted solid (density $\rho_{\text{solid}} > 0$) and void (density 0). The most common method to solve topology optimization problems is the so-called density-based or material distribution approach. The material distribution topology optimization has been subject to intense research in different fields such as electromagnetic,[2–4] fluid-structure interaction,[5,6] acoustics,[7,8] additive manufactured structure,[9] and especially (non)linear elasticity.[10–12] In the latter application, the results are outstanding, when considering conceptual designs of components under idealized

conditions. For example, the corresponding technique is used for designing advanced lightweight components in the car and aeronautical industries.[13] In the material distribution setting, the decision variable in the optimization problem is a design function $\alpha : \Omega \to \{0, 1\}$, that defines the physical density ρ.

Typically, topology optimization problems are solved numerically by discretizing Ω into N elements and α by an element-wise constant function with element values $\boldsymbol{\alpha} = [\alpha_1, \alpha_2, \ldots, \alpha_N]$. After discretization, the use of a binary-valued material indicator function is not feasible for many reasons, one is that the problem becomes computationally intractable. A standard solution for this issue is to allow the element values α_i to take values in the range $[0, 1]$ and to define the physical density ρ by using penalization and filtering. In particular, the physical density ρ is defined as $\rho(x) = \underline{\rho} + (\rho_{\text{solid}} - \underline{\rho})g(\mathcal{F}(\alpha)(x))$, where $\mathcal{F}$ is a filter operator, g is a penalty function, and $\underline{\rho} \geq 0$ is a constant. The relaxation of the material indicator function enables the use of gradient-based optimization methods, which, together with the efficient adjoint-based computation of design sensitivities, empowers the efficient solution of large-scale problems. The standard approach to attain a unique solution for the state problem is to set $\underline{\rho}$ to a small but strictly positive value. Berggren and Kasolis[14] studied a linearly elastic boundary value problem (BVP) defined on Ω and proved that the approximation error is bounded by the sum of three terms: a standard finite element (FE) approximation error term and two additional terms, both tending to zero as $\underline{\rho} \to 0^+$. However, there are some drawbacks when setting $\underline{\rho}$ to a small positive value, which have been indicated and discussed by Buhl et al.[15] and Bruns and Tortorelli.[16] To avoid those issues, there are few studies for vanishing lower bound. Bruns[17] let $\underline{\rho} = 0$ and treated the problem by using spectral decomposition and singular value decomposition (SD/SVD), to construct a generalized inverse (pseudoinverse) stiffness matrix $\boldsymbol{K}_N$ for solving the FE system of equations. Nguyen et al.[18] have introduced a preconditioning approach that allows the element densities to take values from a continuum and use a specific ad hoc preconditioner when solving a standard test problem. However, this approach has been only proved analytically in one spatial dimension ($d = 1$), by using a suitable spectral analysis. In fact, there is a very limited number of studies that use spectral analysis for treating the considered classes of stiffness matrices $\{\boldsymbol{K}_N\}_N$, arising from FE approximation in topology optimization. Nguyen et al.[18] have used a new spectral analysis method that is the so-called theory of Generalized Locally Toeplitz (GLT) sequences to compute and analyze the asymptotic spectral distribution of the stiffness matrices $\{\boldsymbol{K}_N\}_N$ for one spatial dimension problems.

The GLT sequences theory[19,20] is applied here for computing/analyzing the spectral distribution of matrix sequences arising from, for example, the numerical discretization such as the FE approximation of partial differential equations (PDEs) with proper boundary conditions. In the considered matrix sequences, the size of the given linear systems d_N increases with N, and it tends to infinity as $N \to \infty$. Hence, what needs to be considered is not just a single linear system, but an entire sequence of linear systems with increasing size. Under suitable conditions, the sequence of discretization matrices, such as $\{\boldsymbol{K}_N\}_N$, has an asymptotic spectral distribution. More precisely, for a large set of test functions F (usually, for all continuous functions F with bounded support), it often happens that the following limit relation holds:

$$\lim_{N\to\infty} \frac{1}{d_N} \sum_{j=1}^{d_N} F(\lambda_j(\boldsymbol{K}_N)) = \frac{1}{\mu_t(D)} \int_D F(f(x))\, \mathrm{d}x, \tag{1}$$

where $\lambda_j(\boldsymbol{K}_N)$, $j = 1, \ldots, d_N$ are the eigenvalues of $\boldsymbol{K}_N$, $\mu_t(D) \in (0, \infty)$ is the measure (t-dimensional volume) of D, and $f : D \to \mathbb{C}$ is the spectral symbol of the sequence $\{\boldsymbol{K}_N\}_N$. The spectral symbol f contains spectral information briefly described informally as follows: assuming that N is large enough, the eigenvalues of $\boldsymbol{K}_N$, except possibly for a small number of outliers, are approximately equal to the samples of f over a uniform grid in D. It is then clear that the symbol f provides a "compact" and quite accurate description of the spectrum of the matrices $\boldsymbol{K}_N$ (for N large enough), where, of course, D is related to the design domain Ω and t is related to the dimensionality d of Ω.

Here, we extend the results of our previous work[18] to the two-dimensional case ($d = 2$). A key component is the theory of multilevel block GLT sequences,[21,22] which provides the tools for computing the spectral distribution of block-structured matrices arising from the FE approximation in two-dimensional topology optimization, where the symbol f appearing in relation (1) is matrix-valued instead of scalar-valued, according to the concepts provided in Definition 1. By using this theory and more results,[23,24] we perform a detailed spectral analysis of the linear systems associated with the FE discretization of the governing equation. Moreover, the information obtained from the spectral symbol f is exploited to design a fast multigrid solver. More precisely, the proposed multigrid technique turns out to be optimal, in the sense that the (arithmetic) cost for solving the related linear systems up to a fixed desired accuracy is proportional to the computational cost of the matrix-vector products, which is linear with respect to the corresponding matrix size. The method is tested and the numerical results are promising, in terms of linear cost and

number of iterations, which is bounded by a constant independent of the matrix size and only mildly depending on the desired accuracy.

It is well-known that the spectral condition numbers of FE matrices grows as $O(h^{-2})$, where h is the element size. This type of result is quite old (see Reference 25 and references therein): for instance, in some settings, it is enough to work with elementary tools such as the Gerschgorin theorems, at least for proving the positive definiteness. However, in our setting, the matrices have an expression, which is less standard, due to the tensor structure of the operator $\epsilon(u)$, and, in fact, the involved matrices do not possess the form described by Dorostkar et al.[26] In reality, if one applies the Gerschgorin theorems, the conclusions are less general and parameter (ν) dependent, as it is clear looking at the intricate expression of the matrix in Equation (6); refer also to (7). Hence we need more sophisticated mathematical instruments along the lines followed in Reference 26. Indeed, we observe that our results and those in Reference 26 are similar in the sense that the same block multilevel GLT machinery is employed, but different just because the matrices have different mathematical expressions and the derivation of the symbol has to be performed from scratch. We finally stress that determining the symbol is important for two additional reasons, which go beyond the conditioning analysis:

- the position and the order of the zeros, not only for understanding the behavior of the spectral condition numbers but especially for designing the projection/restriction operators, with the target of obtaining optimal solvers, as done in Section 5;
- the global distribution results for the eigenvalues, which of course are not classical and cannot be recovered using classical tools and, moreover, they have an impact in the modern analysis of the convergence speed of (preconditioned) Krylov methods; see References 27,28.

The article is organized as follows. In Section 2, we describe the continuous problem and the resulting coefficient matrices arising from our FE approximation. Section 3 is devoted to the spectral analysis of the FE matrices in the two-dimensional setting, from the perspective of the GLT theory. In Section 4, we give a brief account of multigrid methods, with special attention to the block case encountered in the present context. Section 5 contains a multigrid proposal based on the spectral information given in Section 3 and on the conditions reported in Section 4. Finally, conclusions are reported in Section 6. In addition, Appendix A is devoted to some relevant model information.

2 | PRELIMINARIES AND DISCRETIZATION

In this section, we report the description of the continuous problem (Section 2.1), its approximation by a basic FE procedure (Section 2.2), and finally, the formal expression of the resulting FE matrices (Section 2.3). We emphasize that the formal expression of the relevant matrices is a key ingredient for applying the GLT theory, in order to produce a global spectral description of the matrix sequences under consideration.

2.1 | A problem in linear elastostatics

A material that deforms under a load and resumes its undeformed shape, when removing the load, is called elastic. Provided that the load-induced deformations are small, many solids are linearly elastic, which means that the relation between the deformation and the applied load is linear. In this article, we consider a linearly elastic structure that (when unloaded) occupies the hyper-rectangular domain $\Omega \subset \mathbb{R}^d$. Henceforth, let $b \in L^2(\Omega)^d$ be a given body load (a volume force) in Ω, $t \in L^2(\Gamma_F)^d$ be the surface traction acting on the non-clamped boundary $\Gamma_F \subset \partial\Omega$ of the solid, and u denote the resulting equilibrium displacement.

In linear elasticity, deformations are characterized by the so-called strain tensor

$$\epsilon(u) = \frac{1}{2}(\nabla u + \nabla u^T).$$

We remark that the skew-symmetric part of the displacement gradient is related to rigid rotations (and not to the deformations) of the body. Hooke's generalized law states the relationship between the strain tensor and the so-called stress tensor σ that is

$$\sigma = E\epsilon(u), \tag{2}$$

with E being the fourth-order plane stress elasticity tensor having in total d^4 components. In this article, we limit our attention to the analysis of FE matrices arising in the two-dimensional plane stress setting, which is the standard in material distribution topology optimization problems.[1] By invoking symmetries, the number of independent components in E can be reduced[29(p. 194–197)]. The equilibrium assumption, which states that the forces acting on any sub-body must be in balance, together with Cauchy's theorem, which states that the surface force density depends linearly on the normal derivative n, and the divergence theorem implies

$$-\nabla \cdot \sigma = b. \tag{3}$$

In this article, we assume that the structure is clamped along the boundary portion $\Gamma_D \subset \partial\Omega$ and subject to a surface traction load t on $\Gamma_F = \partial\Omega \setminus \Gamma_D$. The boundary conditions above together with Equations (2) and (3) yield the BVP

$$-\nabla \cdot (E\epsilon(u)) = b \ \text{ in } \Omega, \tag{4a}$$

$$u = 0 \ \text{ on } \Gamma_D, \tag{4b}$$

$$(E\epsilon(u))n = t \ \text{ on } \Gamma_F. \tag{4c}$$

Here, the application in focus is material distribution based topology optimization. More precisely, we assume that the elasticity tensor is ρE_c, where E_c is a constant fourth-order elasticity tensor and $\rho \in \mathcal{A} \subset L^\infty(\Omega)$. A typical choice is to let $\mathcal{A} = \{\rho \in L^\infty(\Omega) \mid \underline{\rho} \le \rho \le 1 \ \text{ a.e. in } \ \Omega\}$, where $\underline{\rho}$ is a constant that satisfies $0 < \underline{\rho} \ll 1$.

In the context of material distribution topology optimization, the finite element method (FEM) is the standard choice for generating numerical solutions of the BVP (4). This solution process uses the variational form of the BVP. Since the structure is clamped along the boundary portion $\Gamma_D \subset \partial\Omega$, the set of all kinematically admissible displacements of the structure is

$$\mathcal{U} = \left\{ u \in H^1(\Omega)^d \mid u|_{\Gamma_D} \equiv 0 \right\}.$$

Under the above assumptions, the steady-state displacement u of the structure is the solution to

$$\text{Find } u \in \mathcal{U} \text{ such that } a(\rho; u, v) = \ell(v) \ \forall v \in \mathcal{U}. \tag{5}$$

Here, the energy bilinear form a and the load linear form ℓ are defined as

$$a(\rho; u, v) = \int_\Omega \rho \, (E_c\epsilon(u)) : \epsilon(v),$$

$$\ell(v) = \int_{\Gamma_F} t \cdot v + \int_\Omega b \cdot v,$$

where the colon ":" denotes the full contraction between the two tensors. When using the standard basis, the full contraction of the two matrices is their Frobenius scalar product.

2.2 | Discretization by $\mathbb{Q}_1$ finite elements

Let $\mathcal{S}_h$ be a structured grid consisting of tensor product elements. More precisely, we define $\mathcal{U}_h$ to be the space of all continuous functions that are zero on the boundary Γ_D and 2-vectors with each component being linear in each coordinate direction on each element in $\mathcal{S}_h$. More precisely, we set

$$\mathcal{U}_h = \{u \in \mathcal{U} \mid u|_S \in P_1(S)^d, \ \forall S \in \mathcal{S}_h\},$$

where $P_1(S)$ is the space of all functions that are linear in each spatial component on element S. We define $N = n_1 n_2$ to be the number of elements in $\mathcal{S}_h$ and $M = m_1 m_2$ to be the number of basis functions in $\mathcal{U}_h$ where subscripts 1 and 2 denote the number of elements and the number of basis functions in x_1 and x_2 direction, respectively.

