Spectral analysis of coupled PDEs and of their Schur complements via the notion of Generalized Locally Toeplitz sequences

Ali Dorostkar*, Maya Neytcheva*,
Stefano Serra-Capizzano†

Abstract

We consider large linear systems of algebraic equations arising from the Finite Element approximation of coupled partial differential equations. As case study we focus on the linear elasticity equations, formulated as a saddle point problem to allow for modeling of purely incompressible materials. Using the notion of the so-called *spectral symbol* in the Generalized Locally Toeplitz (GLT) setting, we derive the GLT symbol (in the Weyl sense) of the sequence of matrices $\{A_n\}$ approximating the elasticity equations. Further, exploiting the property that the GLT class defines an algebra of matrix sequences and the fact that the Schur complements are obtained via elementary algebraic operation on the blocks of A_n, we derive the symbols $f^{\mathcal{S}}$ of the associated sequences of Schur complements $\{S_n\}$. As a consequence of the GLT theory, the eigenvalues of S_n for large n are described by a sampling of $f^{\mathcal{S}}$ on a uniform grid of its domain of definition. We extend the existing GLT technique with novel elements, related to block-matrices and Schur complement matrices, and illustrate the theoretical findings with numerical tests.

Key words: coupled systems of PDEs, Schur complement, Toeplitz matrix, GLT sequence, joint eigenvalue distribution

1 Introduction: notation and preliminaries

This section consists of two parts. First we fix the notations by providing basic definitions and concepts, regarding Toeplitz and circulant matrices, and noteworthy spectral properties. Second, we introduce a link between structures of Toeplitz type and approximations of partial differential equations.

*Department of Information Technology, Uppsala University, Sweden, ali.dorostkar@it.uu.se, maya.neytcheva@it.uu.se

†Dept. of Science and high Technology, Insubria University, Italy and Dept. of Information Technology, Uppsala University, Sweden, stefano.serrac@uninsubria.it, stefano.serra@it.uu.se

1.1 Notations

Although the notations in the paper are used in their broadly accepted conventional meaning, we include some definitions for clarity and self-consistency of the paper.

Definition 1.1. [Toeplitz matrix] *A finite-dimensional Toeplitz matrix is a square matrix that has constant elements along each descending diagonal from left to right, namely,*

$$T = \begin{bmatrix} a_0 & a_{-1} & a_{-2} & \cdots \\ a_1 & a_0 & a_{-1} & \cdots \\ a_2 & a_1 & \ddots & \ddots \\ \ddots & \ddots & \ddots & \ddots \end{bmatrix}. \tag{1}$$

As (1) indicates, the matrix T can be also infinitely dimensional. A finite-dimensional Toeplitz matrix of dimension n is denoted as T_n,

$$T_n = \begin{bmatrix} a_0 & a_{-1} & a_{-2} & \cdots & \cdots & a_{-n+1} \\ a_1 & a_0 & a_{-1} & \ddots & & \vdots \\ a_2 & a_1 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & a_{-1} & a_{-2} \\ \vdots & & \ddots & a_1 & a_0 & a_{-1} \\ a_{n-1} & \cdots & \cdots & a_2 & a_1 & a_0 \end{bmatrix}. \tag{2}$$

In the sequel we consider also sequences of Toeplitz matrices as a function of their dimension, denoted as $\{T_n\}$, $n = 1, 2, \cdots, \infty$.

Definition 1.2. [Block Toeplitz matrix (multilevel)] *A matrix X_N is a block Toeplitz matrix if it is a Toeplitz matrix with elements that are square blocks of relevant sizes,*

$$X_N = T_n = \begin{bmatrix} A_0 & A_{-1} & A_{-2} & \cdots & \cdots & A_{-n+1} \\ A_1 & A_0 & A_{-1} & \ddots & & \vdots \\ A_2 & A_1 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & A_{-1} & A_{-2} \\ \vdots & & \ddots & A_1 & A_0 & A_{-1} \\ A_{n-1} & \cdots & \cdots & A_2 & A_1 & A_0 \end{bmatrix}, \tag{3}$$

where $A_i, i \in \{-n+1, \cdots, n-1\}$ are square blocks of size k and $N = n \times k$. Note that, the subindex of T_n is the number of blocks in the Toeplitz matrix while the subindex of X_N is the size of the matrix. A special case of block-Toeplitz matrices is the class of two- and multilevel block Toeplitz matrices, where the blocks are Toeplitz (or multilevel Toeplitz) matrices themselves. The standard Toeplitz matrices are sometimes addressed as unilevel Toeplitz.

Definition 1.3. [Toeplitz sequences (generating function of)] *Denote by $f(\theta_1, \cdots, \theta_d)$ a d-variate complex-valued integrable function, defined over the domain $Q^d = [-\pi, \pi]^d, d \geq 1$. Denote by f_k the Fourier coefficients of f,*

$$f_k = \frac{1}{m\{Q^d\}} \int_{Q^d} f(\theta) e^{-i\,(k,\theta)}\, d\theta,\ k = (k_1, \cdots, k_d) \in \mathbb{Z}^d,\ i^2 = -1,$$

where $(k, \theta) = \sum_{j=1}^{d} k_j \theta_j$, $n = (n_1, \cdots, n_d)$, and $N(n) = n_1 \cdots n_d$. By following the multi-index notation in [67][Section 6], with each f we can associate a sequence of Toeplitz matrices $\{T_n\}$, where

$$T_n = \{f_{k-\ell}\}_{k,\ell=\mathbf{e}^T}^{n} \in \mathbb{C}^{N(n)\times N(n)},$$

$\mathbf{e} = [1, 1, \cdots, 1] \in \mathbb{N}^d$.

The function f is referred to as the generating function (or the symbol of) T_n. Using a more compact notation, we say that the function f is the generating function of the whole sequence $\{T_n\}$ and we write $T_n = T_n(f)$.

If $f(\theta_1, \cdots, \theta_d)$ is d-variate, $\mathbb{C}^{s\times t}$ matrix- valued, and integrable over Q^d, $d, s, t \geq 1$, then we can define the Fourier coefficients of f in the same way (now f_k is a matrix of size $s \times t$) and consequently $T_n = \{f_{k-\ell}\}_{k,\ell=\mathbf{e}^T}^{n} \in \mathbb{C}^{sN(n)\times tN(n)}$. If $s = t$ then T_n is a d-level block Toeplitz matrix according to Definition 1.2. As in the scalar case, the function f is referred to as the generating function (or the symbol of) T_n, we say that the function f is the generating function of the whole sequence $\{T_n\}$, and we write $T_n = T_n(f)$.

Remark 1.4. [Toeplitz matrices and multiple generating functions] *Let n be even and consider the function $f(\theta) = 2 - 2\cos(\theta)$. According to Definition 1.3, $X_n = T_n(f)$ is of the form so that*

$$X_n = T_n(f) = \begin{bmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & -1 & \ddots & \ddots & \\ & & \ddots & \ddots & -1 \\ & & & -1 & 2 \end{bmatrix}.$$

However, we may look at the same matrix X_n as a block Toeplitz matrix with blocks of size two. According to equation (3) in Definition 1.2, we have

$$X_n = T_{\frac{n}{2}}(f^{[2]}) = \begin{bmatrix} A_0 & A_{-1} & \mathbf{0} & \cdots & \cdots & \mathbf{0} \\ A_1 & A_0 & A_{-1} & \ddots & & \vdots \\ \mathbf{0} & A_1 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & A_{-1} & \mathbf{0} \\ \vdots & & \ddots & A_1 & A_0 & A_{-1} \\ \mathbf{0} & \cdots & \cdots & \mathbf{0} & A_1 & A_0 \end{bmatrix},$$

where

$$A_0 = \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix}, \quad A_1 = A_{-1}^T = \begin{bmatrix} 0 & -1 \\ 0 & 0 \end{bmatrix}.$$

Thus, $X_n = T_{\frac{n}{2}}\left(A_0 + A_1 e^{i\ \theta} + A_1^T e^{-i\ \theta}\right)$. *The new function* $A_0 + A_1 e^{i\ \theta} + A_1^T e^{-i\ \theta}$ *is denoted as* $f^{[2]}$, *to remind that it is defined from* f *via the above procedure. Of course, if* $k \geq 3$ *is fixed and* n *is a multiple of* k, *then we can define in a similar way the function* $f^{[k]}$. *In our specific context we have*

$$f^{[k]}(\theta) = T_k(f) - \mathbf{e}_1\mathbf{e}_k^T e^{i\ \theta} - \mathbf{e}_k\mathbf{e}_1^T e^{-i\ \theta}$$

with $\mathbf{e}_j$, $j = 1, \ldots, k$, *being the canonical basis of* $\mathbb{C}^k$, *so that* $X_n = T_{\frac{n}{k}}(f^{[k]})$.

It is evident that analogous multiple representation holds for every function f.

Definition 1.5. [Circulant, Block circulant matrix, Strang preconditioner (of a Toeplitz matrix)] *A circulant matrix of size* n *is a special Toeplitz matrix* C_n *defined as in (2) where* $(C_n)_{j,k} = b_{(j-k) \bmod n}$, *with given complex coefficients* $b_0, \ldots, b_{n-1}$.

A block circulant matrix is a circulant matrix with elements that are square blocks of relevant sizes,

$$C_n = \begin{bmatrix} B_0 & B_{n-1} & B_{n-2} & \cdots & \cdots & B_1 \\ B_1 & B_0 & B_{n-1} & \ddots & & \vdots \\ B_2 & B_1 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & B_{n-1} & B_{n-2} \\ \vdots & & \ddots & B_1 & B_0 & B_{n-1} \\ B_{n-1} & \cdots & \cdots & B_2 & B_1 & B_0 \end{bmatrix}. \tag{4}$$

A special case of block-circulant matrices is the class of two- and multilevel block circulant matrices, where the blocks are circulant (or multilevel circulant) matrices themselves. The standard circulant matrices are sometimes addressed as unilevel circulant. See [19] for more details on the subject.

Given a unilevel either scalar or block Toeplitz matrix T_n, *according either to (2) or (3), its circulant Strang preconditioner* $C(T_n)$ *is defined as (4) where*

$$B_j = \begin{cases} A_j & \text{if } j \leq \lceil \frac{n}{2} \rceil, \\ A_{j-1-n} & \text{if } j > \lceil \frac{n}{2} \rceil. \end{cases} \tag{5}$$

If T_n *is multilevel then* $C(T_n)$, *its circulant Strang preconditioner [61], is defined as (4) where*

$$B_j = \begin{cases} C(A_j) & \text{if } j \leq \lceil \frac{n}{2} \rceil, \\ C(A_{j-1-n}) & \text{if } j > \lceil \frac{n}{2} \rceil, \end{cases} \tag{6}$$

and here $C(A_k)$ *is recursively defined according to the previous rules in (5)-(6), see also [67][Section 6].*

In the sequel, when mentioning asymptotic spectral properties, we refer to the asymptotics of the extremal eigenvalues and to global properties as existence of the joint asymptotic distribution of $\{T_n\}$ in the Weyl sense (see [14, 56] and the references therein). We recall that for a sequence of matrices $\{X_n\}$, X_n of increasing order n and with spectrum $\Lambda_n \subset \mathbb{C}$, the measure σ is said to describe a joint asymptotic spectrum if for all functions $F \in \mathcal{C}_0(\mathbb{C})$, that is, the subset of functions from $\mathcal{C}(\mathbb{C})$ having bounded support, there holds

$$\lim_{n\to\infty} \frac{1}{n} \sum_{\lambda\in\Lambda_n} F(\lambda) = \int F(\lambda)\, d\sigma(\lambda), \tag{7}$$

where each eigenvalue is counted according to its multiplicity. Hence, σ is a probability measure supported on the extended complex plane $\overline{\mathbb{C}} = \mathbb{C} \cup \{\infty\}$. It should be observed that often (7) is written in a more informative way as

$$\lim_{n\to\infty} \frac{1}{n} \sum_{\lambda\in\Lambda_n} F(\lambda) = \frac{1}{s} \int_D \text{trace}\,(F(f(\theta)))\, \frac{d\theta}{m(D)}, \tag{8}$$

where f is a $s \times s$ matrix-valued function, trace$(\cdot)$ denotes the trace of its argument (that is the sum of all its eigenvalues), D is a domain with a finite positive Lebesgue measure in $\mathbb{R}^d$, $d \geq 1$, and $m(\cdot)$ is the standard Lebesgue measure (see [46]). In that case f is said to be the symbol of $\{X_n\}$ and we write $\{X_n\} \sim_\lambda f$.