In order to obtain a discrete equation system corresponding to the variational problem (5), we approximate functions $u \in \mathcal{U}$ and $v \in \mathcal{U}$ by functions $u_h \in \mathcal{U}_h$ and $v_h \in \mathcal{U}_h$, and approximate the physical density ρ by a function ρ_h that is constant on each element in $\mathcal{S}_h$. We let $\boldsymbol{u} \in \mathbb{R}^M$ and $\boldsymbol{v} \in \mathbb{R}^M$ be the coefficient vectors of u_h and v_h, respectively. We define $\boldsymbol{\rho} \in \mathbb{R}^N$ to be the vector whose entries coincide with the values of ρ_h in each element. In other words, we define $\mathcal{A}_h \subset \mathbb{R}^N$ to be the set of vectors $\boldsymbol{\rho}$ that correspond to element-wise constant functions ρ_h that are in $\mathcal{A}$. By applying the above approximations, we deduce that the variational problem (5) reduces to the linear system

$$\boldsymbol{K}(\boldsymbol{\rho})\boldsymbol{u} = \boldsymbol{f},$$

where the stiffness matrix $\boldsymbol{K}(\boldsymbol{\rho})$ and the right hand vector $\boldsymbol{f}$ have entries

$$K_{ij} = \int_\Omega \rho_h (E_c \epsilon(\varphi_i)) : \epsilon(\varphi_j) \qquad \text{and} \qquad f_j = \int_{\Gamma_F} t \cdot \varphi_j + \int_\Omega b \cdot \varphi_j,$$

respectively.

A standard procedure for assembling the stiffness matrix $\boldsymbol{K}$ is to loop over each element so that

$$\boldsymbol{K} = \sum_{n=1}^{N} \rho_n \boldsymbol{K}_n,$$

where $\boldsymbol{K}_n$ is the elementary stiffness matrix for the element S_n. In the studied setup, we remark that all elements have the same shape and size, and hence all elements in the stiffness matrices possess the same non-zero values. The positions holding these values in the element matrices correspond to the indices of the basis function having support in the element.

2.3 | Explicit expressions for the element stiffness matrix

In this article, we limit our attention to two spatial dimensions. Hence, if not otherwise stated, by default $d = 2$. In the subsequent lines, we give an explicit expression of the element stiffness matrix $\boldsymbol{K}_n$ with respect to the reference element illustrated in Figure 1A, which will be assembled into stiffness matrix $\boldsymbol{K}$ using the global node numbering illustrated in Figure 1B. For example, for the first element, the nodes with local node numbers 5, 6, 7, and 8 have the global node numbers $2m_2 + 1$, $2m_2 + 2$, $2m_2 + 3$, and $2m_2 + 4$, respectively.

Here, we consider the so-called plane stress case. In this setting, the element stiffness matrix

$$\boldsymbol{K}_n = \frac{E_0}{1-\nu^2} \begin{bmatrix} k_1 & k_2 & k_3 & k_4 & k_5 & k_6 & k_7 & k_8 \\ k_2 & k_1 & k_6 & k_5 & k_4 & k_3 & k_8 & k_7 \\ k_3 & k_6 & k_1 & k_8 & k_7 & k_2 & k_5 & k_4 \\ k_4 & k_5 & k_8 & k_1 & k_2 & k_7 & k_6 & k_3 \\ k_5 & k_4 & k_7 & k_2 & k_1 & k_8 & k_3 & k_6 \\ k_6 & k_3 & k_2 & k_7 & k_8 & k_1 & k_4 & k_5 \\ k_7 & k_8 & k_5 & k_6 & k_3 & k_4 & k_1 & k_2 \\ k_8 & k_7 & k_4 & k_3 & k_6 & k_5 & k_2 & k_1 \end{bmatrix}, \tag{6}$$

where ν is the Poisson ratio, $E_0/(1-\nu^2)$ is a constant that will be neglected without affecting spectral analysis in the following sections, and the entries are

$$\begin{aligned} &k_1 = \tfrac{1}{2}\left(1 - \tfrac{\nu}{3}\right), \quad k_3 = \tfrac{\nu}{6}, && k_5 = -\tfrac{1}{4}\left(1 + \tfrac{\nu}{3}\right), \quad k_7 = \tfrac{1}{4}\left(-1 + \tfrac{\nu}{3}\right), \\ &k_2 = \tfrac{1}{8}(1+\nu), && k_4 = \tfrac{1}{8}(1-3\nu), \quad k_6 = -\tfrac{1}{8}(1-3\nu), \quad k_8 = -\tfrac{1}{8}(1+\nu). \end{aligned} \tag{7}$$

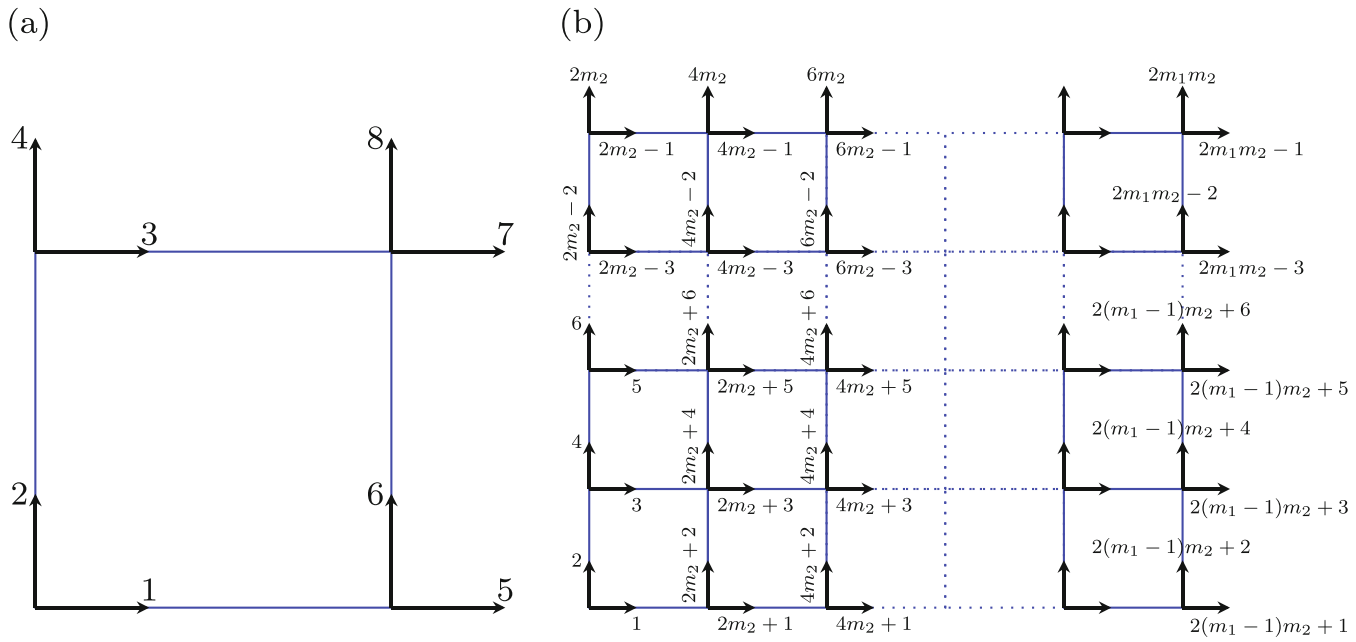

FIGURE 1 Local node numbering on the reference element (left) and the global node numbering for the full discretization (right)

Appendix A provides more information regarding the stress–strain relation and the assumptions used to arrive at the element stiffness matrix above. Moreover, Appendix A includes a motivation why the Poisson's ratio must be in the range $\nu \in (-1, 1)$ in the two-dimensional setting. This is in contrast to the typical bound $\nu \in (-1, 0.5)$, which holds for the three-dimensional setting. From an application viewpoint, we are mainly interested in the case where the material is isotropic. Therefore, we primarily consider Poisson's ratios in the range $\nu \in [0, 0.5)$.

3 | SPECTRAL ANALYSIS

The current section is devoted to the spectral analysis of the FE coefficient matrices derived in the previous section and is complemented by numerical tests, that confirm the theoretical analysis. In particular, Section 3.1 contains the necessary preliminary concepts and tools, while Sections 3.2 and 3.3 focus on the specific study in 2D, where Section 3.2 covers the constant coefficient case and Section 3.3 covers the variable coefficient case.

3.1 | Premises

The premises include the formal definition of eigenvalue and singular value distribution, the notion of multi-indexing, the concepts of multilevel block Toeplitz matrices, multilevel block sampling matrices, and multilevel block GLT matrix sequences.

3.1.1 | Singular value/eigenvalue distributions

We first give the formal definitions, and then we briefly discuss the informal and practical meaning.

Definition 1. Let $\{A_n\}_n$ be a sequence of matrices, with A_n of size d_n, and let $f : D \subset \mathbb{R}^t \to \mathbb{C}^{r\times r}$ 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

$$\lim_{n\to\infty} \frac{1}{d_n} \sum_{i=1}^{d_n} F(\sigma_i(A_n)) = \frac{1}{\mu_t(D)} \int_D \frac{\sum_{i=1}^{r} F(\sigma_i(f(\mathbf{x})))}{r} \, d\mathbf{x}, \quad \forall\, F \in C_c(\mathbb{R}). \tag{8}$$

- 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

$$\lim_{n\to\infty} \frac{1}{d_n}\sum_{i=1}^{d_n} F(\lambda_i(A_n)) = \frac{1}{\mu_t(D)}\int_D \frac{\sum_{i=1}^{r} F(\lambda_i(f(\mathbf{x})))}{r}\,\mathrm{d}\mathbf{x}, \quad \forall\, F \in C_c(\mathbb{C}). \tag{9}$$

If $\{A_n\}_n$ has both a singular value and an eigenvalue distribution described by f , then we write $\{A_n\}_n \sim_{\sigma,\lambda} f$. Notice that (9) is a generalization of (1). More specifically (1) reduces to (9) when the size r of the matrix-valued symbol f is equal to 1. (In practice in the Toeplitz setting and often in the GLT setting, the parameter r can be read at a matrix level as the size of the elementary blocks which form the global matrix A_n. This also holds for the problem studied in this article, as will be further discussed in terms of the stiffness matrices in Sections 3.2 and 3.3 as well as in the block Toeplitz/block diagonal sampling structures in Section 3.1.2.)

As already mentioned in the introduction, the symbol f contains spectral/singular value information briefly described informally as follows. With reference to (9), assuming that N is large enough, the eigenvalues of $\boldsymbol{K}_N$ are partitioned into r subsets of the same cardinality, except possibly for a small number of outliers, and the ith subset is approximately formed by the samples of $\lambda_i(f)$ over a uniform grid in D, $i = 1, \ldots, r$. It is then clear that the symbol f provides a "compact" and a quite accurate description of the spectrum of the matrices $\boldsymbol{K}_N$ for N large enough. Relation (8) has the same meaning when referring to the singular values of $\boldsymbol{K}_N$ and by replacing $\lambda_i(f)$ with $\sigma_i(f)$, $i = 1, \ldots, r$.

3.1.2 | Multilevel block Toeplitz matrices, multilevel block diagonal sampling matrices, and multilevel block GLT sequences

We start by introducing the multi-index notation, which is quite useful for treating sequences of matrices arising from the discretization of PDEs. A multi-index $\boldsymbol{i} \in \mathbb{Z}^d$, also called a d-index, is simply a (row) vector in $\mathbb{Z}^d$; its components are denoted by $i_1, \ldots, i_d$.

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

$$\Big[\ \ldots\ \Big[\ \big[\ (j_1, \ldots, j_d)\ \big]_{j_d=h_d,\ldots,k_d}\ \Big]_{j_{d-1}=h_{d-1},\ldots,k_{d-1}}\ \ldots\ \Big]_{j_1=h_1,\ldots,k_1}. \tag{10}$$

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

(Multilevel) Block Toeplitz matrices.

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

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

with $\mathbf{e}$ vector of all ones, with blocks $A_{\mathbf{k}} \in \mathbb{C}^{r\times r}$, $\mathbf{k} = -(\mathbf{n}-\mathbf{e}), \ldots, \mathbf{n}-\mathbf{e}$, is called a multilevel block Toeplitz matrix, or, more precisely, a d-level r-block 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

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

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

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

or, equivalently, as

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

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 $\{T_{\mathbf{n}}(f)\}_{\mathbf{n}\in\mathbb{N}^d}$ the family of (multilevel block) Toeplitz matrices associated with f, which, in turn, is called the generating function of $\{T_{\mathbf{n}}(f)\}_{\mathbf{n}\in\mathbb{N}^d}$.

(Multilevel) block diagonal sampling matrices. For $n \in \mathbb{N}$ and $a : [0,1] \to \mathbb{C}^{r\times r}$, we define the block diagonal sampling matrix $D_n(a)$ as the diagonal matrix

$$D_n(a) = \underset{i=1,\ldots,n}{diag}\ a\left(\frac{i}{n}\right) = \begin{bmatrix} a(\frac{1}{n}) & & & \\ & a(\frac{2}{n}) & & \\ & & \ddots & \\ & & & a(1) \end{bmatrix} \in \mathbb{C}^{rn\times rn}.$$

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

$$D_{\boldsymbol{n}}(a) = \underset{\boldsymbol{i}=\mathbf{1},\ldots,\boldsymbol{n}}{diag}\ a\left(\frac{\boldsymbol{i}}{\boldsymbol{n}}\right) \in \mathbb{C}^{rN(\boldsymbol{n})\times rN(\boldsymbol{n})},$$

with the lexicographical ordering (10) as discussed at the beginning of Section 3.1.2.

Zero-distributed sequences. According to Definition 1, a sequence of matrices $\{Z_n\}_n$ such that

$$\{Z_n\}_n \sim_\sigma 0,$$

is referred to as a zero-distributed sequence and we emphasize that this notion applies in the sense of the singular values. Note that, for any $r \geq 1$, $\{Z_n\}_n \sim_\sigma 0$ is equivalent to $\{Z_n\}_n \sim_\sigma O_r$ (notice that O_m and I_m denote the $m\times m$ zero matrix and the $m\times m$ identity matrix, respectively). Proposition 1 provides an important characterization of zero-distributed sequences together with a useful sufficient condition for detecting such sequences. Throughout this article, we use the natural convention $1/\infty = 0$.

Proposition 1. *Let $\{Z_n\}_n$ be a sequence of matrices, with Z_n of size d_n, and let $\|\cdot\|$ be the standard spectral matrix norm (the one induced by the Euclidean vector norm).*

- *$\{Z_n\}_n$ is zero-distributed if and only if $Z_n = R_n + N_n$ with $\mathrm{rank}(R_n)/d_n \to 0$ and $\|N_n\| \to 0$ as $n \to \infty$.*
- *$\{Z_n\}_n$ is zero-distributed if there exists a $p \in [1,\infty]$ such that $\|Z_n\|_p/(d_n)^{1/p} \to 0$ as $n \to \infty$.*

(Multilevel) Block GLT matrix sequences.

Now we give a very concise and operational description of the multilevel block GLT sequences, from which it will be clear that the multilevel block Toeplitz structures, the zero-distributed matrix sequences, and the multilevel block diagonal sampling matrices represent the basic building components.

Let $d, r \geq 1$ be fixed positive integers. A multilevel r-block GLT sequence (or simply a GLT sequence if d, r can be inferred from the context or we do not need/want to specify them) is a special r-block matrix sequence $\{A_n\}_n$ equipped with a measurable function $\kappa : [0,1]^d \times [-\pi,\pi]^d \to \mathbb{C}^{r\times r}$, the so-called symbol. We use the notation $\{A_n\}_n \sim_{\mathrm{GLT}} \kappa$ to indicate that $\{A_n\}_n$ is a GLT sequence with symbol κ. The symbol of a GLT sequence is unique in the sense that if $\{A_n\}_n \sim_{\mathrm{GLT}} \kappa$ and $\{A_n\}_n \sim_{\mathrm{GLT}} \varsigma$ then $\kappa = \varsigma$ a.e. in $[0,1]^d \times [-\pi,\pi]^d$. The main properties of r-block GLT sequences proved in

Reference 21 are listed below. If A is a matrix, we denote by $A^\dagger$ the Moore–Penrose pseudoinverse of A (recall that $A^\dagger = A^{-1}$ whenever A is invertible).

GLT 1 If $\{A_n\}_n \sim_{\text{GLT}} \kappa$, then $\{A_n\}_n \sim_\sigma \kappa$. If moreover each A_n is Hermitian, then $\{A_n\}_n \sim_\lambda \kappa$.
GLT 2 We have:
(a) $\{T_n(\phi)\}_n \sim_{\text{GLT}} \kappa(\mathbf{x},\boldsymbol{\theta}) = \phi(\boldsymbol{\theta})$ if $\phi : [-\pi,\pi]^d \to \mathbb{C}^{r\times r}$ is in $L^1([-\pi,\pi]^d)$;
(b) $\{D_n(a)\}_n \sim_{\text{GLT}} \kappa(\mathbf{x},\boldsymbol{\theta}) = a(\mathbf{x})$ if $a : [0,1]^d \to \mathbb{C}^{r\times r}$ is Riemann-integrable;
(c) $\{Z_n\}_n \sim_{\text{GLT}} \kappa(\mathbf{x},\boldsymbol{\theta}) = O_r$ if and only if $\{Z_n\}_n \sim_\sigma 0$ (zero-distributed sequences coincide exactly with the GLT sequences having GLT symbol equal to O_r a.e. and hence equal to O_s a.e. for any positive integer s).
GLT 3 If $\{A_n\}_n \sim_{\text{GLT}} \kappa$ and $\{B_n\}_n \sim_{\text{GLT}} \varsigma$, then
(a) $\{A_n^*\}_n \sim_{\text{GLT}} \kappa^*$;
(b) $\{\alpha A_n + \beta B_n\}_n \sim_{\text{GLT}} \alpha\kappa + \beta\varsigma$ for all $\alpha,\beta \in \mathbb{C}$;
(c) $\{A_nB_n\}_n \sim_{\text{GLT}} \kappa\varsigma$;
(d) $\{A_n^\dagger\}_n \sim_{\text{GLT}} \kappa^{-1}$ provided that κ is invertible a.e.

3.2 | The constant coefficient case ($\rho \equiv 1$)

Hereafter we are looking for identifying the symbol underlying the constant coefficient case $\rho \equiv 1$, according to the standard notion of generating function in the Toeplitz theory (see Section 3.1.2 and Reference 22 for more details) and according to the notion of symbol reported in Definition 1.

3.2.1 | Symbol definition

Let $A_n(1,\mathtt{DN}^3)$ be the stiffness matrix obtained with a $\mathbb{Q}_1$ FEs approximation as described in Section 2.2 with proper boundary conditions, Dirichlet in one side "D" and Neumann "N" in the other three (and hence the formal notation $A_n(1,\mathtt{DN}^3)$), where we have chosen a uniform meshing with n intervals in the x_1 direction and n intervals in the x_2 direction. According to the previously considered ordering of the nodes, the matrix $A_n(1,\mathtt{DN}^3)$ is a two-level block tridiagonal structure of size n with tridiagonal blocks of size $n+1$, whose elements are small matrices of size 2. We notice that the size is dictated by all the mesh points (including those lying in the boundaries) when considering Neumann boundary conditions, and by all the internal mesh points (excluding those lying in the boundaries), when considering Dirichlet boundary conditions, so that the precise structure is the following:

$$A_n(1,\mathtt{DN}^3) = \begin{bmatrix} \hat{A}_0 & \hat{A}_{-1} & & & \\ \hat{A}_{-1}^T & \hat{A}_0 & \hat{A}_{-1} & & \\ & \ddots & \ddots & \ddots & \\ & & \hat{A}_{-1}^T & \hat{A}_0 & \hat{A}_{-1} \\ & & & \hat{A}_{-1}^T & \widetilde{A_0} \end{bmatrix},$$

with $\hat{A}_1 = \hat{A}_{-1}^T$ and $\widetilde{A_0}$ slightly differing from $\hat{A}_0$ due to the Neumann boundary condition on the right vertical border, and where

$$\hat{A}_0 = \begin{bmatrix} \tilde{a}_{0,0} & a_{0,-1} & & & \\ a_{0,-1} & a_{0,0} & a_{0,-1} & & \\ & \ddots & \ddots & \ddots & \\ & & a_{0,-1} & a_{0,0} & a_{0,-1} \\ & & & a_{0,-1} & \tilde{a}_{0,0} \end{bmatrix}, \qquad \hat{A}_{-1} = \begin{bmatrix} \tilde{a}_{-1,0} & a_{-1,-1} & & & \\ a_{-1,1} & a_{-1,0} & a_{-1,-1} & & \\ & \ddots & \ddots & \ddots & \\ & & a_{-1,1} & a_{-1,0} & a_{-1,-1} \\ & & & a_{-1,1} & \tilde{a}_{-1,0} \end{bmatrix}.$$

Again with the superscript "$\backsim$" we are denoting slightly different entries due to Neumann boundary conditions. We notice that the dimension of the blocks A_i equals double the number of nodes on the vertical border, while their number equals

the number of nodes on the horizontal border minus one due to the Dirichlet boundary condition on the left vertical border. Finally, every $a_{s,t}$, $s,t \in \{-1,0,1\}$ is a 2×2 matrix, because two degrees of freedom are associated to each node of the mesh. More precisely, it holds

$$a_{0,0} = 2\begin{bmatrix} 2k_1 & k_2+k_8 \\ k_2+k_8 & 2k_1 \end{bmatrix}, \qquad a_{0,1} = a_{0,-1} = \begin{bmatrix} 2k_3 & k_4+k_6 \\ k_4+k_6 & 2k_5 \end{bmatrix}, \qquad a_{1,-1} = a_{-1,1} = \begin{bmatrix} k_7 & k_2 \\ k_2 & k_7 \end{bmatrix},$$
$$a_{1,0} = a_{-1,0} = \begin{bmatrix} 2k_5 & k_4+k_6 \\ k_4+k_6 & 2k_3 \end{bmatrix}, \qquad a_{1,1} = a_{-1,-1} = \begin{bmatrix} k_7 & k_8 \\ k_8 & k_7 \end{bmatrix}.$$

Now setting $\boldsymbol{n} = (n_1, n_2)$, $n_1 = n, n_2 = n+1$ we define

$$T_{\boldsymbol{n}}(f_{\mathbb{Q}_1}) = \begin{bmatrix} A_0 & A_{-1} & & & \\ A_{-1}^T & A_0 & A_{-1} & & \\ & \ddots & \ddots & \ddots & \\ & & A_{-1}^T & A_0 & A_{-1} \\ & & & A_{-1}^T & A_0 \end{bmatrix},$$

where $A_1 = A_{-1}^T$ and

$$A_0 = \begin{bmatrix} a_{0,0} & a_{0,-1} & & & \\ a_{0,-1} & a_{0,0} & a_{0,-1} & & \\ & \ddots & \ddots & \ddots & \\ & & a_{0,-1} & a_{0,0} & a_{0,-1} \\ & & & a_{0,-1} & a_{0,0} \end{bmatrix}, \qquad A_{-1} = \begin{bmatrix} a_{-1,0} & a_{-1,-1} & & & \\ a_{-1,1} & a_{-1,0} & a_{-1,-1} & & \\ & \ddots & \ddots & \ddots & \\ & & a_{-1,1} & a_{-1,0} & a_{-1,-1} \\ & & & a_{-1,1} & a_{-1,0} \end{bmatrix}.$$

Therefore, in accordance with the multi-index notation and with the definition of multilevel block Toeplitz matrices in Section 3.1.2, we can easily read the generating function $f_{\mathbb{Q}_1}$, just by having in mind that we are dealing with a matrix-valued one of size 2×2. More precisely, we have

$$\begin{aligned} f_{\mathbb{Q}_1}(\theta_1,\theta_2) &= f_{A_0}(\theta_1) + f_{A_{-1}}(\theta_1)e^{-\hat{\imath}\theta_2} + f_{A_1}(\theta_1)e^{\hat{\imath}\theta_2} \\ &= (a_{0,0} + a_{0,-1}e^{-\hat{\imath}\theta_1} + a_{0,-1}e^{\hat{\imath}\theta_1}) + (a_{-1,0} + a_{-1,-1}e^{-\hat{\imath}\theta_1} + a_{-1,1}e^{\hat{\imath}\theta_1})e^{-\hat{\imath}\theta_2} + (a_{1,0} + a_{1,-1}e^{-\hat{\imath}\theta_1} + a_{1,1}e^{\hat{\imath}\theta_1})e^{\hat{\imath}\theta_2} \\ &= a_{0,0} + 2a_{0,-1}\cos\theta_1 + 2a_{-1,0}\cos\theta_2 + a_{-1,-1}(e^{-\hat{\imath}\theta_1}e^{-\hat{\imath}\theta_2} + e^{\hat{\imath}\theta_1}e^{\hat{\imath}\theta_2}) + a_{-1,1}(e^{\hat{\imath}\theta_1}e^{-\hat{\imath}\theta_2} + e^{-\hat{\imath}\theta_1}e^{\hat{\imath}\theta_2}) \\ &= \begin{bmatrix} f_{11}(\theta_1,\theta_2) & f_{12}(\theta_1,\theta_2) \\ f_{21}(\theta_1,\theta_2) & f_{22}(\theta_1,\theta_2) \end{bmatrix}, \end{aligned} \tag{11}$$

where

$$\begin{aligned} f_{11}(\theta_1,\theta_2) &= 4k_1 + 4k_3\cos\theta_1 + 4k_5\cos\theta_2 + 4k_7\cos\theta_1\cos\theta_2, \\ f_{12}(\theta_1,\theta_2) &= 2(k_2+k_8) + 2(k_4+k_6)(\cos\theta_1+\cos\theta_2) + k_8(e^{-\hat{\imath}\theta_1}e^{-\hat{\imath}\theta_2} + e^{\hat{\imath}\theta_1}e^{\hat{\imath}\theta_2}) + k_2(e^{\hat{\imath}\theta_1}e^{-\hat{\imath}\theta_2} + e^{-\hat{\imath}\theta_1}e^{\hat{\imath}\theta_2}), \\ f_{22}(\theta_1,\theta_2) &= 4k_1 + 4k_5\cos\theta_1 + 4k_3\cos\theta_2 + 4k_7\cos\theta_1\cos\theta_2, \\ f_{21}(\theta_1,\theta_2) &= f_{12}(\theta_1,\theta_2). \end{aligned}$$

Finally, by recalling the expression of k_i in (7), we deduce that

$$\begin{aligned} f_{11}(\theta_1,\theta_2) &= 2\left(1-\frac{\nu}{3}\right) + 2\frac{\nu}{3}\cos\theta_1 - \left(1+\frac{\nu}{3}\right)\cos\theta_2 + \left(-1+\frac{\nu}{3}\right)\cos\theta_1\cos\theta_2, \\ f_{12}(\theta_1,\theta_2) = f_{21}(\theta_1,\theta_2) &= \frac{1}{8}(1+\nu)(-e^{-\hat{\imath}\theta_1}e^{-\hat{\imath}\theta_2} - e^{\hat{\imath}\theta_1}e^{\hat{\imath}\theta_2} + e^{\hat{\imath}\theta_1}e^{-\hat{\imath}\theta_2} + e^{-\hat{\imath}\theta_1}e^{\hat{\imath}\theta_2}), \\ f_{22}(\theta_1,\theta_2) &= 2\left(1-\frac{\nu}{3}\right) - \left(1+\frac{\nu}{3}\right)\cos\theta_1 + 2\frac{\nu}{3}\cos\theta_2 + \left(-1+\frac{\nu}{3}\right)\cos\theta_1\cos\theta_2. \end{aligned}$$

Of course, when the boundary conditions are uniformly of Dirichlet type, then $A_n(1, \mathtt{D}^4) = T_{\boldsymbol{n}}(f_{\mathbb{Q}_1})$ with $\boldsymbol{n} = (n_1, n_2)$, $n_1 = n_2 = n - 1$, since only the internal meshpoints are involved in the matrix.