Similarly, we write $\{X_n\} \sim_\sigma f$ if for all functions $F \in \mathcal{C}_0(\mathbb{R}_0^+)$, that is, the subset of functions from $\mathcal{C}(\mathbb{R}_0^+)$ having bounded support, there holds

$$\lim_{n\to\infty} \frac{1}{n} \sum_{\sigma\in\Sigma_n} F(\sigma) = \frac{1}{s} \int_D \text{trace}\left(F\left[(f^*(\theta)f(\theta))^{\frac{1}{2}}\right]\right) \frac{d\theta}{m(D)}, \tag{9}$$

with Σ_n being the set of the singular values of X_n and f as in (8).

Remark 1.6. *The informal interpretation of formula (8) is rather insightful. When $\{X_n\}_n \sim_\lambda f$ and f is a scalar-valued and smooth enough function, it means that, for n large enough, the eigenvalues of X_n are given by equispaced evaluations of the function f on its domain of definition, up to infinitesimal errors and possibly up to a few outliers (at most $o(n)$ of them, but often much fewer). When the symbol f is $s \times s$ matrix-valued and smooth enough, denote by $\lambda_1(f), \lambda_2(f), \ldots, \lambda_s(f)$ its s eigenvalues, each of them being a scalar-valued function. Let n be the size of X_n, n being an increasing sequence. Then the connection between the eigenvalues of X_n and f is as follows. For large enough n, n/s eigenvalues of X_n are described by equispaced evaluations of $\lambda_1(f)$, n/s eigenvalues of X_n are described by equispaced evaluations of $\lambda_2(f)$, and so on until the last n/s eigenvalues of X_n, which are described by equispaced evaluations of the last eigenvalue function $\lambda_s(f)$. Clearly, as in the case of $s = 1$, it is possible that infinitesimal errors and $o(n)$ outliers of the spectrum of X_n remain uncaptured by the symbol.*

The informal interpretation of formula (9) is exactly the same of formula (8), but with singular values in place of eigenvalues.

Remark 1.7. *In the (block) multilevel Toeplitz case (cf. [67][Section 6] and [64]), the symbol of $\{T_n(f)\}$ in the sense of (9) coincides with the (matrix-valued) multivariate generating function f, according to Definition 1.3, whose Fourier coefficients determine the entries of any matrix $T_n(f)$ (see [34, 66, 64]): see also [50, 64, 55, 63, 56, 57] for more general and more advanced results.*

If f is Hermitian-valued (real-valued in the scalar case), then the symbol of $\{T_n(f)\}$ in the sense of (8) coincides with the Hermitian-valued (real-valued in the scalar case) multivariate generating function f, according to Definition 1.3.

Concerning Remark 1.6 in the context of Toeplitz sequences $\{T_n(f)\}$, the equispaced evaluations of the symbol are made on a grid that discretizes $Q = (-\pi, \pi)$ as $\{x_j^{(n)}\}$, $x_j^{(n)} = -\pi + \frac{2\pi j}{n}$, $j = 1, \ldots, n$. If in addition the symbol is also even then a better choice of a sampling grid is given by $\{x_j^{(n)}\}$, with $x_j^{(n)} = \frac{\pi j}{n+1}$, $j = 1, \ldots, n$.

Remark 1.8. *With reference to Remark 1.6, in the Hermitian Toeplitz case, when considering a smooth scalar-valued symbol, it is worth noting that no outliers show up (see e.g. [34, 14]). When the symbol f is $s \times s$ matrix-valued, d-variate and smooth, then a bound (although, often pessimistic) of the number of outliers is given by $c(s)n^{\frac{d-1}{d}}$, where $c(s)$ is a constant growing linearly with s. The latter statement is easy to see by using the decay of the Fourier coefficients according to the Riemann-Lebesgue lemma [46] and the interlacing theorem [13, 36], when writing $T_n(f)$ in the form of a Strang preconditioner plus an error term (see Definition 1.5 and [61]) in the block multilevel circulant algebra (see e.g [31][Section 2.5]).*

When the generating function is either complex-valued or matrix-valued, then, as already observed, the resulting matrices are either multilevel Toeplitz or block multilevel Toeplitz, but they are in general non-Hermitian. In that case, formula (8) is known to hold for the singular values (see e.g. [66, 64]) with $|f| = (f^*f)^{1/2}$ instead of f, see (9), and for the eigenvalues again with f, but only in some special cases [65]. For more details we refer to the fundamental paper by Tilli [65] and to [24] for the extension of Tilli's results to the matrix-valued setting. A similar technique applied to the preconditioning can be found in [23].

1.2 Toeplitz matrices in the context of discrete partial differential equations

Consider a differential boundary value problem of the general form

$$\mathcal{L}u = f \quad \text{on } \Omega, \tag{10}$$

where $\mathcal{L}$ is a given differential operator, $\Omega \subset \mathbb{R}^d$, $d \geq 1$ is some open, bounded, connected domain, and assume that the equation is equipped with proper boundary conditions on the domain boundary $\partial\Omega$.

When discretizing this problem for a sequence of discretization parameters h_n we obtain a corresponding sequence of matrices $\{A_n\}$ of size n that grows to infinity as the approximation error tends to zero. The approximate solution to (10) is a vector $\mathbf{u}_n$ solving a system of the type $A_n\mathbf{u}_n = \mathbf{b}_n$ with some $\mathbf{b}_n$ related to the problem data, the discretization parameter, the boundary conditions, the chosen method etc. The more precise the approximation is, the larger the related size and, therefore, the more difficult the solution of the resulting linear system becomes. The difficulties of that task are largely influenced by the structure and by the spectral properties of the underlying matrices. In particular, the study of the spectrum of A_n for fixed dimension and its behavior in an asymptotic sense is often a prerequisite for designing efficient solvers and preconditioners. As an example, in the case of symmetric positive definite (spd) matrices both the asymptotic distribution (cf. [10]) and the conditioning of the finite dimensional matrices (cf. [5, 68]) play a role for the convergence speed of the conjugate gradient (CG) method. Similar results are shown for the convergence of the GMRES method for general normal matrices. In the latter case, the estimates include also the conditioning of the corresponding eigenvector matrices (see e.g [48] for both CG and GMRES methods).

An appropriate knowledge of the spectrum is needed also for establishing stability conditions for time-dependent problems, discretized in space using the above mentioned discretization schemes.

In this work, for simplicity of the presentation, we consider constant coefficient partial differential equations (PDEs), square domains and uniform grids, and explicitly mention how the derived results generalize to variable coefficients, domains of arbitrary shape and nonequidistand discretization meshes.

The techniques to approximate PDEs by local methods such as the Finite Difference method (FDM) (cf. [62, 35, 3], the Finite Element method (FEM) (cf. [29, 16, 37]) or via the Isogeometric Analysis (IgA) (cf. [17]) lead to sequences of matrices that admit a Toeplitz structure. For instance, in the FDM and IgA settings, if the domain is d-dimensional, then the generating function f is a real-valued trigonometric polynomial in d variables, which are the corresponding Fourier variables, and its degree is related to the accuracy of the used approximation formulas, see [58, 30].

It has been recently shown that, compared to FDM and IgA, FEM discretizations exhibit somewhat different properties. Namely, even for scalar PDEs, discretized by higher order FEM, the generating function is $s \times s$ matrix-valued (so we are in a block context) and s grows exponentially with the degree p of the polynomial basis functions, defined on the local elements (triangle, quadrilateral etc.), and with the dimensionality d, e.g., when considering quadrilateral Finite Elements of degree p, we have $s = p^d$ [31]. We encounter a particular example of such a situation with $s = 4$ induced by $p = d = 2$, when dealing with Taylor-Hood elements in Section 4.3. In view of Remark 1.6, this entails an exponential scattering of the spectrum, i.e., we have $s = p^d$ different functions, $\lambda_1(f), \lambda_2(f), \ldots, \lambda_{p^d}(f)$, that describe the spectrum of the related large matrices, refrains from using p much larger than 2 even in low dimensionality d. In the Engineering community this phenomenon is known as the appearance of branches in the spectrum, that are not related to the true spectrum of the

continuous operator, but are artificially introduced by the chosen approximation method. We note that such a behavior is not present when using either the FDM or the IgA approach or FEM with minimal degree of the basis functions $p = 0, 1$, even though the case $p = 0$ has to be treated differently.

As discussed above, the insight, obtained when applying the theory of block multilevel Toeplitz matrix sequences could give additional and deeper understanding of the spectral behavior of large matrices, arising from the approximation of PDEs by local methods. However, the limitations on square domains, uniform gridding or triangulations, and constant coefficients are quite strong. A substantial step for overcoming these limiting factors, has been done by Tilli [63] and by the third author [56, 57]. In [63, 56, 57] the notion of 'locality' in the Toeplitz context has been introduced, leading to the definition of Generalized Locally Toeplitz (GLT) sequences. There are five main features of the GLT class of matrices that are also of high relevance to the problems, considered in this paper.

GLT1 Each GLT sequence has a symbol f. If the sequence is Hermitian then (8) holds. Otherwise, as specified in (9), the same formula also holds true with the singular values in place of the eigenvalues and $|f| = (f^*f)^{1/2}$ in place of f.

GLT2 The set of GLT sequences form a $*$-algebra that is close under linear combinations, conjugation, products, inversion (whenever the symbol vanishes, at most, in a set of zero Lebesgue measure). Hence, the sequence obtained via algebraic operations on a finite set of input GLT sequences is still a GLT sequence and its symbol is obtained by the same algebraic manipulations on the corresponding symbols of the input GLT sequences.

GLT3 Every Toeplitz sequence generated by a L^1 function f is a GLT sequence and its symbol is f, possessing the properties from **GLT1**. Every sequence of matrices $\{X_n\}$ which is distributed as the zero function in the singular value sense, i.e. $\{X_n\} \sim_\sigma 0$ according to (9), is a GLT sequence with symbol $f \equiv 0$.

GLT4 The approximation of PDEs with non-constant coefficients, general domains, nonuniform gridding by local methods (FDM, FEM, IgA etc), under very mild assumptions leads also to GLT sequences (see [63, 56, 57] for the case of FDM, [11, 31] for the FEM setting, and [22, 30] for the case of IgA approximations).

GLT5 We encounter GLT structures for certain matrix sequences, related to preconditioners, based on approximations of PDEs by local methods. Moreover, the symbol includes information about the coefficients and the domain of the PDE, as well as information on the discretization schemes for the derivatives including the used meshes, which have to be described, at least asymptotically, as a map of a reference equispaced mesh (see [44, 63, 58, 51] for the one-dimensional setting and [53, 52, 59, 56, 57, 11, 20, 21, 31] for the two-dimensional and multi-dimensional settings. Furthermore, also in presence of non-dominating advection terms the distribution result for the eigenvalues can be recovered, thanks to ad hoc results in [33, 32], heavily based on the majorization theory well explained in the remarkable book [13].

In this paper we focus on the case of coupled PDEs, more precisely, on the linear elasticity problem in saddle point form. To the best of our knowledge, this is the first time when the Toeplitz and GLT machinery is applied to the case of coupled or vector PDEs, even though the idea is mentioned in [57][Section 3.3].

Our aim with this work is two-fold.

1. Show how to use the Toeplitz and GLT technology for deriving spectral information on the approximating matrices and on the related Schur complement that is widely used when constructing preconditioners for discrete differential operators of two-by-two block form, including saddle point matrices.

2. Use the spectral properties of the corresponding symbol to analyze known numerical preconditioning methods and to suggest new preconditioners. Given the complexity of the first item, this part is a subject of future investigations.

The paper is organized as follows. In Section 2 we present two hypothetical PDE problems in order to exemplify the idea of the methodology and the aim we want to achieve for the linear elasticity problem in saddle point form. Section 3 describes the state-of- the-art preconditioning techniques for two-by-two block matrices of saddle point form, the role of the Schur complement, including the substantial difficulties encountered in the non-symmetric case when analyzing the quality of a particular Schur complement approximation. Section 4 represents the central part of the paper, where the GLT machinery is used to study, at least asymptotically, the spectrum of the matrices arising from the target elasticity problem and that of the related Schur complement. Numerical results support the theoretical findings. Final remarks and open issues to be addressed in a future work are given in Section 5.

2 Basic examples of symbols in a PDE context

In this section we introduce the GLT technique for deriving spectral properties of the related matrices and of the corresponding Schur complement with the help of two simplified examples, the first in one dimension, discretized using FEM, the second in two dimensions, discretized using FDM. Both examples are in a vector form in order to indicate how to handle the elasticity problem, also in mixed form and using stable finite element approximation spaces.