By collecting all the previous results, the following proposition links in a precise way the Toeplitz structures with matrices $A_n(1, \mathtt{D}^4)$ and $A_n(1, \mathtt{DN}^3)$.

Proposition 2. *Let $f_{\mathbb{Q}_1}(\theta_1, \theta_2)$ be the symbol defined in (11). Let us consider a uniform meshing in both directions with n sub-intervals. Then we have the following relationships*

$$\begin{aligned} A_n(1, \mathtt{D}^4) &= T_{\boldsymbol{n}}(f_{\mathbb{Q}_1}), & \boldsymbol{n} &= (n_1, n_2), & n_1 &= n_2 = n - 1, \\ A_n(1, \mathtt{DN}^3) &= T_{\boldsymbol{n}}(f_{\mathbb{Q}_1}) + R_{\boldsymbol{n}}, & \boldsymbol{n} &= (n_1, n_2), & n_1 &= n, n_2 = n + 1, \end{aligned}$$

where $R_{\boldsymbol{n}}$ has rank bounded by $2(n + 1)$ due to the term $\widetilde{A_0} - A_0$ plus $4(n - 1)$ due to the terms $\hat{A}_j - A_j$, $j = -1, 0, 1$.

If the more general setting is considered with n_1 sub-intervals in the x_1 direction and n_2 sub-intervals in the x_2 direction, then analogous relations are true:

$$\begin{aligned} A_{\boldsymbol{n}}(1, \mathtt{D}^4) &= T_{\boldsymbol{n}}(f_{\mathbb{Q}_1}), & \boldsymbol{n} &= (n_1 - 1, n_2 - 1), \\ A_{\boldsymbol{n}}(1, \mathtt{DN}^3) &= T_{\boldsymbol{n}}(f_{\mathbb{Q}_1}) + R_{\boldsymbol{n}}, & \boldsymbol{n} &= (n_1, n_2 + 1), \end{aligned}$$

where $R_{\boldsymbol{n}}$ has rank bounded by $2(n_2 + 1)$ due to the term $\widetilde{A_0} - A_0$ plus $4(n_1 - 1)$ due to the terms $\hat{A}_j - A_j$, $j = -1, 0, 1$.

In the following Section 3.2.2, we show by numerical evidences that the function $f_{\mathbb{Q}_1}$ is the eigenvalue symbol, in the sense of Definition 1, of the matrix sequences $\{A_n(1, \mathtt{D}^4)\}_n$ and $\{A_n(1, \mathtt{DN}^3)\}_n$. These visualizations find a theoretical ground in Section 3.2.3, where the related statements are proven in Proposition 3.

3.2.2 | Numerics: The eigenvalue distribution

We start our analysis by performing a few numerical tests. First of all, in Figure 2 we draw the two eigenvalues surfaces of the symbol $f_{\mathbb{Q}_1}$ with a sampling in $(-\pi, \pi) \times (-\pi, \pi)$ with respect to different ν values together with corresponding contour lines. The figure clearly shows that there is a minimum at $(0, 0)$ for both surfaces, regardless of the value of ν. Moreover, the eigenvalue surfaces for different values of ν share the same general features. However, there are some noticeable trends that one might find interesting. In particular, for the second surface, both the maximum value and the minimum decrease as ν increases in $(0, 0.5)$ with the minimum decreasing faster than the maximum. Concerning the maximum values of the two surfaces, Figure 3 highlights that the maximum of the first surface is 4 independently of ν, while the maximum of the second surface is decreasing from 2 to $1.\overline{6}$ as long as ν increases in $(0, 0.5)$. The picture is similar if we consider the full interval $(-1, 0.5)$.

We have computed the minimal eigenvalue and the spectral condition number of matrices of increasing dimension with $\nu = 0.4$, both in the case of Dirichlet boundary conditions and Dirichlet+Neumann boundary conditions (see Table 1). In both cases it is evident that the ratio $\lambda_{\min}(h)/\lambda_{\min}(h/2)$ tends to 4 and the ratio $\kappa_2(h)/\kappa_2(h/2)$ to $1/4$ in accordance with the above conjecture. No significant dependency on ν is present.

Finally, we would like to stress as the match between the samplings of the symbol eigenvalues and the matrix eigenvalues is really sharp even for the moderate size of the involved matrices. To this end in Figure 4, we report the ordered union of equispaced samplings of the two surfaces $\lambda_1\left(f_{\mathbb{Q}_1}(\theta_1, \theta_2)\right)$ and $\lambda_2\left(f_{\mathbb{Q}_1}(\theta_1, \theta_2)\right)$ (red line) side by side to the ordered eigenvalues of the matrix $A_n(1, \mathtt{D}^4)$ for $N(n) = 7938$ and different values of ν. No significant dependency on ν is observed for both types of boundary conditions (see Figure 5).

3.2.3 | Symbol spectral analysis: Distribution, extremal eigenvalues, and conditioning

This section comprises three propositions that focus on spectral distribution, extremal eigenvalues, and conditioning of our matrix sequences in 2D. All the results are based on the symbol and its analytical features.

We start with a result devoted to the spectral distribution, whose numerical evidences have been discussed already in Section 3.2.2.

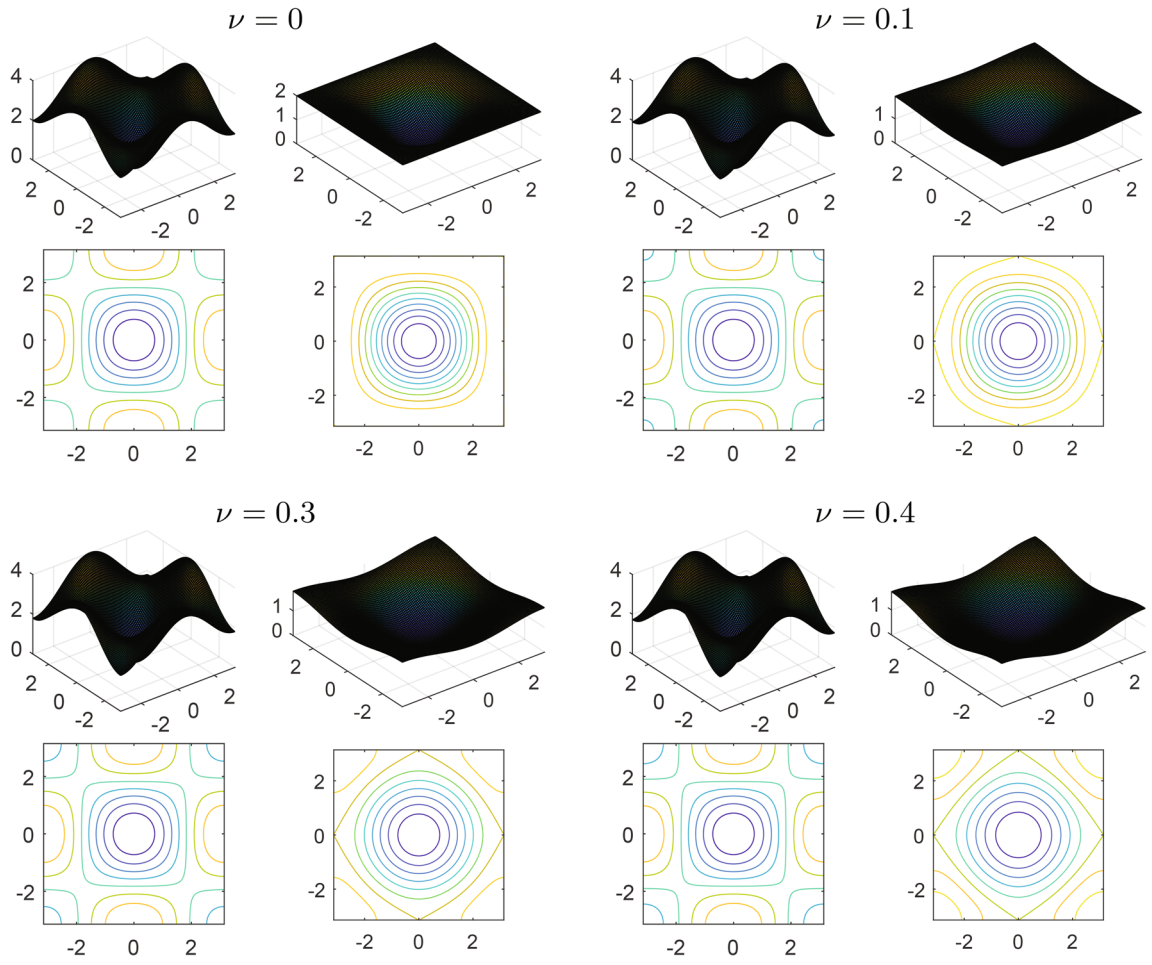

FIGURE 2 Eigenvalues surfaces of the symbol $f_{\mathbb{Q}_1}$ for different ν values and corresponding contour lines

Proposition 3. *Let $f_{\mathbb{Q}_1}(\theta_1, \theta_2)$ be the symbol defined in (11). According to the notation in Proposition 2, all the matrix sequences $\{T_{\boldsymbol{n}}(f_{\mathbb{Q}_1})\}_{\boldsymbol{n}}$, $\{A_n(1, \mathtt{D}^4)\}_n$, $\{A_{\boldsymbol{n}}(1, \mathtt{D}^4)\}_{\boldsymbol{n}}$, $\{A_n(1, \mathtt{DN}^3)\}_n$, $\{A_{\boldsymbol{n}}(1, \mathtt{DN}^3)\}_{\boldsymbol{n}}$ are spectrally distributed as $f_{\mathbb{Q}_1}$ in the sense of Definition 1.*

Proof. For the multilevel block Toeplitz sequences $\{T_{\boldsymbol{n}}(f_{\mathbb{Q}_1})\}_{\boldsymbol{n}}$, $\{A_n(1, \mathtt{D}^4)\}_n$, $\{A_{\boldsymbol{n}}(1, \mathtt{D}^4)\}_{\boldsymbol{n}}$ refer to Item **GLT 2.**, part 1, and Item **GLT 1.**, part 2 in Section 3.1.2. For the sequences $\{A_n(1, \mathtt{DN}^3)\}_n$ and $\{A_{\boldsymbol{n}}(1, \mathtt{DN}^3)\}_{\boldsymbol{n}}$, first observe that the matrix sequence $\{R_{\boldsymbol{n}}\}_{\boldsymbol{n}}$ defined in Proposition 2 is zero-distributed thanks to Propositions 1 and 2, since $\text{rank}(R_{\boldsymbol{n}})/\text{size}(R_{\boldsymbol{n}}) \to 0$, as $\boldsymbol{n} \to \infty$. Then the claim follows thanks to the $*$-algebra structure of the GLT sequences and more specifically thanks to Item **GLT 3.**, part 2, Item **GLT 2.**, part 1, and Item **GLT 1.**, part 2. ∎

Now we analyze few key analytical features of the spectral symbol, which is shared by all the matrix sequences mentioned in Proposition 3. Such a study is important for the analysis of the extremal eigenvalues and of the conditioning of the same matrix sequences and, in Section 5, it will be the main ingredient for designing ad hoc multigrid solvers, when dealing with the related large linear systems.

Theorem 1. *Let $f_{\mathbb{Q}_1}(\theta_1, \theta_2)$ be the symbol defined in (11). The following statements hold true:*

1. $f_{\mathbb{Q}_1}(0,0)\mathbf{e} = 0$, $\mathbf{e} = [1,\ 1]^T$;
2. *both eigenvalues of $f_{\mathbb{Q}_1}$ have a zero of order 2 at $(0,0)$.*

Proof. Claim 1. The function $f_{\mathbb{Q}_1}$ evaluated at $(0,0)$ equals

$$f_{\mathbb{Q}_1}(0,0) = 4\begin{bmatrix} k_1+k_3+k_5+k_7 & k_2+k_4+k_6+k_8 \\ k_2+k_4+k_6+k_8 & k_1+k_3+k_5+k_7 \end{bmatrix},$$

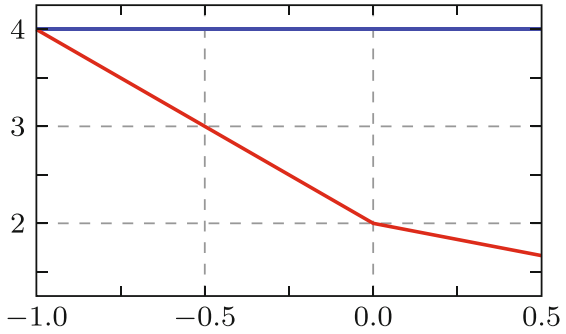

FIGURE 3 Maxima of eigenvalues surfaces of the symbol $f_{\mathbb{Q}_1}$ as functions of ν

TABLE 1 Spectral data for matrices $A_n(1, \mathtt{D}^4)$ and $A_n(1, \mathtt{DN}^3)$ of increasing dimension $N(n)$—case $\nu = 0.4$

$A_n(1, \mathtt{D}^4)$				
$N(n)$	$\lambda_{\min}$	$\lambda_{\min}(h)/\lambda_{\min}(h/2)$	$\mu(h)$	$\mu(h)/\mu(h/2)$
18	6.5599e-01	-	4.8455e+00	-
98	1.8112e-01	3.6219e+00	2.0809e+01	2.3286e-01
450	4.6397e-02	3.9036e+00	8.4925e+01	2.4502e-01
1922	1.1670e-02	3.9759e+00	3.4148e+02	2.4870e-01
7938	2.9218e-03	3.9940e+00	1.3677e+03	2.4967e-01
$A_n(1, \mathtt{DN}^3)$				
40	1.2678e-02	-	5.1460e+00	-
144	4.0891e-03	3.1005e+00	2.1003e+01	2.4501e-01
544	1.1807e-03	3.4632e+00	8.5034e+01	2.4700e-01
2112	3.1877e-04	3.7040e+00	3.4154e+02	2.4897e-01
8320	8.2930e-05	3.8438e+00	1.3678e+03	2.4971e-01

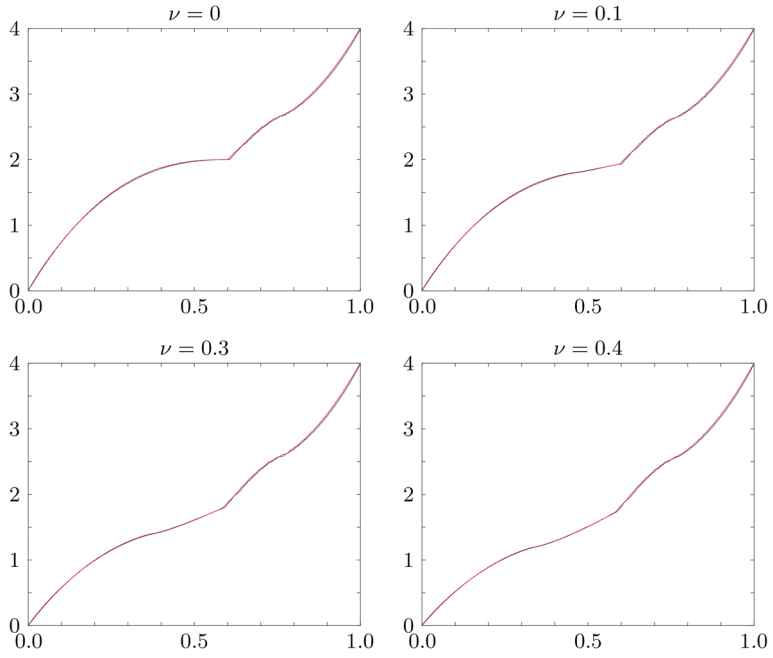

FIGURE 4 Ordered equispaced samplings of $\lambda_j\left(f_{\mathbb{Q}_1}(\theta_1, \theta_2)\right), j = 1, 2$ (red line) and ordered eigenvalues $\lambda_l(A_n(1, \mathtt{D}^4)), l = 1, \ldots, N(n)$, $N(n) = 7938$ (blue line)

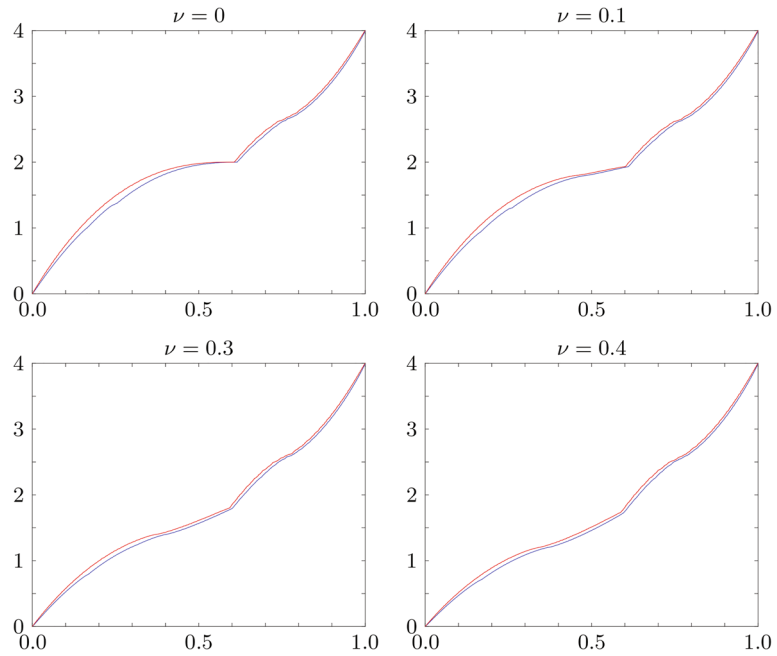

FIGURE 5 Ordered equispaced samplings of $\lambda_j\left(f_{\mathbb{Q}_1}(\theta_1,\theta_2)\right), j = 1, 2$ (red line) and ordered eigenvalues $\lambda_l(A_n(1, \mathtt{DN}^3)), l = 1, \ldots, N(n)$, $N(n) = 8320$ (blue line)

and the row-sum vanishes as $\sum_{i=1}^{8} k_i = 0$ according to (7).

Claim 2. It is just a direct check. Indeed we can evaluate the eigenvalues of the symbol as

$$\lambda_{1,2}(\theta_1,\theta_2) = \frac{1}{2}\left(\mathrm{tr}(f_{\mathbb{Q}_1}(\theta_1,\theta_2)) \pm \sqrt{\mathrm{tr}^2(f_{\mathbb{Q}_1}(\theta_1,\theta_2)) - 4\det(f_{\mathbb{Q}_1}(\theta_1,\theta_2))}\right),$$

where $\mathrm{tr}(f_{\mathbb{Q}_1}(\theta_1,\theta_2))$ and $\det(f_{\mathbb{Q}_1}(\theta_1,\theta_2))$ denote the trace and the determinant of the matrix-valued symbol respectively. By expanding the previous expression, we obtain that

$$\lambda_{1,2}(\theta_1,\theta_2) = \frac{1}{2}\left(\left(\frac{\nu}{3} - 1\right)(\cos\theta_1 + \cos\theta_2 + 2\cos\theta_1\theta_2 - 4) \pm (\nu+1)|\cos\theta_1\cos\theta_2 - 1|\right)$$

with $\nu > -1$. Finally, by considering the Taylor expansion centered at $(0, 0)$ we have

$$\lambda_{1,2}(0,0) = \frac{\partial\lambda_{1,2}(\theta_1,\theta_2)}{\partial\theta_1}_{|(0,0)} = \frac{\partial\lambda_{1,2}(\theta_1,\theta_2)}{\partial\theta_2}_{|(0,0)} = 0 \quad \text{for all } \nu,$$

whereas the second order derivatives form a positive definite Hessian matrix at $(0, 0)$. ∎

Since the minimal eigenvalue of the symbol $f_{\mathbb{Q}_1}(\theta_1,\theta_2)$ has a unique zero of order two at $(\theta_1,\theta_2) = (0, 0)$ (in fact both eigenvalues of the symbol $f_{\mathbb{Q}_1}(\theta_1,\theta_2)$ have a unique zero of order two at $(\theta_1,\theta_2) = (0, 0)$) and since the symbol is positive semi-definite, we can deduce and discuss important information on the extremal eigenvalues and on the conditioning of the corresponding matrix sequences.

Proposition 4. *Let $f_{\mathbb{Q}_1}(\theta_1,\theta_2)$ be the symbol defined in (11). Let us consider a uniform meshing in both directions with n subintervals. Then we have the following relationships*

$$\begin{aligned}
&\lambda_{\min}(A_n(1, \mathtt{D}^4)) \sim n^{-2},\\
&\max\lambda_{\max}(f_{\mathbb{Q}_1}) - \lambda_{\max}(A_n(1, \mathtt{D}^4)) \sim n^{-2},\\
&\kappa_2(A_n(1, \mathtt{D}^4)) \sim n^2,\\
&A_n(1, \mathtt{D}^4) = T_{\boldsymbol{n}}(f_{\mathbb{Q}_1}),\ \boldsymbol{n} = (n_1, n_2),\ n_1 = n_2 = n - 1,
\end{aligned}$$

with $\kappa_2(\cdot)$ denoting the condition number in spectral norm (the one induced by the Euclidean vector norm $\|\cdot\|_2$), and $a_n \sim b_n$ for $a_n, b_n \geq 0$ indicating that there exist positive constants c, C independent of n such that $c \cdot a_n \leq b_n \leq C \cdot a_n$ for any index n (and analogously in the case of multi-indices).

If the more general setting is considered with n_1 subintervals in the x_1 direction and n_2 subintervals in the x_2 direction, then analogous relations are true

$$\begin{aligned}
&\lambda_{\min}(A_{\boldsymbol{n}}(1, \mathbb{D}^4)) \sim [\min n_j]^{-2},\\
&\max \lambda_{\max}(f_{\mathbb{Q}_1}) - \lambda_{\max}(A_{\boldsymbol{n}}(1, \mathbb{D}^4)) \sim [\min n_j]^{-2},\\
&\kappa_2(A_{\boldsymbol{n}}(1, \mathbb{D}^4)) \sim [\min n_j]^{2},\\
&A_{\boldsymbol{n}}(1, \mathbb{D}^4) = T_{\boldsymbol{n}}(f_{\mathbb{Q}_1}),\ \boldsymbol{n} = (n_1 - 1, n_2 - 1).
\end{aligned}$$

Proof. The proof relies essentially on the fact that the involved operators are linear and positive (see Reference 30 for a general treatment of the subject in a matrix theoretic context).

By Theorem 1, second item, it follows that $f_{\mathbb{Q}_1}(\theta_1, \theta_2)$ can be factored as

$$l(\theta_1, \theta_2) g_{\mathbb{Q}_1}(\theta_1, \theta_2), \quad l(\theta_1, \theta_2) = 4 - 2\cos(\theta_1) - 2\cos(\theta_2),$$

where l is the generating function of the standard 2D discrete Laplacian by centered finite differences of order two and minimal bandwidth, and where $g_{\mathbb{Q}_1}$ has eigenvalues given by

$$s_j(\theta_1, \theta_2) = \frac{\lambda_j(f_{\mathbb{Q}_1})}{l(\theta_1, \theta_2)}, \quad j = 1, 2,$$

which are bounded away from zero and infinity since $\lambda_j(f_{\mathbb{Q}_1}), j = 1, 2$, $l(\theta_1, \theta_2)$ are all nonnegative and with a unique zero of order 2 at $(0, 0)$. More precisely there exist $C > c > 0$ such that $c \leq s_1(\theta_1, \theta_2), s_2(\theta_1, \theta_2) \leq C$ uniformly with respect to (θ_1, θ_2) with

$$c = \min\{\min s_1, \min s_2\}, \quad C = \max\{\max s_1, \max s_2\}.$$

Therefore we infer

$$c l(\theta_1, \theta_2) I_2 \leq f_{\mathbb{Q}_1}(\theta_1, \theta_2) \leq C l(\theta_1, \theta_2) I_2,$$

in the sense of the partial ordering in the (real) vector space of the Hermitian matrices. Now, since $T_{\boldsymbol{n}}(\cdot)$ is a linear positive operator, it is also monotone (see e.g., References 23,24), and hence

$$c\lambda_{\min}(T_{\boldsymbol{n}}(l(\theta_1, \theta_2) I_2)) \leq \lambda_{\min}(A_n(1, \mathbb{D}^4)) \leq C\lambda_{\min}(T_{\boldsymbol{n}}(l(\theta_1, \theta_2) I_2)), \tag{12}$$

where

$$\begin{aligned}
\lambda_{\min}(T_{\boldsymbol{n}}(l(\theta_1, \theta_2) I_2)) &= 4 - 2\cos\left(\frac{\pi}{n_1 + 1}\right) - 2\cos\left(\frac{\pi}{n_2 + 1}\right)\\
&= 4\sin^2\left(\frac{\pi}{2(n_1 + 1)}\right) + 4\sin^2\left(\frac{\pi}{2(n_2 + 1)}\right)\\
&\sim n_1^{-2} + n_2^{-2} \sim [\min n_j]^{-2}.
\end{aligned} \tag{13}$$

By combining (12) and (13), we deduce

$$\lambda_{\min}(A_{\boldsymbol{n}}(1, \mathbb{D}^4)) \sim [\min n_j]^{-2},$$

which reduces to

$$\lambda_{\min}(A_{\boldsymbol{n}}(1, \mathbb{D}^4)) \sim n^{-2},$$

if $n_1 = n_2 = n - 1$. The other claim that is

$$\max \lambda_{\max}(f_{\mathbb{Q}_1}) - \lambda_{\max}(A_{\boldsymbol{n}}(1, \mathtt{D}^4)) \sim [\min n_j]^{-2},$$

follows in the same way by duality, simply because

$$\max \lambda_{\max}(f_{\mathbb{Q}_1}) - \lambda_{\max}(A_{\boldsymbol{n}}(1, \mathtt{D}^4)) = \lambda_{\min}(T_{\boldsymbol{n}}(\lambda_{\max}(f_{\mathbb{Q}_1})I_2 - f_{\mathbb{Q}_1})).$$

Therefore the statements on the conditioning represent a direct consequence of the estimates on the extreme of the spectrum, given the positive definite character of the involved matrices. ∎

Remark 1. For the case of $A_n(1, \mathtt{DN}^3)$ the numerical experiments reported in Table 1 confirms that the extremal eigenvalue behavior and the conditioning should be the same as in the case of the pure Dirichlet boundary conditions. Hence the following relations should hold:

$$\begin{aligned}
&\lambda_{\min}(A_n(1, \mathtt{DN}^3)) \sim n^{-2},\\
&\max \lambda_{\max}(f_{\mathbb{Q}_1}) - \lambda_{\max}(A_n(1, \mathtt{DN}^3)) \sim n^{-2},\\
&\kappa_2(A_n(1, \mathtt{DN}^3)) \sim n^2.
\end{aligned}$$

The proof could be conducted using the Sherman-Morrison-Woodbury formula since $A_n(1, \mathtt{DN}^3) = T_{\boldsymbol{n}}(f_{\mathbb{Q}_1}) + R_{\boldsymbol{n}}$, $\boldsymbol{n} = (n_1, n_2)$, $n_1 = n, n_2 = n + 1$, with $R_{\boldsymbol{n}}$ of relatively small rank with respect to the matrix size, as in Proposition 2. The idea relies on a clever writing of $R_{\boldsymbol{n}}$ as XY, with X, Y^T rectangular matrices of the same sizes, the smaller dimension coinciding with the rank of $R_{\boldsymbol{n}}$. The same splitting should be computed also in the general case of $A_{\boldsymbol{n}}(1, \mathtt{DN}^3) = T_{\boldsymbol{n}}(f_{\mathbb{Q}_1}) + R_{\boldsymbol{n}}$, $\boldsymbol{n} = (n_1, n_2 + 1)$. We believe that this direction is worth to be explored in a future investigation.