Problem 2.1. *Consider the coupled system of scalar equations*

$$\begin{cases} -(\kappa(x)u')' + v' &= g_1(x), \\ u' - \rho v &= g_2(x), \end{cases} \tag{11}$$

with Dirichlet boundary conditions. Here $\rho > 0$ and the function $\kappa(x)$ is positive and continuous on the domain $\Omega = [0, 1]$.

Assume first that $\kappa(x) = \kappa_0$ is constant in Ω. The use of linear FEM basis functions on a uniform mesh with a step size h and a proper scaling leads to a linear system of equations with a coefficient matrix that admits the following structure

$$\mathcal{A} = \begin{bmatrix} K & B^T \\ B & -\rho M \end{bmatrix}.$$

Here, the blocks K, B, M are square of size n and are depicted in (12) and (13).

$$K = \kappa_0 \cdot \begin{bmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & -1 & \ddots & \ddots & \\ & & \ddots & \ddots & -1 \\ & & & -1 & 2 \end{bmatrix}, \quad M = \frac{h^2}{6} \begin{bmatrix} 4 & 1 & & & \\ 1 & 4 & 1 & & \\ & 1 & \ddots & \ddots & \\ & & \ddots & \ddots & 1 \\ & & & 1 & 4 \end{bmatrix}, \tag{12}$$

$$B = h \begin{bmatrix} 1 & & & \\ -1 & 1 & & \\ & -1 & \ddots & \\ & & \ddots & \ddots \\ & & & -1 \quad 1 \end{bmatrix}. \tag{13}$$

Clearly, all the blocks have a Toeplitz character, according to Definition 1.1. Following Definition 1.3 in the specific context of Problem 2.1, we find the corresponding representations of the matrix blocks as Toeplitz matrices and their generating symbol:

$$\begin{array}{rclcrcl} K & = & \kappa_0\, T_n(2 - 2\cos(\theta)), & & B & = & h\, T_n(1 - \mathrm{e}^{\mathrm{i}\theta}), \\ B^T & = & h\, T_n(1 - \mathrm{e}^{-\mathrm{i}\theta}), & & M & = & \frac{h^2}{3}\, T_n(2 + \cos(\theta)). \end{array} \tag{14}$$

The negative Schur complement of $\mathcal{A}$ is defined as follows

$$\mathcal{S} = \rho M + B^T K^{-1} B. \tag{15}$$

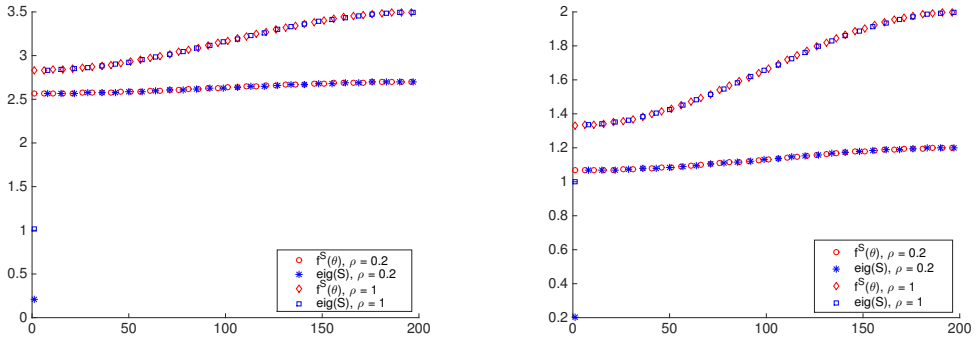
We note that as $\kappa > 0$, the matrix K is invertible.

Considering (15) in terms of Toeplitz structures, we are in a position to construct the symbol of $\mathcal{S}$, $f^{\mathcal{S}}$. First, $\mathcal{S}$ expressed by Toeplitz matrices, reads as

$$\mathcal{S} = \frac{\rho}{3} T_n(2 + \cos(\theta)) + \frac{1}{\kappa_0} T_n(1 - \mathrm{e}^{-\mathrm{i}\theta}) T_n^{-1}(2 - 2\cos(\theta)) T_n(1 - \mathrm{e}^{\mathrm{i}\theta}). \tag{16}$$

According to **GLT3**, the sequences $\{T_n(2 + \cos(\theta))\}$, $\{T_n(1 - \mathrm{e}^{\mathrm{i}\theta})\}$, $\{T_n(1 - \mathrm{e}^{-\mathrm{i}\theta})\}$ and $\{T_n(2 - 2\cos(\theta))\}$ are GLT sequences with symbols $2 + \cos(\theta)$, $1 - \mathrm{e}^{\mathrm{i}\theta}$, $1 - \mathrm{e}^{-\mathrm{i}\theta}$, $2 - 2\cos(\theta)$, respectively. Furthermore, according to the structure of the $*$-algebra in **GLT2**, $\{\mathcal{S}_n\}$ is a GLT sequence generated by the symbol

$$f^{\mathcal{S}}(\theta) = \frac{\rho}{3}(2 + \cos(\theta)) + \frac{1}{\kappa_0}(1 - \mathrm{e}^{-\mathrm{i}\theta}) \frac{1}{2 - 2\cos(\theta)} (1 - \mathrm{e}^{\mathrm{i}\theta}) = \frac{\rho}{3}(2 + \cos(\theta)) + \frac{1}{\kappa_0}. \tag{17}$$



(a) $\kappa_0 = 0.4$ (b) $\kappa_0 = 1$

Figure 1: Problem 2.1: Spectrum of $\mathcal{S}$ vs sampling of its symbol $f^{\mathcal{S}}$, constant coefficient κ_0

Since $\mathcal{S}_n$ is Hermitian independently of its size, according to **GLT1**, we deduce that (8) holds for $\{\mathcal{S}\}$, i.e., $\{\mathcal{S}\} \sim_\lambda f^{\mathcal{S}}$, where $f^{\mathcal{S}}$ is an even trigonometric polynomial. Figure 1 shows the agreement between the asymptotic forecast and the eigenvalues of $\mathcal{S}$ for a couple of values of the parameters κ_0 and ρ, where both the evaluations of $f^{\mathcal{S}}$ over $x_j^{(n)} = \frac{\pi j}{n+1}$, $j = 1, \ldots, n$. In the plots, the eigenvalues of $\mathcal{S}$ are sorted in an increasing order.

Next we consider equation (11) with a non-constant diffusivity coefficient $\kappa(x)$. The resulting matrix K is no longer Toeplitz but, as shown in [11], the related sequence $\{K\}$ belongs to the GLT class with symbol $\kappa(x)(2 - 2\cos(\theta))$. Therefore, following verbatim the reasoning for the case of $\kappa(x) = \kappa_0$, we deduce that the sequence of Schur complements forms a GLT sequence with the symbol

$$\begin{aligned} f^{\mathcal{S}}(x,\theta) &= \frac{\rho}{3}(2+\cos(\theta)) + \frac{1}{\kappa(x)}(1-\mathrm{e}^{-\mathrm{i}\theta})\frac{1}{\kappa(x)(2-2\cos(\theta))}(1-\mathrm{e}^{\mathrm{i}\theta}) \\ &= \frac{\rho}{3}(2+\cos(\theta)) + \frac{1}{\kappa(x)}. \end{aligned} \tag{18}$$

The latter is illustrated in Figure 2, showing the agreement between the asymptotic forecast and the eigenvalues of $\mathcal{S}$ for a pair of choices of ρ and $\kappa(x)$, where both the evaluations of $f^{\mathcal{S}}$ over a n-sized uniform gridding over $[0,1] \times [0,\pi]$ of size n and the eigenvalues of $\mathcal{S}$ have been sorted in an increasing order.

We formulate next a two-dimensional coupled test system of two PDEs and two unknowns, the first of which, referred to as the *displacements* is a vector with two components.

Problem 2.2. *Consider an elasticity-like problem in saddle point form, defined in* $\Omega = [0,1]^2$,

$$\mathcal{A} = \begin{bmatrix} K & B^T \\ B & -\rho M \end{bmatrix} \begin{matrix} \}\textit{displacements} \\ \}\textit{pressure}, \end{matrix}$$

where K and M are symmetric and positive definite.

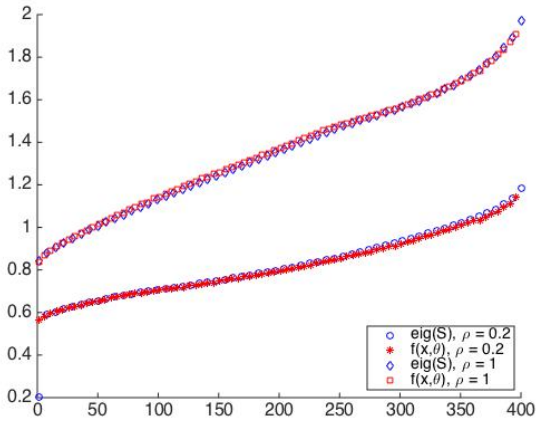


Figure 2: Problem 2.1: Spectrum of $\mathcal{S}$ vs sampling of its symbol $f^{\mathcal{S}}$, variable coefficient $\kappa_0(x) = 1 + x$.

Under the so-called separate displacement ordering (SDO) of the components of the displacements, K itself attains a two-by-two block structure,

$$K = \begin{bmatrix} K_{11} & K_{12} \\ K_{21} & K_{22} \end{bmatrix} \begin{matrix} \} \textit{displacements in } x \\ \} \textit{displacements in } y. \end{matrix}$$

The SDO ordering of the displacements induces a corresponding block structure in the block B, which we denote as $B = \begin{bmatrix} B_1 & B_2 \end{bmatrix}$. Thus, we have

$$\mathcal{A} = \begin{bmatrix} K_{11} & K_{12} & B_1^T \\ K_{21} & K_{22} & B_2^T \\ B_1 & B_2 & -\rho M \end{bmatrix}. \tag{19}$$

Further, we assume that, respectively, K_{11}, K_{22} approximate two anisotropic Laplacians of the form $-\left(2\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}\right)$, $-\left(\frac{\partial^2}{\partial x^2} + 2\frac{\partial^2}{\partial y^2}\right)$, K_{12}, K_{21} approximate the operators $-\frac{\partial^2}{\partial y \partial x}$, $-\frac{\partial^2}{\partial x \partial y}$ and B_1^T, B_2^T approximate the operators $\frac{\partial}{\partial x}$, $\frac{\partial}{\partial y}$. M is the mass matrix, approximating the identity operator.

To convey the idea in the simplest possible way, we consider standard FDM with a square mesh, even though this approximation does not possess the stability properties, required for

mixed problems. The blocks of $\mathcal{A}$ are described as follows:

$$\begin{aligned}
K_{11} &= 2\,T_n(2-2\cos(\theta_1))\otimes I_n + I_n\otimes T_n(2-2\cos(\theta_2)),\\
K_{12} &= T_n(1-\mathrm{e}^{-\mathrm{i}\theta_1})\otimes T_n(1-\mathrm{e}^{-\mathrm{i}\theta_2}),\\
K_{21} &= K_{12}^T,\\
K_{22} &= T_n(2-2\cos(\theta_1))\otimes I_n + 2I_n\otimes T_n(2-2\cos(\theta_2)),\\
B_1 &= h\,T_n(1-\mathrm{e}^{-\mathrm{i}\theta_1})\otimes I_n,\\
B_2 &= h\,I_n\otimes T_n(1-\mathrm{e}^{-\mathrm{i}\theta_2}),\\
M &= h^2\,I_n.
\end{aligned}$$

Here $\otimes$ denotes the Kronecker product (cf. [36]). We notice that the Kronecker product induces a two-level Toeplitz structure and in fact every matrix is Toeplitz, where each 'entry' along the diagonal is a standard unilevel Toeplitz matrix, see Definition 1.2. Hence, using the two- level notation introduced in [31, 67], we construct the two- variate generating functions as in Definition 1.3, associated with each block. More precisely,

$$\begin{aligned}
K_{11} &= T_n((6-4\cos(\theta_1)-2\cos(\theta_2))), &(20)\\
K_{12} &= T_n((1-\mathrm{e}^{-\mathrm{i}\theta_1})(1-\mathrm{e}^{-\mathrm{i}\theta_2})), &(21)\\
K_{21} &= T_n((1-\mathrm{e}^{\mathrm{i}\theta_1})(1-\mathrm{e}^{\mathrm{i}\theta_2})), &(22)\\
K_{22} &= T_n((6-2\cos(\theta_1)-4\cos(\theta_2))), &(23)\\
B_1 &= h\,T_n(1-\mathrm{e}^{-\mathrm{i}\theta_1}), &(24)\\
B_2 &= h\,T_n(1-\mathrm{e}^{-\mathrm{i}\theta_2}), &(25)\\
M &= h^2\,T_n(1). &(26)
\end{aligned}$$

Consider next the negative Schur complement of $\mathcal{A}$, $S=\rho M+BK^{-1}B^T$. Given the rich block structure of the matrix $\mathcal{A}$, the formal expression of the Schur complement involves inversion of the spd block K and multiplication by rectangular blocks. Since we want to use the symbols (20)–(26) of the related sub-blocks, we utilize the following exact block-factorization of K and of its inverse:

$$K=\begin{bmatrix}K_{11} & K_{12}\\ K_{21} & K_{22}\end{bmatrix}=\begin{bmatrix}I & \\ K_{21}K_{11}^{-1} & I\end{bmatrix}\begin{bmatrix}K_{11} & \\ & S_K\end{bmatrix}\begin{bmatrix}I & K_{11}^{-1}K_{12}\\ & I\end{bmatrix}, \tag{27}$$

where $S_K=K_{22}-K_{21}K_{11}^{-1}K_{12}$. Since the positive definite character of S_K is guaranteed by the positive definite character of K, we find

$$K^{-1}=\begin{bmatrix}I & -K_{11}^{-1}K_{12}\\ & I\end{bmatrix}\begin{bmatrix}K_{11}^{-1} & \\ & S_K^{-1}\end{bmatrix}\begin{bmatrix}I & \\ -K_{21}K_{11}^{-1} & I\end{bmatrix}. \tag{28}$$

Clearly, the latter factorization holds for any nonsingular matrix K. Therefore, as it is well known, the explicit formal expression of $\mathcal{S}$ is given by

$$\begin{aligned}
\mathcal{S} &= \rho M+\begin{bmatrix}B_1 & B_2\end{bmatrix}K^{-1}\begin{bmatrix}B_1^T\\ B_2^T\end{bmatrix}\\
&= \rho M+B_1K_{11}^{-1}B_1^T+(B_2-B_1K_{11}^{-1}K_{12})S_K^{-1}(B_2-B_1K_{11}^{-1}K_{12})^T. &(29)
\end{aligned}$$

As expected, constructing the symbol for Problem 2.2 is by far more complicated than that in Problem 2.1. Nevertheless, the expression of $\mathcal{S}$ in (29) can be seen as a 'rational noncommutative' formula involving the blocks in (20)–(26), which are two-level Toeplitz structures with trigonometric polynomial symbols. As a consequence, in view of **GLT1, GLT2, GLT3**, exactly as done in the one-dimensional example (11), we deduce that $\{\mathcal{S}\}$ is a GLT sequence with $\{\mathcal{S}\} \sim_\lambda f^{\mathcal{S}}$, since $\mathcal{S}$ is Hermitian independently of its size, and $f^{\mathcal{S}}$ is a rational function of the symbols given in (20)–(26). Interestingly enough, setting $\delta(\theta) = 2 - 2\cos(\theta)$ the symbol of the unilevel Laplacian and making tedious algebraic manipulations, we find

$$f^{\mathcal{S}} = \rho + \mu \frac{q_1(\delta(\theta_1), \delta(\theta_2))}{q_2(\delta(\theta_1), \delta(\theta_2))},$$

where q_1, q_2 are two nonnegative homogeneous polynomials of degree two. Therefore, $\rho \leq f(\theta_1, \theta_2) \leq \rho + \phi$ with ϕ being the maximum of q_1/q_2.

Following the same procedure as in the one-dimensional setting we can treat the case of variable coefficients. However, we do not elaborate more on that, since the considered discretization is not suited for approximating the elasticity equation in a stable way.

Instead, based on the structural analysis reported in (28)–(29) and on the properties of the GLT sequences, in Section 4 we study in depth the symbol of the Schur complement, in the framework of stable approximations of the mixed variable linear elasticity problem, cf., e.g. [8].

3 Solving linear elasticity in saddle point form: background and open questions

In this section we first describe the target problem and its discrete formulation and then we briefly discuss related solution methods with special attention to preconditioning.

3.1 Target problem and discrete formulation

To motivate the present study and to formulate the open questions we address using the GLT theory, we consider one particular target application, arising in Geophysics. We consider the so-called Glacial Isostatic Adjustment (GIA) model, used to describe the response of the solid Earth to redistribution of mass due to alternating glaciation and deglaciation periods. The processes that cause submerge or uplift of the Earth surface are active today and studying and obtaining a better insight on them steadily attracts attention. The model gives raise to very large algebraic systems to be solved, thus imposing strong requirements on the efficiency of the numerical solution methods to be used and also on the suitability of those methods for high performance computations.

Here we consider only the purely elastic response of the Earth, which is a building block of a more realistic, but more complex viscoelastic model of the underlying phenomena. The detailed formulation of the GIA problem can be found, for instance, in [69, 40] and references

therein and the references therein. We present the problem in a simplified form, neglecting the effect of self-gravitation and consider a two dimensional model, where the Earth is modeled as flat elastic homogeneous material body, that can be treated as compressible or incompressible. We consider the equilibrium state of the displacement field $\mathbf{u} = \{u_i\}$, $i = 1, \ldots, d$, $d = 2, 3$. The momentum equation for quasi- static perturbations of a homogeneous, elastic continuum in a constant gravity field reads as

$$-\nabla \cdot \sigma - \nabla(\mathbf{u} \cdot \nabla p_0) + (\nabla \cdot \mathbf{u})\nabla p_0 = \mathbf{f} \quad \text{in } \Omega \subset \mathbb{R}^d,\ d = 2, 3, \tag{30}$$

with some appropriate boundary conditions.

In (30), σ is the stress, p_0 is the so-called *pre-stress*, $\mathbf{f}$ is a body force. The body is also subject to surface load. The third term on the left hand side of Equation (30) describes the *buoyancy* of the compressed material, that vanishes for purely incompressible materials since $\nabla \cdot \mathbf{u} = 0$.

We note that the properties of the material body are described via material parameters, either chosen to be the Young modulus E and the Poisson ratio ν or the Lamé coefficients μ and λ, that are related to E and ν as $\mu = E/(2(1+\nu))$, $\lambda = 2\mu\nu/(1-2\nu)$.

In order to compensate for excluding self-gravitation effects, we need to model fully incompressible materials, i.e., for which $\nu = 0.5$. Note, however, that for fully incompressible linear elastic the problem is not well-posed, since λ becomes unbounded. It is seen from the definition of the Lamé coefficient λ that when ν approaches 0.5, λ tends to infinity. Thus, when $\nu \to 0.5$, the problem (30) becomes ill-posed, and its discrete analogue becomes extremely ill-conditioned. This is the mathematical formulation of the phenomenon known as *volumetric locking*, which may lead to erroneous results when solving the discretized Equation (30) in the nearly incompressible limit. See, for example, [15], for further details on the locking effect.

A known remedy to the locking problem is to introduce the so-called *kinematic pressure* $p = \frac{\lambda}{\mu}\nabla \cdot \mathbf{u}$, and reformulate (30) as a coupled system of PDEs (cf., i.e.[8]), We also consider a more general form of the first order terms. Below we state the problem in terms of displacements only. The coupled system of equations yields

$$-\nabla \cdot (2\mu\varepsilon(\mathbf{u})) - \nabla(\mathbf{u} \cdot \mathbf{b}) + (\nabla \cdot \mathbf{u})\mathbf{c} - \mu\nabla p = \mathbf{f} \quad \text{in } \Omega, \tag{31a}$$

$$\mu\nabla \cdot \mathbf{u} - \frac{\mu^2}{\lambda}p = 0 \quad \text{in } \Omega, \tag{31b}$$

with $\varepsilon(\mathbf{u}) = \frac{1}{2}\left(\nabla\mathbf{u} + \nabla\mathbf{u}^T\right)$. It is assumed that $\mathbf{b} = \{b_i\}$, $\mathbf{c} = \{c_i\}$, $i = 1, 2$ are some given vectors, for simplicity with constant coefficients. In the above formulation, when λ becomes unbounded, the coefficient $\frac{\mu^2}{\lambda}$ vanishes.

The system (31) is then first formulated in variational terms and discretized with a stable pair of finite element spaces that satisfy the Ladyzhenskaya-Babuška-Brezzi (LBB) stability condition (cf., e.g. [29]). We note that the target geometry of the problem is rectangular and therefore a discretization with a square or a rectangular mesh is the natural choice. We consider below two stable FEM discretizations, the so-called Modified Taylor-Hood elements, denoted by Q1isoQ1, and the Q2Q1 Taylor-Hood elements. As the notation indicates, in

the Q2Q1 we use biquadratic FEM for the displacements and bilinear FEM for the pressure. For the case Q1isoQ1, both the displacements and the pressure are discretized using bilinear basis functions, however the pressure unknowns are discretized on a twice coarser mesh than that used for the displacements. We note that the number of unknowns in both cases is the same.

As the variational setting and discretization of (31) are straightforward, we present directly the algebraic system of equations to be solved,

$$\mathcal{A}\begin{bmatrix}\mathbf{u}_h\\ \mathbf{p}_h\end{bmatrix} = \begin{bmatrix}\mathbf{f}\\ \mathbf{g}\end{bmatrix} \text{ where } \mathcal{A} = \begin{bmatrix}K & B^T\\ B & -\rho M\end{bmatrix}. \tag{32}$$

Here M is the pressure mass matrix. The coefficient $\rho = \frac{\mu^2}{\lambda}$ is different from zero for compressible materials, it is exactly zero for purely incompressible materials and approaches zero in the nearly incompressible case. The block K is symmetric and positive definite when $\mathbf{b} = \mathbf{c} = \mathbf{0}$, otherwise it is nonsymmetric. The blocks B and B^T correspond to discrete divergence and gradient operators, correspondingly. Imposing SDO for the displacement vector, i.e., ordering first the displacements in x-direction and then the displacements in y-direction, we induce a two-by-two block structure of the block K and on B as $B = \begin{bmatrix}B_1 & B_2\end{bmatrix}$, and the system matrix becomes

$$\mathcal{A} = \begin{bmatrix}K_{11} & K_{12} & B_1^T\\ K_{21} & K_{22} & B_2^T\\ B_1 & B_2 & -\rho M\end{bmatrix}. \tag{33}$$

We note that the matrix in (33) has the same structure as that in (19), however, here the blocks B_1 and B_2 are rectangular.

3.2 Solution method and preconditioning

As we aim at large scale numerical simulations, we exclude the direct methods applied to the whole discrete system as infeasible and advocate preconditioned iterative solution methods. To solve systems with $\mathcal{A}$ we consider preconditioned Krylov subspace iterative solution methods for general matrices, such as `GMRES` (cf. [49]) or rather `FGMRES` or `GCG` (cf. [2, 47]), that are suitable for variable preconditioning schemes.

As is well known, the most important issue in this context is the choice of a good preconditioner that combines high numerical efficiency with low arithmetic costs, ideally linearly proportional to the number of degrees of freedom.

Most often, the preconditioners for block matrices and in particular for two-by-two block systems utilize in some way the block structure and are based on the exact block factorization of $\mathcal{A}$,

$$\mathcal{A} = \begin{bmatrix}A_{11} & A_{12}\\ A_{21} & A_{22}\end{bmatrix} = \begin{bmatrix}A_{11} & 0\\ A_{21} & S\end{bmatrix}\begin{bmatrix}I_1 & A_{11}^{-1}A_{12}\\ 0 & I_2\end{bmatrix}. \tag{34}$$

The preconditioner can be of block-multiplicative form (35, left), of block-triangular form, say as in (35, middle), or of block-diagonal form as in (35, right),

$$\mathcal{B}_F = \begin{bmatrix} A_{11} & 0 \\ A_{21} & \widehat{S} \end{bmatrix} \begin{bmatrix} I_1 & Z_{12} \\ 0 & I_2 \end{bmatrix}, \quad \mathcal{B} = \begin{bmatrix} A_{11} & 0 \\ A_{21} & \widehat{S} \end{bmatrix}, \quad \mathcal{B}_D = \begin{bmatrix} A_{11} & 0 \\ 0 & \widehat{S} \end{bmatrix}. \tag{35}$$

Here $\widehat{S}$ is an approximation of the exact Schur complement S of $\mathcal{A}$, $S = A_{22} - A_{21}A_{11}^{-1}A_{12}$, Z_{12} is either exactly equal to or approximates the matrix product $A_{11}^{-1}A_{12}$ and I_1, I_2 are identity matrices of corresponding order.

For more details on the spectral properties of the above preconditioners and recommendations regarding the suitability of one or another, we refer to [6, 12].