3.3 | The non-constant coefficient case

The natural extension of the previous analysis refers to the case of a non-constant coefficient ρ. In the physical model, we are assuming the function ρ piecewise constant with respect to the mesh elements $\tau \in \mathcal{T}$. Thus, according to a standard assembling procedure, we can write the stiffness matrix as

$$A_n(\rho) = \sum_{\tau \in \mathcal{T}} \rho_\tau A^{El}_{n,\tau}, \tag{14}$$

where $A^{El}_{n,\tau}$ is the elementary matrix $\boldsymbol{K}_n$ in (6) (possibly properly cut when nodes on the boundary are involved), but widened to size $N(\boldsymbol{n})$ following the chosen ordering of nodes. Here $\boldsymbol{n} = (n - 1, n - 1)$ in the case of Dirichlet boundary conditions and n subintervals in all the directions so that $A_n(\rho) = A_n(\rho, \mathtt{D}^4)$, while $\boldsymbol{n} = (n, n + 1)$ in the case of Dirichlet boundary conditions on one side, Neumann boundary conditions in the other three with n subintervals in all the directions so that $A_n(\rho) = A_n(\rho, \mathtt{DN}^3)$.

Clearly the elementary matrix $\boldsymbol{K}_n$ is positive semi-definite. More precisely according to values k_i in (7), the non-zero eigenvalues are

$$(1 - \nu) \text{ (double)}, \ \frac{1}{2}\left(1 - \frac{\nu}{3}\right) \text{ (double)}, \ 1 + \nu \text{ (simple)}.$$

In addition, every elementary matrix, which has been cut due to nodes on the boundary, is positive semi-definite as well being a principal submatrix of $\boldsymbol{K}_n$ in (6).

Thus, based on (14), we infer

$$\rho_{\min} x^T A_n(1) x \le x^T A_n(\rho) x \le \rho_{\max} x^T A_n(1) x, \quad \text{for all } \ x \neq 0,$$

with $\rho_{\min}$ and $\rho_{\max}$ minimum and maximum of ρ, respectively, so that by applying the Courant–Fischer theorem we claim that

$$\begin{aligned} \rho_{\min}\lambda_{\min}(A_n(1)) &\le \lambda_{\min}(A_n(\rho)) \le \rho_{\max}\lambda_{\min}(A_n(1)),\\ \rho_{\min}\lambda_{\max}(A_n(1)) &\le \lambda_{\max}(A_n(\rho)) \le \rho_{\max}\lambda_{\max}(A_n(1)). \end{aligned} \tag{15}$$

Now putting together the inequalities in (15) and Proposition 4, we deduce that the extremal eigenvalues and the conditioning have the same asymptotical behavior, as in the constant coefficient setting with Dirichlet boundary conditions. When Neumann boundary conditions are used, the result would be true if the conjecture reported in Remark 1 will be proven.

A deeper analysis of the whole eigenvalues distribution can be conveniently performed by referring to the quite general and powerful GLT theory,[22] whose basics have been introduced in Section 3.1.2. Let $D_n(\rho)$ be a multilevel block diagonal sampling matrix according to the notions introduced in Section 3.1.2 and let $A_n(1)$ be the multilevel block Toeplitz matrix $T_{\boldsymbol{n}}(f_{\mathbb{Q}_1})$ if Dirichlet boundary conditions are used or $T_{\boldsymbol{n}}(f_{\mathbb{Q}_1}) + R_{\boldsymbol{n}}$ in the other case. Then the following facts hold:

Fact 1 $\{D_n(\rho)\}_n \sim_{\text{GLT}} \rho$ according to Item **GLT 2.**

Fact 2 $\{R_{\boldsymbol{n}}\}_n \sim_{\text{GLT}} 0$ according to Proposition 2, Proposition 1, and Item **GLT 2.**

Fact 3 $\{T_{\boldsymbol{n}}(f_{\mathbb{Q}_1})\}_n \sim_{\text{GLT}} f_{\mathbb{Q}_1}$ according to Item **GLT 2.**

Fact 4 $\{A_n(1)\}_n \sim_{\text{GLT}} f_{\mathbb{Q}_1}$ according to Fact 2, Fact 3, and to the $*$-algebra structure of GLT sequences that is Item **GLT 3.**

Fact 5 given $\Delta_n = A_n(\rho) - D_n(\rho)A_n(1)$ a simple check shows that $\{\Delta_n\}_n \sim_{\text{GLT}} 0$ since it is obviously true that $\{\Delta_n\}_n \sim_\sigma 0$, by invoking Proposition 1.

Fact 6 $\{A_n(\rho)\}_n \sim_{\text{GLT}} \rho\, f_{\mathbb{Q}_1}$ as a consequence of Fact 5, Fact 1, Fact 4, and of the $*$-algebra structure of GLT sequences that is Item **GLT 3.**; moreover since the matrix sequence $\{A_n(\rho)\}_n$ is made up by Hermitian matrices, by Item **GLT 1.** it follows that $\{A_n(\rho)\}_n \sim_{\sigma,\lambda} \rho\, f_{\mathbb{Q}_1}$.

Fact 7 $\{A_n^{-1}(1)A_n(\rho)\}_n \sim_{\text{GLT}} \rho$ as a direct consequence of Fact 4 and Fact 6, taking into account $*$-algebra structure of GLT sequences that is Item **GLT 3.** Here, for the present preconditioned matrix sequences, we can conclude that the eigenvalues are distributed as ρ, even if the involved matrices are not Hermitian. The reasoning is as follows: $A_n^{-1}(1)A_n(\rho)$ is similar to $A_n^{-1/2}(1)A_n(\rho)A_n^{-1/2}(1)$ which is Hermitian for any size and $\{A_n^{-1/2}(1)A_n(\rho)A_n^{-1/2}(1)\}_n \sim_{\text{GLT}} \rho$ since the square root of a positive definite GLT matrix sequence is still a positive definite GLT matrix sequence (see Reference 19). Therefore

$$\{A_n^{-1/2}(1)A_n(\rho)A_n^{-1/2}(1)\}_n \sim_{\sigma,\lambda} \rho,$$

and by similarity, we infer that

$$\{A_n^{-1}(1)A_n(\rho)\}_n \sim_{\sigma,\lambda} \rho.$$

In the preceding series of statements we have used the notation $\{A_n(\rho)\}_n$ for several matrix sequences. More in detail, we stress that the main facts that is Fact 6 and Fact 7 apply to all the matrix sequences

$$\{A_n(\rho)\}_n \in \left\{\{A_n(\rho, \mathtt{D}^4)\}_n, \{A_{\boldsymbol{n}}(\rho, \mathtt{D}^4)\}_{\boldsymbol{n}}, \{A_n(\rho, \mathtt{DN}^3)\}_n, \{A_{\boldsymbol{n}}(\rho, \mathtt{DN}^3)\}_{\boldsymbol{n}}\right\}.$$

In fact, the previous matrix sequences represent the variable coefficient versions of the different matrix sequences reported in Proposition 3, for the constant coefficient case when $\rho \equiv 1$.

4 | TWO-GRID AND MULTIGRID METHODS

In this section, we concisely report few relevant results concerning the convergence theory of algebraic multigrid methods[31], with special attention to the case of multilevel block Toeplitz structures generated by a matrix-valued symbol f.

We start by taking into consideration the generic linear system $A_m x_m = b_m$ with large dimension m, where $A_m \in \mathbb{C}^{m\times m}$ is a Hermitian positive definite matrix and $x_m, b_m \in \mathbb{C}^m$. Let $m_0 = m > m_1 > \cdots > m_s > \cdots > m_{s_{\min}}$ and let $P_s^{s+1} \in \mathbb{C}^{m_{s+1}\times m_s}$ be a full-rank matrix for any s. At last, let us denote by $\mathcal{V}_s$ a class of stationary iterative methods for given linear systems of dimension m_s.

In accordance with Reference 32, the algebraic Two-Grid Method (TGM) can be easily seen as a stationary iterative method whose generic steps are reported below, where we refer to the dimension m_s by means of its subscript s. More specifically $x_s^{\text{out}} = \text{TGM}(s, x_s^{\text{in}}, b_s)$, with $x_s^{\text{pre}} = \mathcal{V}_{s,\text{pre}}^{\nu_{\text{pre}}}(x_s^{\text{in}}, b_s)$ (the *pre-smoothing iteration*), $r_s = A_s x_s^{\text{pre}} - b_s$, $r_{s+1} = P_{m_s}^{m_{s+1}} r_s$, $A_{s+1} = P_{m_s}^{m_{s+1}} A_s (P_{m_s}^{m_{s+1}})^H$, Solve $A_{s+1} y_{s+1} = r_{s+1}$, $\hat{x}_s = x_s^{\text{pre}} - (P_{m_s}^{m_{s+1}})^H y_{s+1}$ (the *exact coarse grid correction*), and finally $x_s^{\text{out}} = \mathcal{V}_{s,\text{post}}^{\nu_{\text{post}}}(\hat{x}_s, b_s)$ (the *post-smoothing iteration*).

The resulting iteration matrix of the TGM is then defined as $TGM_s = V_{s,\text{post}}^{\nu_{\text{post}}} CGC_s V_{s,\text{pre}}^{\nu_{\text{pre}}}$ where $CGC_s = I^{(s)} - (P_{m_s}^{m_{s+1}})^H A_{s+1}^{-1} P_{m_s}^{m_{s+1}} A_s$ and $A_{s+1} = P_{m_s}^{m_{s+1}} A_s (P_{m_s}^{m_{s+1}})^H$. In this setting $V_{s,\text{pre}}$ and $V_{s,\text{post}}$ represent the pre-smoothing and post-smoothing iteration matrices, respectively, and $I^{(s)}$ is the identity matrix at the sth level.

By employing a recursive procedure, the TGM leads to a Multi-Grid Method (MGM). Indeed the standard V-cycle can be expressed in the following way:

$x_s^{\text{out}} = \text{MGM}(s, x_s^{\text{in}}, b_s)$

if $s \le s_{\min}$ then	
Solve $A_s x_s^{\text{out}} = b_s$	Exact solution
else	
$x_s^{\text{pre}} = \mathcal{V}_{s,\text{pre}}^{\nu_{\text{pre}}}(x_s^{\text{in}}, b_s)$	Pre-smoothing iterations
$r_s = A_s x_s^{\text{pre}} - b_s$ $r_{s+1} = P_{m_s}^{m_{s+1}} r_s$ $y_{s+1} = \text{MGM}(s+1, \mathbf{0}_{s+1}, r_{s+1})$ $\hat{x}_s = x_s^{\text{pre}} - (P_{m_s}^{m_{s+1}})^H y_{s+1}$	Coarse grid correction
$x_s^{\text{out}} = \mathcal{V}_{s,post}^{\nu_{\text{post}}}(\hat{x}_s, b_s)$	Post-smoothing iterations

From a computational viewpoint, it is more efficient to compute the matrices $A_{s+1} = P_s^{s+1} A_s (P_s^{s+1})^H$ in the so called *setup phase* as this reduces the total costs.