It is well-known that the necessary condition for all the above preconditioners to work efficiently is that $\widehat{S}$ has to be a high quality approximation of the Schur complement matrix S. This is not sufficient as we also need to approximate accurately the 11-block. To control this, we use an inner solver for A_{11} and an appropriate stopping criterion.

We focus out attention on the task to construct a high quality sparse approximation of the, in general, dense matrix S.

3.3 Schur complement approximations

Various studies have shown (cf. [9, 39, 42]) that one particular approximation of S, obtainable in the finite element context, is very efficient for the target problem we want to simulate. This is to so-called element-wise Schur complement.

For completeness, we briefly describe the construction of the element-wise Schur complement approximation. To this end we assume that the spatial discretization is done by the FEM method on some mesh with characteristic mesh-size h, denoted by $\mathcal{T}_h = \{\tau_\ell^e\}, \ell = 1, \cdots, L$, where τ_ℓ^e denote the individual elements (triangles, quadrilaterals, bricks etc.) and L is their number.

Clearly, the matrix $\mathcal{A}$ can be assembled block by block. We also observe that it can be assembled based on local matrices that have the same structure as $\mathcal{A}$, namely, $\mathcal{A} = \sum_{\ell=1}^{L} R^{(\ell)^T} A^{(\ell)} R^{(\ell)}$, $\mathcal{A} \in \mathbb{R}^{N\times N}$, $A^{(\ell)} \in \mathbb{R}^{n\times n}$, where

$$\mathcal{A}^{(\ell)} = \begin{bmatrix} A_{11}^{(\ell)} & A_{12}^{(\ell)} \\ A_{21}^{(\ell)} & A_{22}^{(\ell)} \end{bmatrix} \begin{matrix} \}n_1 \\ \}n_2 \end{matrix}. \tag{36}$$

Here $n = n_1 + n_2$, $\ell = 1, \cdots, L$ and L denotes the number of the finite (macro-)elements in the discretization mesh. The matrices $R^{(\ell)} \in \mathbb{R}^{n\times N}$ are the standard Boolean matrices which provide the local-to-global correspondence of the numbering of the degrees of freedom.

Based on (36) we can compute the local Schur complements exactly and assemble those into a global matrix that is then used as an approximation of S,

$$S = \sum_{\ell=1}^{L} R_2^{(\ell)^T} S^{(\ell)} R_2^{(\ell)}, \tag{37}$$

where $S^{(\ell)} = A_{22}^{(\ell)} - A_{21}^{(\ell)} {A_{11}^{(\ell)}}^{-1} A_{12}^{(\ell)}$ and $R_2^{(\ell)}$ are the parts of $R^{(\ell)}$ corresponding to the degrees of freedom in A_{22}. The matrix S in (37) is referred to as the element-wise Schur complement approximation.

Remark 3.1. *Without loss of generality we assume that all $A_{11}^{(\ell)}$ are invertible. If these are singular, we circumvent the problem by adding a diagonal perturbation of order h^2, where h is the characteristic discretization parameter. As this is used only for the construction of the preconditioner, it does not affect the order of the discretization error.*

The construction of $S^{(\ell)}$ can be done fully in parallel across ℓ and the resulting matrix is sparse – properties that comprise two important advantages of the method.

We discuss shortly the notion of a macroelement. The easiest way to illustrate the idea is to consider mesh, referred as coarse, and refine it once in a regular manner. Each element on the coarse element is a macroelement on the fine mesh, where we solve the discrete system. This is the setting where the idea of element-by-element Schur approximation has been first developed for stationary elliptic problems, cf. [38, 4]. This has been further generalized, considering the macroelement to be an agglomerate of several elements of the fine mesh and then allowing to compute $S^{(\ell)}$ as an average of several overlapping agglomerates (macroelements). The latter construction for scalar equations allows for showing that the element-wise Schur complement approximation preserves its high quality in the presence of highly oscillating coefficients, cf. [39].

For coupled systems of equations of the form (32) that are discretized with mixed finite elements, the macroelement is tightly related to the choice of the stable finite element pair of spaces we use. For the Q1isoQ1 case we have two meshes, based on one consecutive regular refinement. For the Q2Q1 case we have only one mesh, however the block structure of $\mathcal{A}^{(\ell)}$ is imposed due to the quadratic basis functions for one of the variables.

Using Linear Algebra tools it is possible to explain the experimentally observed high qualities of the element-wise Schur complement for the case when $\mathcal{A}$ is symmetric and positive definite as well as when it is symmetric indefinite and A_{11} is positive semi-definite (cf. [41, 7, 25]). Those tools and the available results for Schur complements are not applicable for both definite or indefinite matrices. Therefore, to get a better insight in the above, we apply the GLT framework.

4 The symbol of the linear elasticity problem in saddle point form

The matrix $\mathcal{A}$ in (32) can be seen as a generalized block Toeplitz matrix. Following the techniques illustrated in the basic example in Section 2, we next derive the related symbols for the blocks and for the whole matrix. As already stated, this is done for the Q1isoQ1 and Q2Q1 FEM discretizations, namely, for

Q1isoQ1 modified Taylor-Hood elements, where both displacements and pressure are discretized using bilinear basis functions but the pressure unknowns live on twice coarser mesh;

Q2Q1 Taylor-Hood elements with biquadratic and bilinear basis functions for the displacements and for the pressure, correspondingly.

Remark 4.1. [Toeplitz matrices up to low rank] *We note that here we deal with matrices which are Toeplitz up to low rank corrections E_n. In other words, the matrices appearing in all the blocks of $\mathcal{A}$ in (33) can be written as $T_n(f) + E_n$ for some function f and some low rank perturbation E_n. If the matrices are unilevel then rank(E_n) is bounded by a constant independent of n. In the two-level setting, the matrices have size proportional to n^2 and the rank of E_n grows at most proportionally to the square root of the size. By direct checkup, this implies that $\{E_n\} \sim_\sigma 0$ in the sense of (9) and so by* **GLT3** *the sequence $\{E_n\}$ is GLT with symbol identically zero. Therefore by* **GLT2***, the whole sequence $\{T_n(f) + E_n\}$ is a GLT sequence with the same symbol as $\{T_n(f)\}$: hence, again by* **GLT2***, we deduce that the symbol of $\{T_n(f) + E_n\}$ is the generating function of Toeplitz part i.e. the function f.*

As a consequence, in view of Remark 4.1, in the rest of the paper, when dealing with the symbols, we always neglect the low rank corrections.

4.1 The symbol of the mass matrix for bilinear FEM

As the pressure unknown in both cases is discretized using bilinear basis functions, the mass matrix M for Q1isoQ1 and Q2Q1 is the same. For completeness we include the element mass matrix $M^{(e)}$,

$$M^{(e)} = \frac{H^2}{36}\begin{bmatrix} 4 & 2 & 1 & 2 \\ 2 & 4 & 2 & 1 \\ 1 & 2 & 4 & 2 \\ 2 & 1 & 2 & 4 \end{bmatrix}.$$

Here H indicates the discretization parameter of the mesh on which p is discretized, in contrast to the discretization parameter $h = H/2$ for the displacements. As we consider square meshes, the assembled mass matrix can be seen as based on a stencil[1], shown in Figure 3.

We see from the stencil in Figure 3, that the mass matrix M is block-tridiagonal and each block has a tridiagonal structure. The block-symbol of M, $f^M(\theta_1, \theta_2)$ is computed as follows,

$$\begin{array}{rcl} f_0^{M^b}(\theta) & = & 8(2 + \cos(\theta)), \quad f_1^{M^b}(\theta) \; = \; f_{-1}^{M}(\theta) \; = \; 4 + 2\cos(\theta), \\ f^M(\theta_1, \theta_2) & = & 4(2 + \cos(\theta_1))(2 + \cos(\theta_2)), \end{array} \tag{38}$$

where θ_1 and θ_2 are generic angles between 0 and π.

The following remarks are useful, both for understanding formula (38) and for the further derivations in the rest of the paper.

[1] In the caption of the figures, showing stencils, we indicate in brackets the coefficient that multiplies the corresponding matrix block.

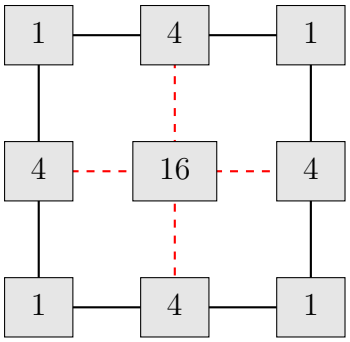


Figure 3: The stencil of the mass matrix ($\frac{H^2}{36}\frac{\mu^2}{\lambda}$)

Remark 4.2. *From Figure 3, it is insightful to observe that the stencil can be seen as a dyad, that is, as a matrix of rank one. In fact, it can be written as* $[1\ 4\ 1]^T[1\ 4\ 1]$. *Notice, that the symbol of the univariate stencil* $[1\ 4\ 1]$ *is exactly* $g(\theta) = 4 + 2\cos(\theta)$ *and so the symbol associated to* $[1\ 4\ 1]^T[1\ 4\ 1]$ *is exactly* $g(\theta_1)g(\theta_2)$, *which fully agrees with the symbol in (38) derived by using the canonical method. The following three general properties are worth noting. (i) If the stencil is a dyad then the bivariate symbol is separable, i.e., it is a product of two univariate functions, the first in the variable* θ_1, *the second in the variable* θ_2. *(ii) If the stencil is a sum of* k *dyads, then the symbol is the sum of* k *separable functions. (iii) The same idea applies unchanged in* d *dimensions, since the stencil becomes a* d*-dimensional tensor and its decomposition in sums of dyads gives immediately the formal expression of the symbol.*

Remark 4.3. *We note that the Lamé coefficients* μ *and* λ *depend on the material properties and can vary through the domain. The good news is that as already mentioned in* **GLT4** *the GLT machinery works also in presence of variable coefficients. For an example, see Figure 2 regarding the agreement of the spectra of the Schur complement with a sampling of its symbol (18), in the variable coefficient univariate problem (11).*

Figure 4 illustrates how well the symbol f^M describes the spectral properties of the mass matrix M. As we observe, there is a very good superposition of the two curves, as theoretically expected. However, the spectrum of M is uniformly a bit below the symbol f^M. This is easily explained by remark 4.1. As a result of the FEM assembly procedure, the Toeplitz matrix $T_n(f^M)$ can be seen as $T_n(f^M) = M + E$, where E is a low rank nonnegative definite matrix, related to lesser contributions along the boundary of Ω. Therefore, the Cauchy interlacing theorem (see e.g. [13]) gives the reason of the observed phenomenon.

4.2 Symbols of K, B and the Schur complement for Q1isoQ1

As in Problem 2.2, the block K is itself a two-by-two block matrix, $K = \begin{bmatrix} K_{11} & K_{12} \\ K_{21} & K_{22} \end{bmatrix}$, where, due to symmetry, $K_{21} = K_{12}^T$. Similarly to the mass matrix, the blocks have a block tri-

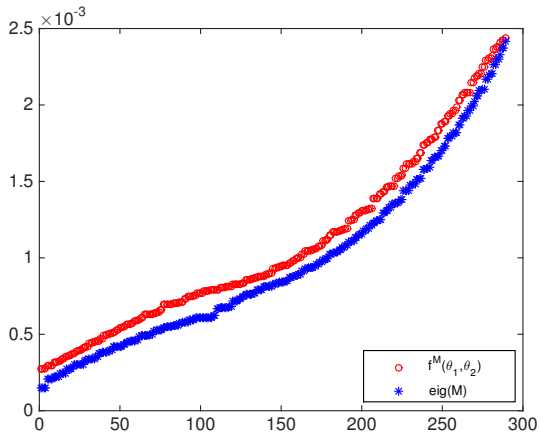
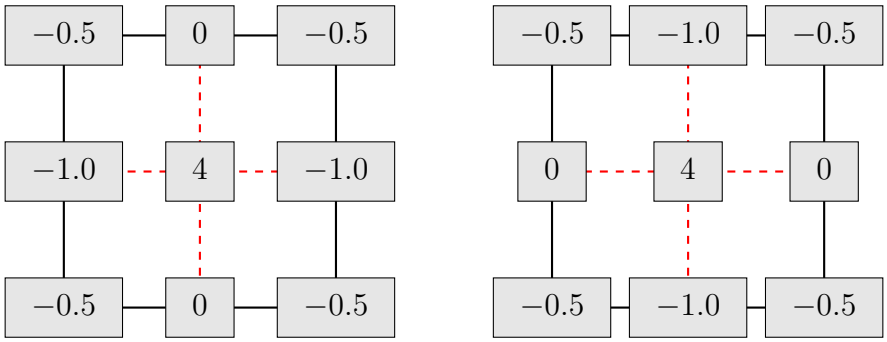


Figure 4: Bilinear basis functions: The spectrum of M vs sampling of its symbol $T(f^M)$

diagonal structure, with each sub-block being tridiagonal itself. The stencils, associated with the assembled matrices are shown in Figures 5 and 6.



(a) K_{11} (b) K_{22}

Figure 5: Q1isoQ1: The stencil of the blocks K_{11} and K_{22} (μ)

The symbols for K_{11} and K_{22} read as follows:

$$\begin{aligned} f_0^{K_{11}^b}(\theta) &= 4 - 2\cos(\theta), \quad f_1^{K_{11}^b}(\theta) = f_{-1}^{K_{11}^b}(\theta) = -\cos(\theta), \\ f^{K_{11}}(\theta_1,\theta_2) &= 4 - 2\cos(\theta_1)(1+\cos(\theta_2)), \end{aligned} \tag{39}$$

$$\begin{aligned} f_0^{K_{22}^b}(\theta) &= 4, \quad f_1^{K_{22}^b}(\theta) = f_{-1}^{K_{22}^b}(\theta) = -(1+\cos(\theta), \\ f^{K_{22}}(\theta_1,\theta_2) &= 4 - 2(1+\cos(\theta_1))\cos(\theta_2). \end{aligned} \tag{40}$$

Notice that both stencils can be seen as a sum of two dyads and in fact, according to Remark 4.2, the two symbols $f^{K_{11}}(\theta_1,\theta_2)$ and $f^{K_{22}}(\theta_1,\theta_2)$ can be regarded as the sum of exactly two separable functions.

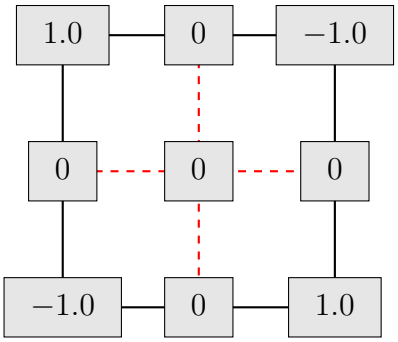


Figure 6: Q1isoQ1: The stencil of the block K_{12} $(\frac{\mu}{4})$

The corresponding symbol for K_{12} is given in (41),

$$\begin{array}{rcl} f_0^{K_{12}^b}(\theta) & = & 0, \quad f_1^{K_{12}^b}(\theta) = -f_{-1}^{K_{12}^b}(\theta) \; = \; 2\sin(\theta), \\ f^{K_{12}}(\theta_1,\theta_2) & = & 4\sin(\theta_1)\sin(\theta_2), \end{array} \tag{41}$$

which is a separable function since the stencil is a dyad.

Based on (39), (40) and(41), we construct now the symbol of the block K, that has the matrix form

$$f^K = \mu \begin{bmatrix} 4 - 2\cos(\theta_1)(1+\cos(\theta_2)) & \sin(\theta_1)\sin(\theta_2) \\ \sin(\theta_1)\sin(\theta_2) & 4 - 2(1+\cos(\theta_1))\cos(\theta_2) \end{bmatrix}. \tag{42}$$

Figure 7 illustrates that f^K represents well the spectral properties of the matrix K.

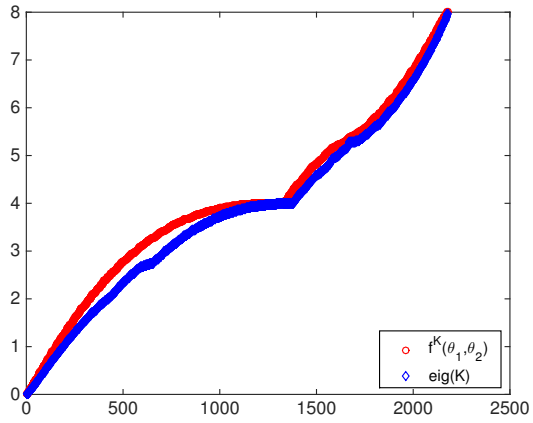


Figure 7: Q1isoQ1: The spectrum of K vs sampling of its symbol f^K, three refinements

We continue the derivation of the symbols of the blocks B_1 and B_2. For the case of Q1isoQ1 and also for Q2Q1, the blocks B_ℓ^T, $\ell = 1, 2$ are of size $n^2 \times m^2$, where m and n are the

number of mesh points in one direction, on two consecutive meshes, i.e., $n = 2(m-1)+1$. (We recall that we assume square meshes and row-wise lexicographical ordering of the unknowns.)

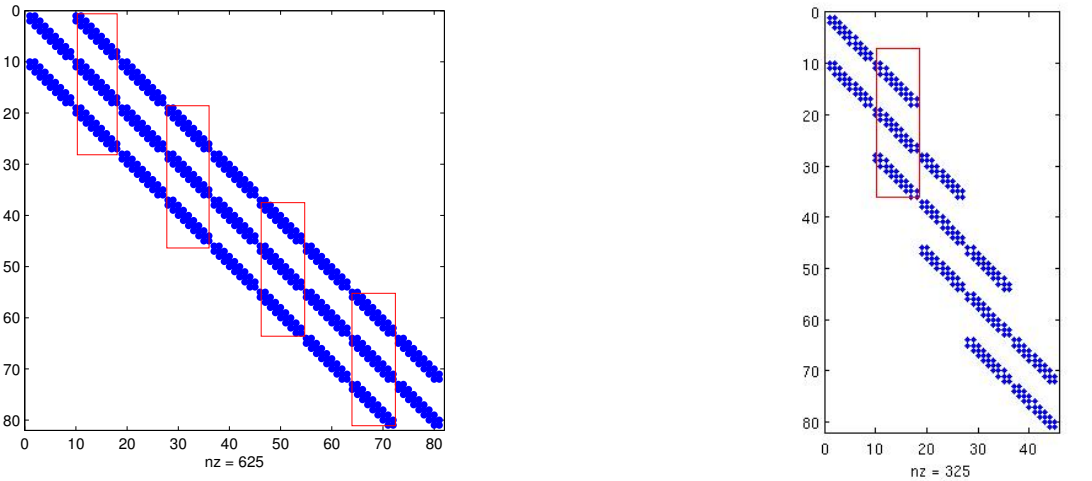
As the symbol can be related only to square matrices, in order to use the technique, we represent B_ℓ as a result of *downsampling* of larger square matrices $\widetilde{B}_\ell$ of size $n \times n$, namely,

$$B_\ell(n, m) = \widetilde{B}_\ell(n, n)H(n, m),$$

where H has a special structure, depicted in Figures 8, 9 and 10. The matrix H is a special instance of 2-level Toeplitz matrix (cf. [43] and references therein) and is used in various contexts, from wavelets and subdivision schemes [18, 26, 43] to multigrid methods [28, 54, 60], where it referred to as the *cutting* matrix. In our setting, in order to determine the symbol of the sequence of Schur complements, we use results from the multigrid framework.

For the considered discretization, $\widetilde{B}_\ell$ are five-diagonal block matrices, where each block is itself five-diagonal of size (n, n). The term *downsampling* describes a particular size reduction of a square matrix (of odd size), obtained by deleting each second column (single column sampling). Analogously, H can be constructed to sample blocks of columns (block-column sampling), or even to perform reduction block-wise and within the blocks simultaneously.

In Figures 8, 9 and 10 the idea of sampling is illustrated on small matrices.



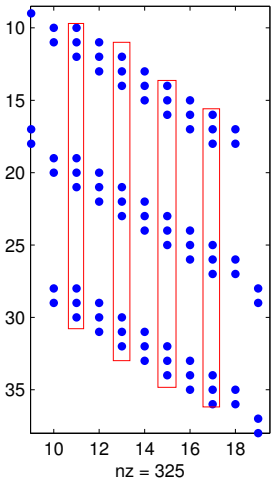
(a) Square matrix

(b) Block-column sampling

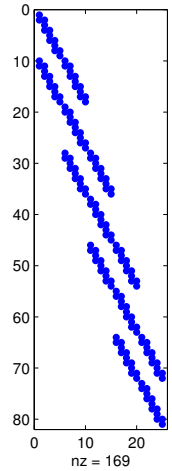
Figure 8: Two-level downsampling

Figure 8(a) represents the original square matrix, in this case a block tri-diagonal matrix. Figure 8(b) shows the result of a block-column downsampling, where all the blocks, marked in Figure 8(a) by rectangles, are deleted. Figure 9(a) shows the structure of one block-column of the already sampled matrix and Figure 9(b) depicts the effect of single column sampling, performed on the blocks in Figure 8(b).

Figure 10 shows the nonzero structure of the sampling matrices: 10(left) - block column sampling, where the elements are identity matrices of corresponding size and 10(right) - single column sampling, where the elements are ones.



(a) Zoom of a block-column



(b) Single-column sampling

Figure 9: Two-level downsampling, cont.

For the particular blocks B_ℓ, the sampling matrix H combines both block-column and single column sampling. It samples blocks of n columns and then samples the columns in the remaining blocks of size $n \times n$. Therefore H can be expressed as $H = H_b \times H_b$, where H_b performs a single-column sampling.

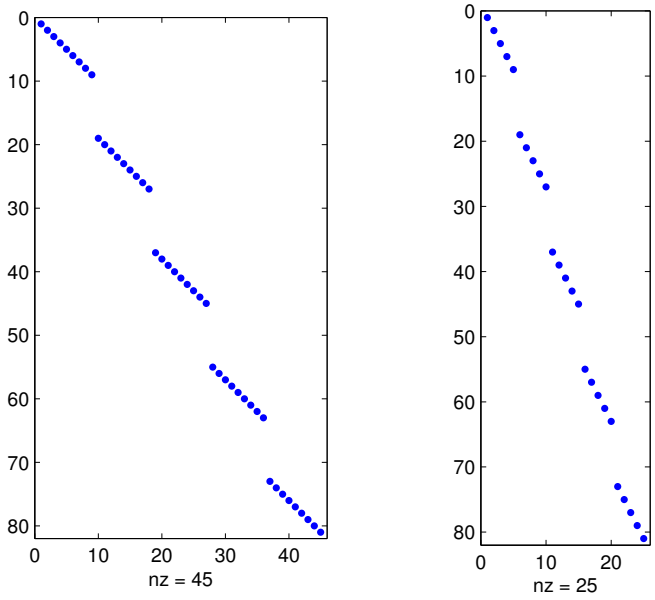
The blocks $\widetilde{B}_\ell$ have the following structure

$$\widetilde{B}_\ell = \begin{bmatrix} \widetilde{B}_0^\ell & \widetilde{B}_{-1}^\ell & \widetilde{B}_{-2}^\ell & & \\ \widetilde{B}_1^\ell & \widetilde{B}_0^\ell & \widetilde{B}_{-1}^\ell & \widetilde{B}_{-2}^\ell & \\ \widetilde{B}_2^\ell & \widetilde{B}_1^\ell & \widetilde{B}_0^\ell & \widetilde{B}_{-1}^\ell & \widetilde{B}_{-2}^\ell \\ & \ddots & \ddots & \ddots & \ddots \end{bmatrix}.$$

Each of the blocks $\widetilde{B}_k^\ell$, $k = -2, -1, 0, 1, 2$ is five-diagonal with entries as shown in Figure 11.

The corresponding symbols read as follows:

$$\begin{aligned} f_0^{B_1}(\theta) &= -20e^{i\theta} + 20e^{-i\theta} - 10e^{i2\theta} + 10e^{i2\theta} = -20\,i(2\sin(\theta) + \sin(2\theta)), \\ f_1^{B_1}(\theta) &= f_{-1}^{B_1} = -12\,i(2\sin(\theta) + \sin(2\theta)), \\ f_2^{B_1}(\theta) &= f_{-2}^{B_1} = -2\,i(\sin(\theta) + \sin(2\theta), \\ f_0^{B_2}(\theta) &= 0, \\ f_1^{B_2}(\theta) &= -f_{-1}^{B_2} = -4(5 + 6\cos(\theta) + \cos(2\theta)), \\ f_2^{B_2}(\theta) &= -f_{-2}^{B_2} = -2(5 + 6\cos(\theta) + \cos(2\theta)). \end{aligned}$$



(a) Examples of sampling matrices

Figure 10: Two-level downsampling

Then, the symbols of the whole blocks $\widetilde{B}_\ell$ become

$$\begin{aligned} f^{\widetilde{B}_1}(\theta_1,\theta_2) &= -4i\phi(\theta_1)\psi(\theta_2), \\ f^{\widetilde{B}_2}(\theta_1,\theta_2) &= -4i\psi(\theta_1)\phi(\theta_2), \end{aligned}$$

where $\phi(\theta) = 2\sin(\theta) + \sin(2\theta)$ and $\psi(\theta) = 5 + 6\cos(\theta) + \cos(2\theta)$. Notice that both symbols are separable thanks to the rank-one structure of the related stencils, in accordance with Remark 4.2. Furthermore, the function $\phi(\cdot)$ is a sine function, because it is related to an odd derivative (the first derivative in this case), and $\phi(\cdot)$ is a cosine function, because it is related to a even derivative (the zero order derivative in this case).