According to the previous setting, the global iteration matrix of the MGM is recursively defined as

$$MGM_{s_{\min}} = O \in \mathbb{C}^{s_{\min} \times s_{\min}},$$
$$MGM_s = V_{s,\text{post}}^{\nu_{\text{post}}} \left[I^{(s)} - (P_{m_s}^{m_{s+1}})^H \left(I^{(s+1)} - MGM_{s+1} \right) A_{s+1}^{-1} P_{m_s}^{m_{s+1}} A_s \right] V_{s,\text{pre}}^{\nu_{\text{pre}}}, \quad s = s_{\min} - 1, \ldots, 0.$$

Remark 2. In the relevant literature (see, for instance, Reference 33), the convergence analysis of the TGM splits into the validation of two separate conditions: the smoothing property and the approximation property. Regarding the latter, with reference to scalar structured matrices,[33,34] the TGM optimality is given in terms of choosing the proper conditions that the symbol p of a family of projection operators has to fulfill. Indeed, let $T_n(f)$ with $n = (2^t - 1)$, f a nonnegative trigonometric polynomial. Let θ^0 be the unique zero of f . Then the TGM optimality applied to $T_n(f)$ is guaranteed if we choose the symbol p of the family of projection operators such that

$$\limsup_{\theta \to \theta^0} \frac{|p(\eta)|^2}{f(\theta)} < \infty, \ \eta \in \mathcal{M}(\theta),$$
$$\sum_{\eta \in \Omega(\theta)} |p(\eta)|^2 > 0, \tag{16}$$

where, for $d = 1$, the sets $\Omega(\theta)$ and $\mathcal{M}(\theta)$ are the following corner and mirror points

$$\Omega(\theta) = \{\eta \in \{\theta, \theta + \pi\}\}, \quad \mathcal{M}(\theta) = \Omega(\theta) \setminus \{\theta\},$$

respectively. In the general case of $d > 1$, we have

$$\Omega(\boldsymbol{\theta}) = \{\boldsymbol{\eta} \in \{\boldsymbol{\theta} + \pi \boldsymbol{s}\}, \boldsymbol{s} = (s_1, \ldots, s_d), s_j \in \{0, 1\}, j = 1, \ldots, d\}$$

with $\mathcal{M}(\boldsymbol{\theta}) = \Omega(\boldsymbol{\theta}) \setminus \{\boldsymbol{\theta}\}$, so that the cardinality of $\Omega(\boldsymbol{\theta})$ and $\mathcal{M}(\boldsymbol{\theta})$ is 2^d and $2^d - 1$, respectively.

TABLE 2 Number of multigrid iterations for matrices $A_n(1, \mathtt{D}^4)$ and $A_n(1, \mathtt{DN}^3)$ of increasing dimension $N(n)$, $\varepsilon = 10^{-6}$

$\boldsymbol{A_n(1, \mathtt{D}^4)}$	$\boldsymbol{A_n(1, \mathtt{DN}^3)}$						
	$\boldsymbol{\nu = 0.1}$						
$N(n)$	Twogrid	Vcycle	Wcycle	$N(n)$	Twogrid	Vcycle	Wcycle
18	4	1	1	40	8	1	1
98	6	6	6	144	7	7	8
450	7	7	7	544	8	8	8
1922	7	7	7	2112	8	9	8
7938	7	7	7	8320	8	9	8
32,258	7	7	7	33,024	8	9	8
	$\nu = 0.2$						
18	5	-	-	40	8	-	-
98	6	6	6	144	8	8	8
450	7	7	7	544	8	9	8
1922	7	8	7	2112	8	9	8
7938	7	8	7	8320	8	9	8
32,258	7	8	7	33,024	8	9	8
	$\nu = 0.4$						
18	6	-	-	40	9	-	-
98	8	8	8	144	9	9	9
450	9	9	9	544	9	10	9
1922	9	9	9	2112	9	10	9
7938	9	9	9	8320	9	11	9
32,258	9	10	9	33,024	9	11	9

Informally, for $d = 1$, it means that the optimality of the two-grid method is obtained by choosing the family of projection operators associated to a symbol p such that $|p|^2(\vartheta) + |p|^2(\vartheta + \pi)$ does not have zeros and $|p|^2(\vartheta + \pi)/f(\vartheta)$ is bounded (if we require the optimality of the V-cycle then the second condition is a bit stronger); see Reference 33. When discretizing differential operators, the previous conditions mean that p has a zero of order at least α at $\vartheta = \pi$, whenever f has a zero at $\theta^0 = 0$ of order 2α, since the involved Toeplitz-like structures have a spectral symbol with a unique zero at $\theta^0 = 0$ of order 2α.

In our specific block setting, by interpreting the analysis given in Reference 35, all the involved symbols are matrix-valued and the conditions which generalize (16) and are sufficient for the TGM convergence and optimality are the following:

A) zero of order 2 at all the mirror points of the eigenvalue functions of the symbol of the projector for our matrix sequences having common symbol $f_{\mathbb{Q}_1}$ (mirror point theory[33,34]),
B) positive definiteness of $\sum_{\boldsymbol{\eta} \in \Omega(\theta)} pp^H(\boldsymbol{\eta})$,
C) commutativity of all $p(\boldsymbol{\eta})$ for $\boldsymbol{\eta}$ varying in the corner points.

Even if the theoretical extension to the V-Cycle and W-Cycle convergence and optimality is not given, in the subsequent section we propose specific choices of the projection operators numerically showing how this leads to two-grid, V-cycle, W-cycle procedures converging optimally or quasi-optimally with respect to all the relevant parameters (size, dimensionality, polynomial degree k).

Our choices are in agreement with the mathematical conditions set in items **A)**, **B)**, and **C)**. We remark that the violation of condition **C)** is discussed in the conclusion section in Reference 35.

5 | MULTIGRID PROPOSALS

According to the theory in Section 4, we define the prolongation operator as follows

$$P_{2h}^h = P \otimes P \otimes I_2,$$

in the case of matrices $A_n(1, \mathtt{D}^4)$ and

$$P_{2h}^h = P_c \otimes P_t \otimes I_2,$$

in the case of matrices $A_n(1, \mathtt{DN}^3)$. In the expressions above,

$$P = \begin{bmatrix} \frac{1}{2} & & & & \\ 1 & & & & \\ \frac{1}{2} & \frac{1}{2} & & & \\ & 1 & & & \\ & \frac{1}{2} & \ddots & \frac{1}{2} & \\ & & & 1 & \\ & & & \frac{1}{2} & \end{bmatrix}, \quad P_t = \begin{bmatrix} 1 & & & & & \\ \frac{1}{2} & \frac{1}{2} & & & & \\ & 1 & & & & \\ & \frac{1}{2} & \frac{1}{2} & & & \\ & & 1 & & & \\ & & \frac{1}{2} & \ddots & \frac{1}{2} & \\ & & & & 1 & \\ & & & & \frac{1}{2} & \frac{1}{2} \\ & & & & & 1 \end{bmatrix}, \quad \text{and} \quad P_c = \begin{bmatrix} \frac{1}{2} & & & & & \\ 1 & & & & & \\ \frac{1}{2} & \frac{1}{2} & & & & \\ & 1 & & & & \\ & \frac{1}{2} & \ddots & \frac{1}{2} & & \\ & & & 1 & & \\ & & & \frac{1}{2} & \frac{1}{2} & \\ & & & & 1 & \end{bmatrix}.$$

Note that $A_{2h} = P^T A_h P$ equals the same FEM approximation on the coarse mesh in both cases, independently of the boundary conditions.

Again, independently of the boundary conditions, the symbol of the prolongation operator is

$$p(\theta_1, \theta_2) = (1 + \cos(\theta_1))(1 + \cos(\theta_2))I_2.$$

Since the matrix-valued function $p(\theta_1, \theta_2)$ is a scalar function times the identity, it is evident that the quite delicate condition **C)** is automatically satisfied. Furthermore, taking into account Theorem 1, since the scalar function $1 + \cos(\theta)$ has a zero of order 2 at $\theta = \pi$, we deduce that condition **A)** holds, while the positive definiteness of $\sum_{\boldsymbol{\eta} \in \Omega(\theta)} pp^H(\boldsymbol{\eta})$ is also trivially verified so that also condition **B)** is met.

In addition, one iteration of Gauss–Seidel as pre and post smoother is considered. In view of the analysis reported in the previous section, we expect that our multigrid method is convergent in an optimal way that is with a convergence speed independent of the matrix size. Table 2 confirms such a theoretical forecast.

We have performed numerical experiments also with respect to other solvers. The incomplete Cholesky factorization preconditioning and the block circulant preconditioning with proper Strang-type correction (see Table 3), as expected, both improve the convergence with respect to the classical conjugate gradient. However, the iteration count grows (according to the theory) as the square root of the matrix size of the involved matrices. On the other hand, our multigrid strategy is not only optimal (number of iterations independent of the matrix size), but also the concrete iteration count is very small (less than 10 iterations) and the computational cost is optimal that is linear and proportional to the cost of the matrix-vector product.

In the last numerical tests reported in Tables 4 and 5, we compare our approach both as a multigrid solver and a one-step iteration in a preconditioned flexible conjugate gradient (FCG) to professional software (see References 36–40 and references therein). As we can see from the subsequent tables, our procedures employ fewer iterations, with a comparable cost per iteration, and show clear optimality. The optimality is not observed using the Notay codes. For the sake of completeness, we again remark that the CPU timings are not reported since our codes are really basic and indeed they would need further optimization for a fair comparison with professional and well-established codes. At any rate, even if the CPU times are not reported, we are sure that also the cost of a single iteration is favorable in our approach both in terms of theoretical complexity and in terms of timing. The reason is that we do not manipulate matrices when using our prolongation/restriction operators but only polynomial symbols: whence only $O(1)$ coefficients are used, and the related timings are very compressed.

TABLE 3 Number of PCG iterations for matrices $A_n(1, \mathrm{D}^4)$ and $A_n(1, \mathrm{DN}^3)$ of increasing dimension $N(n)$, $\epsilon = 10^{-6}$

$A_n(1, \mathrm{D}^4)$				$A_n(1, \mathrm{DN}^3)$			
$\nu = 0.1$							
$N(n)$	I	IC	C	$N(n)$	I	IC	C
18	9	5	11	40	32	9	25
98	20	8	17	144	60	15	36
450	41	12	27	544	119	28	54
1922	73	21	46	2112	231	53	90
7938	128	41	60	8320	450	104	152
$\nu = 0.2$							
18	10	5	11	40	32	8	25
98	21	8	18	144	65	16	37
450	42	14	28	544	126	28	56
1922	76	24	47	2112	240	53	94
7938	133	41	61	8320	462	106	154
$\nu = 0.4$							
18	11	5	11	40	33	10	25
98	24	9	19	144	70	17	39
450	44	14	31	544	134	30	61
1922	77	25	49	2112	260	58	97
7938	144	45	61	8320	497	114	163

TABLE 4 Number of HSL-MI20 iterations (pure multigrid and AMG preconditioner[37]) and AMG-Notay[36] for the matrices $A_n(1, \mathrm{D}^4)$ and $A_n(1, \mathrm{DN}^3)$ of increasing dimension $N(n)$, $\epsilon = 10^{-6}$, where AMG-Notay means FCG with one iteration of the AMG designed by Notay as smoother

$A_n(1, \mathrm{D}^4)$				$A_n(1, \mathrm{DN}^3)$			
$\nu = 0.1$							
$N(n)$	HSL-MI20	HSL-MI20-prec	AMG-Notay	$N(n)$	HSL-MI20	HSL-MI20-prec	AMG-Notay
18	4	3	1	40	17	6	1
98	6	4	1	144	57	8	1
450	9	6	9	544	171	13	15
1922	20	10	10	2112	475	21	20
7938	44	17	11	8320	970	39	22
32,258	96	23	11	33,024	1149	54	24
$\nu = 0.2$							
18	4	3	1	40	19	6	1
98	6	5	1	144	57	8	1
450	11	7	10	544	178	13	12
1922	20	11	13	2112	457	22	21
7938	59	18	16	8320	719	41	24
32,258	108	28	17	33,024	1369	72	27
$\nu = 0.4$							
18	5	3	1	40	22	6	1
98	7	5	1	144	66	9	1
450	14	7	10	544	212	14	16
1922	37	12	13	2112	605	25	19
7938	85	21	15	8320	1389	46	21
32,258	141	31	16	33,024	1891	86	23

TABLE 5 Number of FCG iterations with one iteration of multigrid (either Twogrid or V-cycle or W-cycle) as preconditioner for the matrices $A_n(1, \mathtt{D}^4)$ and $A_n(1, \mathtt{DN}^3)$ of increasing dimension $N(n)$, $\epsilon = 10^{-6}$

$A_n(1, \mathtt{D}^4)$				$A_n(1, \mathtt{DN}^3)$			
$\nu = 0.1$							
$N(n)$	Twogrid	Vcycle	Wcycle	$N(n)$	Twogrid	Vcycle	Wcycle
18	4	-	-	40	6	-	-
98	6	6	6	144	7	7	7
450	6	6	6	544	8	8	8
1922	6	6	6	2112	8	9	8
7938	6	6	6	8320	8	9	8
32,258	6	6	6	33,024	8	9	8
$\nu = 0.2$							
18	4	-	-	40	6	-	-
98	6	6	6	144	7	7	7
450	6	7	6	544	8	8	8
1922	6	7	6	2112	8	9	8
7938	6	7	6	8320	8	9	8
32,258	6	6	6	33,024	8	9	8
$\nu = 0.4$							
18	5	-	-	40	7	-	-
98	7	7	7	144	8	8	8
450	7	7	7	544	9	9	9
1922	7	7	7	2112	9	10	9
7938	7	7	7	8320	9	10	9
32,258	7	7	7	33,024	9	10	9

6 | CONCLUSIONS

In this work, we have considered a problem that stems from topology optimization, which aims to find the best material layout subject to assigned constraints. When solving the related governing equation using the FEM, a large number of elements is employed to discretize the design domain, and an element-wise constant function approximates the coefficient field in the considered 2D design domain. First, we have provided a spectral analysis of the coefficient matrices associated with the linear systems stemming from the FE discretization of a linearly elastic problem, for an arbitrary coefficient field. Based on the spectral information, we have proposed a specialized multigrid method, which turned out to be optimal, in the sense that the (arithmetic) cost for solving the related linear systems, up to a fixed desired accuracy, is proportional to the matrix-vector cost. The method has been tested, and the numerical results are very satisfactory, in terms of linear cost number of iterations, which is bounded by a constant independent of the matrix size and lightly influenced by the desired accuracy. We finally remark that the used tools are very flexible, and in particular, they do not depend on the number of spatial dimensions of the underlying problem. Therefore, we are confident that a generalization of our spectral analysis and the related algorithmic proposals can be obtained in a three-dimensional setting as well, even if to the price of quite heavy notations.