The symbol of the whole matrix $\mathcal{A}$ with no advection term is found to be as follows:

$$f^{\mathcal{A}} = \mu \begin{bmatrix} 4 - 2\cos(\theta_1)(1+\cos(\theta_2)) & \sin(\theta_1)\sin(\theta_2) & \left[\frac{h}{48} f^{\widetilde{B}_1}\right] \\ \sin(\theta_1)\sin(\theta_2) & 4 - 2(1+\cos(\theta_1))\cos(\theta_2) & \left[\frac{h}{48} f^{\widetilde{B}_2}\right] \\ \left[\frac{h}{48}\overline{f^{\widetilde{B}_1}}\right] & \left[\frac{h}{48}\overline{f^{\widetilde{B}_2}}\right] & -\frac{\mu}{\lambda}\frac{H^2}{9}(2+\cos(\theta_1))(2+\cos(\theta_2)) \end{bmatrix}. \tag{43}$$

In (43), the notation $[\cdot]$ indicates that in order to incorporate the effect of the matrices H the sampling for those blocks has to be performed in a special way (see [28, 54] for details).

$$\underbrace{\boxed{-1,-2,0,2,1}}_{f_2^{B_1}}\quad\underbrace{\boxed{-6,-12,0,12,6}}_{f_1^{B_1}}\quad\underbrace{\boxed{-10,-20,0,20,10}}_{f_0^{B_1}}\quad\underbrace{\boxed{-6,-12,0,12,6}}_{f_{-1}^{B_1}}\quad\underbrace{\boxed{-1,-2,0,2,1}}_{f_{-2}^{B_1}}$$

(a) $\widetilde{B}^1$

$$\underbrace{-\boxed{1,6,10,6,1}}_{f_2^{B_2}}\quad\underbrace{-\boxed{2,12,20,12,2}}_{f_1^{B_2}}\quad\underbrace{\boxed{0,0,0,0,0}}_{f_0^{B_2}}\quad\underbrace{\boxed{2,12,20,12,2}}_{f_{-1}^{B_2}}\quad\underbrace{\boxed{1,6,10,6,1}}_{f_{-2}^{B_2}}$$

(b) $\widetilde{B}^2$

Figure 11: The symbol of the matrices $\widetilde{B}^1$ and $\widetilde{B}^2$ ($\mu\frac{h}{48}$)

4.2.1 The symbol of the symmetric Schur complement

The next step is to compute the symbol for the exact Schur complement $\mathcal{S}$ of $\mathcal{A}$. To this end we compute the inverse of the symbol of the block K, as shown in (42). Via symbolic computations in `Matlab` we find the expression of $(f^K)^{-1}$ as

$$\frac{1}{\det(f^K)}\begin{bmatrix}4-2\,\cos(\theta_1)\,\cos(\theta_2)-2\,\cos(\theta_2) & -\sin(\theta_1)\,\sin(\theta_2)\\ -\sin(\theta_1)\,\sin(\theta_2) & 4-2\,\cos(\theta_1)\,\cos(\theta_2)-2\,\cos(\theta_1)\end{bmatrix}, \tag{44}$$

where

$$\begin{aligned}\det(f^K) &= 4\left(\cos(\theta_1)\,\cos(\theta_2)-1\right)^2-8\,\cos(\theta_2)-sin(\theta_1)^2\sin(\theta_2)^2-8\,\cos(\theta_1)+\\ &\quad 4\,\cos(\theta_1)\,\cos(\theta_2)\,(\cos(\theta_1)+\cos(\theta_2)-1)+12.\end{aligned}$$

Further, we have

$$\mathcal{S}=\rho M+\begin{bmatrix}B_1 & B_2\end{bmatrix}K^{-1}\begin{bmatrix}B_1^T\\ B_2^T\end{bmatrix}=\rho M+H^T\begin{bmatrix}\tilde{B}_1 & \tilde{B}_2\end{bmatrix}K^{-1}\begin{bmatrix}\tilde{B}_1^T\\ \tilde{B}_2^T\end{bmatrix}H, \tag{45}$$

where $H=H_b\otimes H_b$ and H_b performs the desired single column sampling. Based on **GLT1**, **GLT2**, **GLT3**, and observing that the involved blocks are of Toeplitz type up to low rank corrections (see Remark 4.1), we compute the symbol of

$$\tilde{B}K^{-1}\tilde{B}^T=\begin{bmatrix}\tilde{B}_1 & \tilde{B}_2\end{bmatrix}K^{-1}\begin{bmatrix}\tilde{B}_1^T\\ \tilde{B}_2^T\end{bmatrix}$$

as $v^*(f^K)^{-1}v$ with the vector v such that $v_1=f^{\widetilde{B}_1}$ and $v_2=f^{\widetilde{B}_2}$. After some simplifications and setting $G=f^{\tilde{B}K^{-1}\tilde{B}^T}$, we obtain

$$\begin{aligned}G &= 512\,(\cos(\theta_1)+1)^2\,(\cos(\theta_2)+1)^2\,\big(2\cos(\theta_1)^3\cos(\theta_2)^3+5\cos(\theta_1)^3\cos(\theta_2)^2\\ &\quad +3\cos(\theta_1)^3\,\cos(\theta_2)-\cos(\theta_1)^3+5\cos(\theta_1)^2\cos(\theta_2)^3+4\cos(\theta_1)^2\cos(\theta_2)^2\\ &\quad -8\cos(\theta_1)^2\,\cos(\theta_2)-10\cos(\theta_1)^2+3\,\cos(\theta_1)\cos(\theta_2)^3-8\,\cos(\theta_1)\cos(\theta_2)^2\\ &\quad -8\,\cos(\theta_1)\,\cos(\theta_2)+4\,\cos(\theta_1)-\cos(\theta_2)^3-10\cos(\theta_2)^2+4\,\cos(\theta_2)+16\big).\end{aligned}$$

Finally we consider the effect of H and H^T on the underlying symbol: indeed, making use of [60][Proposition 5.1, item 1] with $d = 2$ (or [54][Proposition 7.2, item 1] with $d = 2$ in the case of even symbols as G), the symbol of the exact Schur $f^{\mathcal{S}}$ is computed by direct evaluation of the formula

$$f^{\mathcal{S}}(\theta_1, \theta_2) = f^{M}(\theta_1, \theta_2) + \frac{1}{4}\left(\sum_{l=0}^{1}\sum_{m=0}^{1} G\left(\frac{\theta_1}{2} + l\pi, \frac{\theta_2}{2} + m\pi\right)\right). \tag{46}$$

The detailed derivation of the symbol $f^{\mathcal{S}}$ is presented in Proposition 5.5 in Appendix **B**.

Figure 12 illustrates the eigenvalues of the true Schur complement matrix and the match with the sampling of its symbol $f^{\mathcal{S}}$, performed as in (46).

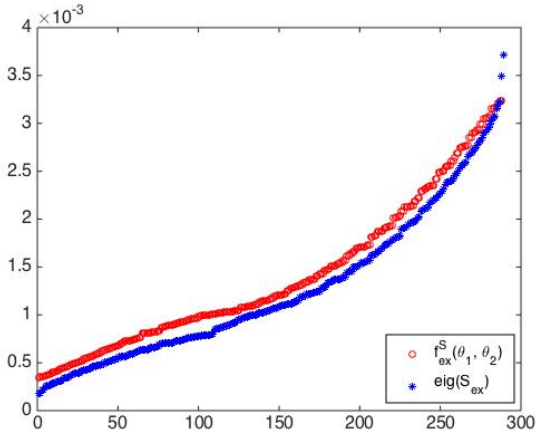


Figure 12: Q1isoQ1: The spectrum of the symmetric $\mathcal{S}$ vs sampling of its symbol $f^{\mathcal{S}}$, three refinements

4.2.2 The symbol of the advection term $\nabla(\mathbf{b}\cdot\mathbf{u})$ and $(\nabla\cdot\mathbf{u})\,\mathbf{c}$

Next we deal with the advection term in the 11-block of the matrix $\mathcal{A}$. To this end we consider first a term of the form $\nabla(\mathbf{b}\cdot\mathbf{u})$, where the advection vector is $\mathbf{b} = [b_1, b_2]$. We denote the matrix, arising from the discretization of $\nabla(\mathbf{b}\cdot\mathbf{u})$ by $A^{(1)}$. Similarly to K, the separate displacement ordering induces a two-by-two structure on $A^{(1)}$. Under lexicographical ordering the blocks of $A^{(1)}$, $A^{(1)}_{k,\ell}$, $k, \ell = 1, 2$ are block-tridiagonal and each block is again block tridiagonal. Similarly to the blocks of K, the assembled matrices $A^{(1)}_{k,\ell}$ can be related to stencils, depicted in Figure 13. We derive below the symbols for the blocks $A^{(1)}_{k,\ell}$.

$$\begin{aligned} f_0^{A_{11}^{1,b}}(\theta) &= -8\,i\,\sin(\theta), \quad f_1^{A_{11}^{1,b}}(\theta) = f_{-1}^{A_{11}^{1,b}}(\theta) = -2\,i\,\sin(\theta), \\ f^{A_{11}^{(1)}}(\theta_1, \theta_2) &= -4\,i\,\sin(\theta_1)(2 + \cos(\theta_2)), \end{aligned} \tag{47}$$

$$\begin{aligned} f_0^{A_{21}^{(1),b}}(\theta) &= 0, \quad f_1^{A_{21}^{(1),b}}(\theta) = -f_{-1}^{A_{21}^{(1),b}}(\theta) = 4 + 2\cos(\theta), \\ f^{A_{21}^{(1)}}(\theta_1, \theta_2) &= 4\,i\,\sin(\theta_2)(2 + \cos(\theta_1)), \end{aligned} \tag{48}$$