Finally, what is conjectured in Remark 1 and it is numerically observed in Table 1 will be the subject of future investigations, in order to complete the spectral analysis of the considered matrices, with Neumann boundary conditions on at least one side.

ACKNOWLEDGMENT

This work is financially supported by the Swedish strategic research programme eSSENCE, a strategic collaborative eScience program funded by the Swedish Research Council, and the Italian Institution for High Mathematics (INdAM).

CONFLICT OF INTEREST

This study does not have any conflicts to disclose.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID

Quoc Khanh Nguyen https://orcid.org/0000-0001-9019-2795
Stefano Serra-Capizzano https://orcid.org/0000-0001-9477-109X
Cristina Tablino-Possio https://orcid.org/0000-0003-1424-2767
Eddie Wadbro https://orcid.org/0000-0001-8704-9584

REFERENCES

1. Bendsøe MP, Sigmund O. Topology optimization. Theory, methods, and applications. New York, NY: Springer; 2003.
2. Wadbro E, Engström C. Topology and shape optimization of plasmonic nano-antennas. Comput Methods Appl Mech Eng. 2015;293:155–69.
3. Elesin Y, Lazarov BS, Jensen JS, Sigmund O. Time domain topology optimization of 3D nanophotonic devices. Photonics Nanostructures Fundam Appl. 2014;12(1):23–33.
4. Erentok A, Sigmund O. Topology optimization of sub-wavelength antennas. IEEE Trans Antennas Propag. 2011;59(1):58–69.
5. Andreasen C, Sigmund O. Topology optimization of fluid–structure-interaction problems in poroelasticity. Comput Methods Appl Mech Eng. 2013;258:55–62.
6. Yoon GH. Topology optimization for stationary fluid–structure interaction problems using a new monolithic formulation. Int J Numer Methods Eng. 2010;82:591–616.
7. Kook J, Koo K, Hyun J, Jensen JS, Wang S. Acoustical topology optimization for Zwicker's loudness model — Application to noise barriers. Comput Methods Appl Mech Eng. 2012;237–240:130–51.
8. Christiansen RE, Lazarov BS, Jensen JS, Sigmund O. Creating geometrically robust designs for highly sensitive problems using topology optimization: acoustic cavity design. Struct Multidiscip Optim. 2015;52(4):737–54.
9. Wadbro E, Niu B. Multiscale design for additive manufactured structures with solid coating and periodic infill pattern. Comput Methods Appl Mech Eng. 2019;357:112605.
10. Clausen A, Aage N, Sigmund O. Topology optimization of coated structures and material interface problems. Comput Methods Appl Mech Eng. 2015;290:524–41.
11. Park J, Sutradhar A. A multi-resolution method for 3D multi-material topology optimization. Comput Methods Appl Mech Eng. 2015;285:571–86.
12. Klarbring A, Strömberg N. Topology optimization of hyperelastic bodies including non-zero prescribed displacements. Struct Multidiscip Optim. 2013;47(1):37–48.
13. Zhu JH, Zhang WH, Xia L. Topology optimization in aircraft and aerospace structures design. Arch Comput Methods Eng. 2016;23(4):595–622.
14. Berggren M, Kasolis F. Weak material approximation of holes with traction-free boundaries. SIAM J Numer Anal. 2012;50(4):1827–48.
15. Buhl T, Pedersen CBW, Sigmund O. Stiffness design of geometrically nonlinear structures using topology optimization. Struct Multidiscip Optim. 2000;19(2):93–104.
16. Bruns TE, Tortorelli DA. An element removal and reintroduction strategy for the topology optimization of structures and compliant mechanisms. Int J Numer Methods Eng. 2003;57(10):1413–30.
17. Bruns TE. Zero density lower bounds in topology optimization. Comput Methods Appl Mech Eng. 2006;196(1):566–78.
18. Nguyen QK, Serra-Capizzano S, Wadbro E. On using a zero lower bound on the physical density in material distribution topology optimization. Comput Methods Appl Mech Eng. 2020;359:112669.
19. Garoni C, Serra-Capizzano S. Generalized locally Toeplitz sequences: theory and applications. Vol I. New York, NY: Springer; 2017.
20. Garoni C, Serra-Capizzano S. Generalized locally Toeplitz sequences: theory and applications. Vol II. New York, NY: Springer; 2018.
21. Barbarino G, Garoni C, Serra-Capizzano S. Block generalized locally Toeplitz sequences: theory and applications in the multidimensional case. Electron Trans Numer Anal. 2020;53:113–216.
22. Garoni C, Mazza M, Serra-Capizzano S. Block generalized locally Toeplitz sequences: from the theory to the applications. Axioms. 2018;7(3):49.
23. Serra S. Asymptotic results on the spectra of block Toeplitz preconditioned matrices. SIAM J Matrix Anal Appl. 1999;20-1:31–44.
24. Serra S. Spectral and computational analysis of block Toeplitz matrices having nonnegative definite matrix-valued generating functions. BIT Numer Math. 1999;39-1:152–75.
25. Axelsson O, Barker VA. Finite element solution of boundary value problems. Theory and computation. Computer Science and Applied Mathematics. Orlando, FL: Academic Press; 1984.

26. Dorostkar A, Neytcheva M, Serra-Capizzano S. Spectral analysis of coupled PDEs and of their Schur complements via generalized locally Toeplitz sequences in 2D. Comput Methods Appl Mech Eng. 2016;305:74–105.
27. Kuijlaars ABJ. Convergence analysis of Krylov subspace iterations with methods from potential theory. SIAM Rev. 2006;48(1):3–40.
28. Beckermann B, Serra-Capizzano S. On the asymptotic spectrum of finite element matrix sequences. SIAM J Numer Anal. 2007;45(2):746–69.
29. Gurtin ME. An introduction to continuum mechanics. Vol 158. New York, NY: Academic Press; 1981.
30. Serra S. Some theorems on linear positive operators and functionals and their applications. Comput Math Appl. 2000;39-7(8):139–67.
31. Stüben K. An introduction to algebraic multi-grid. In: Trottenberg U, Oosterlee C, Schüller A, editors. Multigrid. Cambridge, MA: Academic Press; 2001. p. 413–532.
32. Hackbusch W. Multigrid methods and applications. Berlin/Heidelberg, Germany: Springer-Verlag; 1985.
33. Aricò A, Donatelli M, Serra-Capizzano S. V-cycle optimal convergence for certain (multilevel) structured linear systems. SIAM J Matrix Anal Appl. 2004;26:186–214.
34. Fiorentino G, Serra S. Multigrid methods for toeplitz matrices. Calcolo. 1991;28:283–305.
35. Donatelli M, Ferrari P, Furci I, Serra-Capizzano S, Sesana D. Multigrid methods for block-Toeplitz linear systems: convergence analysis and applications. Numer Linear Algebra Appl. 2021;28(4):e2356.
36. Notay Y. AGMG software and documentation. Available from: http://agmg.eu
37. HSL. A collection of Fortran codes for large scale scientific computation. Available from: http://www.hsl.rl.ac.uk/
38. Notay Y. An aggregation-based algebraic multigrid method. Electron Trans Numer Anal. 2010;37:123–46.
39. Notay Y. Aggregation-based algebraic multigrid for convection-diffusion equations. SIAM J Sci Comput. 2012;34:2288–316.
40. Napov A, Notay Y. An algebraic multigrid method with guaranteed convergence rate. SIAM J Sci Comput. 2012;34:1079–109.
41. Ho DT, Park S-D, Kwon S-Y, Park K, Kim SY. Negative Poisson's ratios in metal nanoplates. Nat Commun. 2014;5:3255.
42. Love AEH, Goldstine HH. A treatise on the mathematical theory of elasticity. New York, NY: Dover Publications; 1944.
43. Mott PH, Roland CM. Limits to Poisson's ratio in isotropic materials. Phys Rev B. 2009;80:132104.

How to cite this article: Nguyen QK, Serra-Capizzano S, Tablino-Possio C, Wadbro E. Spectral analysis of the finite element matrices approximating 2D linearly elastic structures and multigrid proposals. Numer Linear Algebra Appl. 2022;29(4):e2433. https://doi.org/10.1002/nla.2433

APPENDIX A. STRESS–STRAIN RELATION AND VARIOUS BOUNDS

In three-dimensions, there are six independent strain components in total at a point in an element and they are written as a vector

$$\boldsymbol{\epsilon} = [\epsilon_{11} \;\; \epsilon_{22} \;\; \epsilon_{33} \;\; 2\epsilon_{12} \;\; 2\epsilon_{23} \;\; 2\epsilon_{31}]^T.$$

Similarly, corresponding to the six strain components above, there are also six independent stress components written in vector form as

$$\boldsymbol{\sigma} = [\sigma_{11} \;\; \sigma_{22} \;\; \sigma_{33} \;\; \sigma_{12} \;\; \sigma_{23} \;\; \sigma_{31}]^T.$$

By the generalized Hooke's law, the most general linear relation among components of the stress and strain tensor can then be written as

$$\boldsymbol{\sigma} = \boldsymbol{E}\boldsymbol{\epsilon}, \tag{A1}$$

where $\boldsymbol{E}$ is a matrix that corresponds to the constant fourth-order elasticity tensor. The relationship between stresses and strains is

$$\begin{aligned} \epsilon_{11} &= \frac{1}{E_0}\left(\sigma_{11} - \nu(\sigma_{22}+\sigma_{33})\right), \quad \epsilon_{12} = \frac{\sigma_{12}}{2G}, \quad \epsilon_{13} = \frac{\sigma_{13}}{2G}, \\ \epsilon_{22} &= \frac{1}{E_0}\left(\sigma_{22} - \nu(\sigma_{11}+\sigma_{33})\right), \quad \epsilon_{23} = \frac{\sigma_{23}}{2G}, \quad \epsilon_{33} = \frac{1}{E_0}\left(\sigma_{33} - \nu(\sigma_{11}+\sigma_{22})\right), \end{aligned} \tag{A2}$$

where ν is Poisson's ratio, E_0 is Young's modulus, and

$$G = \frac{E_0}{2+2\nu}, \tag{A3}$$

is the shear modulus.

However, in this article, we consider the two-dimensions plane stress condition that implies $\frac{\partial}{\partial x_3} = 0$ and $\sigma_{13} = \sigma_{23} = \sigma_{33} = 0$. Then, relations (A2) reduces to

$$\epsilon_{11} = \frac{1}{E_0}(\sigma_{11} - \nu\sigma_{22}), \qquad \epsilon_{22} = \frac{1}{E_0}(\sigma_{22} - \nu\sigma_{11}), \qquad \epsilon_{33} = -\frac{\nu}{E_0}(\sigma_{11} + \sigma_{22}),$$
$$\epsilon_{12} = \frac{\sigma_{12}}{2G}, \qquad \epsilon_{13} = 0, \qquad \epsilon_{23} = 0.$$

Moreover, by combining relations for ϵ_{11} and ϵ_{22}, we find that $\epsilon_{33} = -\frac{\nu}{1-\nu}(\epsilon_{11} + \epsilon_{22})$. Hence, the stress–strain relation can be reduced to the stress and strain components in the x_1x_2-plane.

$$\begin{bmatrix} \epsilon_{11} \\ \epsilon_{22} \\ 2\epsilon_{12} \end{bmatrix} = \frac{1}{E_0}\begin{bmatrix} 1 & -\nu & 0 \\ -\nu & 1 & 0 \\ 0 & 0 & 2(1+\nu) \end{bmatrix}\begin{bmatrix} \sigma_{11} \\ \sigma_{22} \\ \sigma_{12} \end{bmatrix}. \tag{A4}$$

Analogous the three-dimensional case, equation system (A4) can be inverted to obtain Hooke's law (A1) with

$$\boldsymbol{E} = \frac{E_0}{(1-\nu)(1+\nu)}\begin{bmatrix} 1 & \nu & 0 \\ \nu & 1 & 0 \\ 0 & 0 & \frac{1-\nu}{2} \end{bmatrix}.$$

In this case, the bulk modulus K can be expressed as

$$K = \frac{(\sigma_{11} + \sigma_{22})/2}{\epsilon_{11} + \epsilon_{22}} = \frac{E_0}{2(1-\nu)}. \tag{A5}$$

Also in this case, the Young (E_0), shear (G), and bulk (K) moduli need to be positive. Thus, Equations (A3) and (A5) imply that the Poisson ratio in the two dimensional plane stress setting must satisfy $-1 < \nu < 1$. We remark that the upper-bound in this case is larger than the upper-bound in the three dimensional setting.

Remark 3. Doing a similar analysis as above for the two-dimensional plane strain as well as in the three-dimensional case yields that the Poisson's ratio belongs to an open interval of the form $-1 < \nu < 0.5$. As a consequence, the plane stress setting allows for exotic (unphysical) materials with Poisson's ratios larger than 0.5, which is the incompressibility limit for the underlying physical problem (the three-dimensional case). To the best of our knowledge, there are no known natural occurring isotropic materials for $\nu < 0$. On the other hand, one can design isotropic material with negative Poisson's ratio using theoretical analysis and simulations.[41] Nevertheless as a matter of practice,[42] we limit our attention to Poisson's ratio satisfying the following inequalities $0 \leq \nu < 0.5$. However, in some more delicate circumstances, the Poisson ratio should lie in the interval $0.2 \leq \nu < 0.5$, as proved experimentally in view of elastic properties of real isotropic materials.[43]