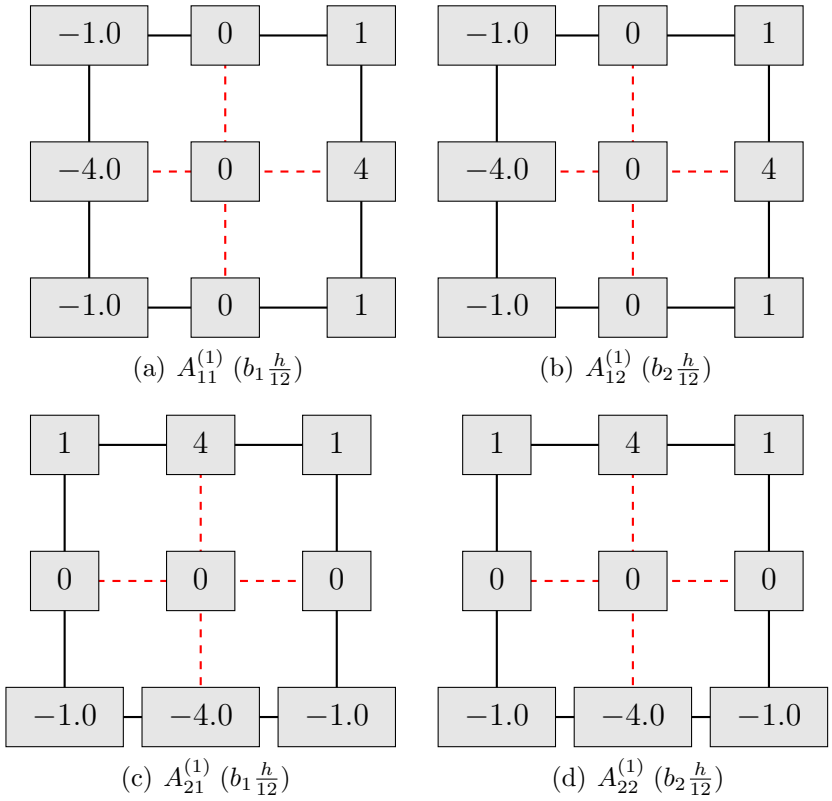


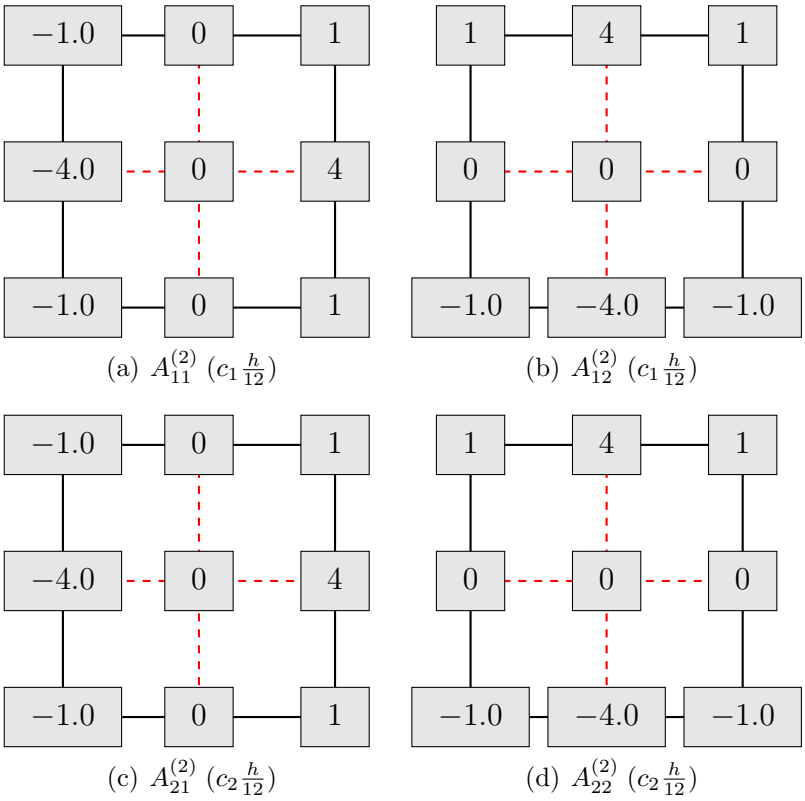
Figure 13: Q1isoQ1: The stencils of the advection block for $\nabla(\mathbf{b} \cdot \mathbf{u})$

We note that $f^{A_{12}^{(1)}}(\theta_1, \theta_2) = f^{A_{11}^{(1)}}(\theta_1, \theta_2)$ and $f^{A_{22}^{(1)}}(\theta_1, \theta_2) = f^{A_{21}^{(1)}}(\theta_1, \theta_2)$. The symbol of the block $A^{(1)}$ then takes the form:

$$f^{A^{(1)}} = -4\,i \begin{bmatrix} b_1 \sin(\theta_1)(2+\cos(\theta_2)) & b_2 \sin(\theta_1)(2+\cos(\theta_2)) \\ b_1 \sin(\theta_2)(2+\cos(\theta_1)) & b_2 \sin(\theta_2)(2+\cos(\theta_1)) \end{bmatrix}. \tag{49}$$

In an analogous way we derive the symbol of the matrix $A^{(2)}$ that arises from the discretization of the term $(\nabla \cdot \mathbf{u})\,\mathbf{c}$ with $\mathbf{c} = [c_1, c_2]$. The stencil of the matrix is shown in Figure 14. The symbol of $A^{(2)}$ has the following form

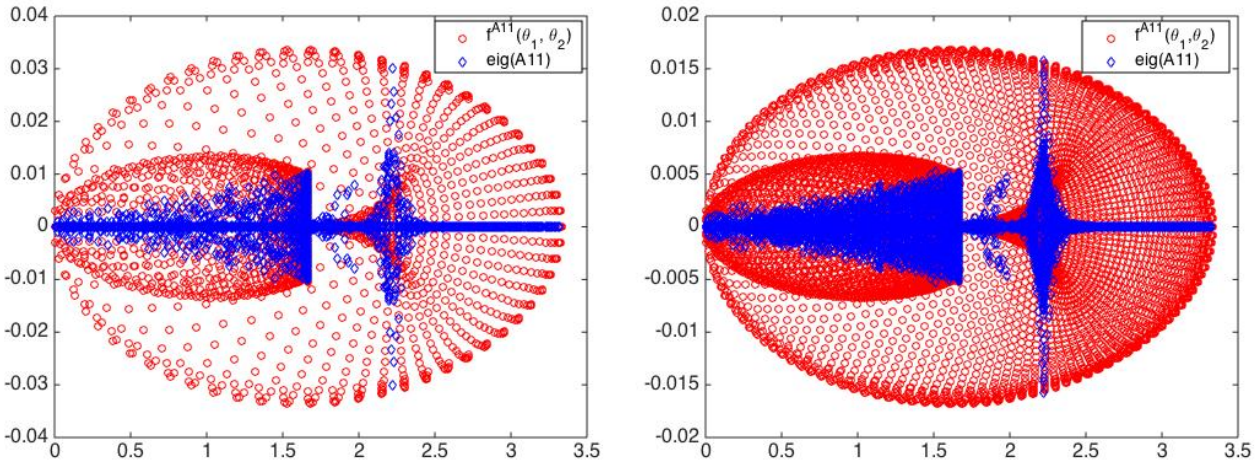
$$f^{A^{(2)}} = -i\,4 \begin{bmatrix} c_1 \sin(\theta_1)(2+\cos(\theta_2)) & c_1 \sin(\theta_2)(2+\cos(\theta_1)) \\ c_2 \sin(\theta_1)(2+\cos(\theta_2)) & c_2 \sin(\theta_2)(2+\cos(\theta_1)) \end{bmatrix}. \tag{50}$$



(a) $A_{11}^{(2)}$ $(c_1 \frac{h}{12})$ (b) $A_{12}^{(2)}$ $(c_1 \frac{h}{12})$

(c) $A_{21}^{(2)}$ $(c_2 \frac{h}{12})$ (d) $A_{22}^{(2)}$ $(c_2 \frac{h}{12})$

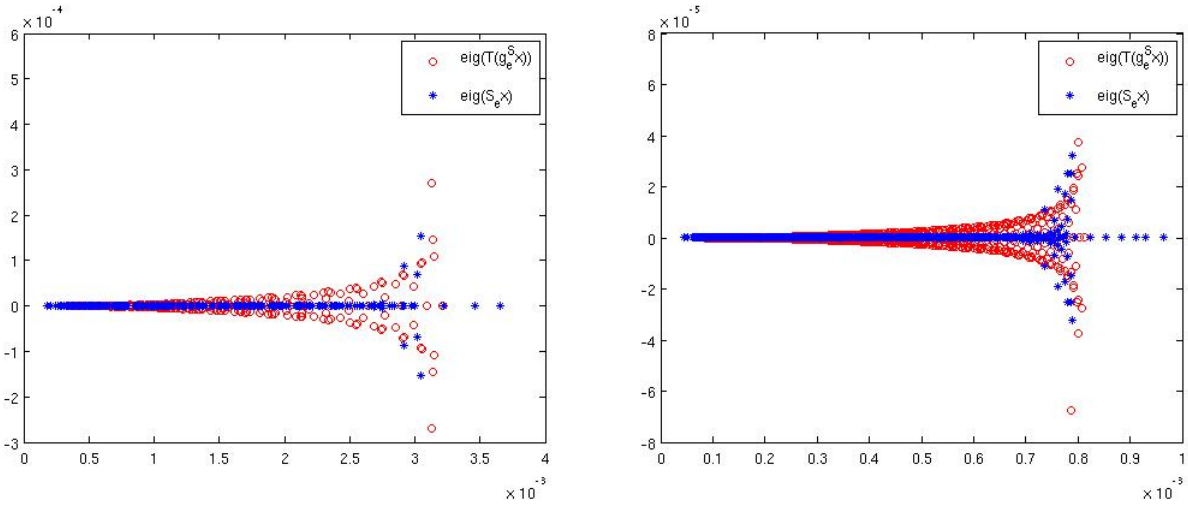
Figure 14: Q1isoQ1: The stencils of the advection block for $(\nabla \cdot \mathbf{u})\mathbf{c}$



(a) Three Refinements (b) Four Refinements

Figure 15: Q1isoQ1: The spectrum of $\mu K + A^{(1)} + A^{(2)}$ vs sampling of its symbol $f^K + f^{A^{(1)}} + f^{A^{(2)}}$; $\mathbf{b} = [0, 1]$, $\mathbf{c} = [0, 0]$

In Figure 15 we show the spectrum of $\mu K + A^{(1)} + A^{(2)}$ for $\mathbf{b} = [0, 1]$ and $\mathbf{c} = [0, 0]$ (in blue) and the sampled symbol (in red). The symbol of the nonsymmetric Schur complement is obtained in the same way as in the symmetric case. Figure 16 shows a very good match between the spectrum of the Schur complement and its sampled symbol.



(a) Three Refinements (b) Four Refinements

Figure 16: Q1isoQ1: The spectrum of the nonsymmetric $\mathcal{S}$ vs its symbol, $\mathbf{b} = [0, 1]$, $\mathbf{c} = [0, 0]$

4.3 Symbols of K, B and the Schur complement for Q2Q1

In this section we construct the symbols of K_{11}, K_{12}, K_{22}, B_1 and B_2 for the Q2Q1 discretization. We recall that K_{ij} and B_i^T are the blocks of the matrix K and B^T. The derivations are done only for the symmetric case. Therefore, $K_{12} = K_{21}^T$ and as the symbol of K_{12} is Hermitian-valued, it equals that of K_{21}. Because of the higher complexity of the expressions, the symbol of $\mathcal{S}$ is not constructed explicitly. When sampling it, we use (29).

The sparsity patterns of the blocks of K are shown in Figure 17. We see that each block matrix K_{ij} itself has a block structure and this agrees with the result in [31], showing that the use of Q_p Lagrangian Finite Elements on a scalar differential equation in d dimensions induces a matrix-valued symbol whose size is p^d. In fact, looking at the symbols associated with any of K_{ij}, $i = 1, 2$, equations (55)-(59), we find that the size is $4 = 2^2$ and indeed in our case $p = d = 2$.

The following remark sheds some light on the intricate structure of the matrices K_{ij}.

Remark 4.4. [Structure of the blocks K_{ij}] *According to Definition 1.2, we consider K_{ij} as a block tridiagonal Toeplitz matrix of size $\frac{n}{2}$ with a diagonal block C_0^{ij}, a lower diagonal block*

C_1^{ij}, *and a upper diagonal block* C_{-1}^{ij}*:*

$$K_{ij} = T_{\frac{n}{2}} = \begin{bmatrix} C_0^{ij} & C_{-1}^{ij} & \mathbf{0} & \cdots & \cdots & \mathbf{0} \\ C_1^{ij} & C_0^{ij} & C_{-1}^{ij} & \ddots & & \vdots \\ \mathbf{0} & C_1^{ij} & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & C_{-1}^{ij} & \mathbf{0} \\ \vdots & & \ddots & C_1^{ij} & C_0^{ij} & C_{-1}^{ij} \\ \mathbf{0} & \cdots & \cdots & \mathbf{0} & C_1^{ij} & C_0^{ij} \end{bmatrix}. \tag{51}$$

Furthermore, each block C_l^{ij} *has size* $2n$ *and has a particular structure, since*

$$C_l^{ij} = \begin{bmatrix} T^{11,ijl} & T^{12,ijl} \\ T^{21,ijl} & T^{22,ijl} \end{bmatrix}.$$

The positions of C_k^{ij} *are depicted in Figure 17. Here, any* $T^{vw,ijl}$ *is a matrix of size* n *and it can be viewed as* $T_{\frac{n}{2}}(\alpha^{vw,ijl})$*, where, taking into account Definition 1.2 and Definition 1.3,* $\alpha^{vw,ijl}$ *is* 2×2 *matrix-valued trigonometric polynomial. More specifically, we obtain*

$$C_l^{ij} = \begin{bmatrix} T_{\frac{n}{2}}(\alpha^{11,ijl}) & T_{\frac{n}{2}}(\alpha^{12,ijl}) \\ T_{\frac{n}{2}}(\alpha^{21,ijl}) & T_{\frac{n}{2}}(\alpha^{22,ijl}) \end{bmatrix}. \tag{52}$$

The structure in (52) is not standard and it does not fit into Definition 1.2. However, it has been already studied in the literature in a context of signal reconstruction from missing data [45]. In [45][Subsection 4.2] it is shown that there exists a permutation matrix Π *of size* $2n$ *(cf. [45][p. 407]) such that*

$$\Pi C_l^{ij} \Pi^T = T_{\frac{n}{2}}(f^{C_l^{ij}}), \tag{53}$$

$$f^{C_l^{ij}} = \begin{bmatrix} \alpha^{11,ijl} & \alpha^{12,ijl} \\ \alpha^{21,ijl} & \alpha^{22,ijl} \end{bmatrix}, \tag{54}$$

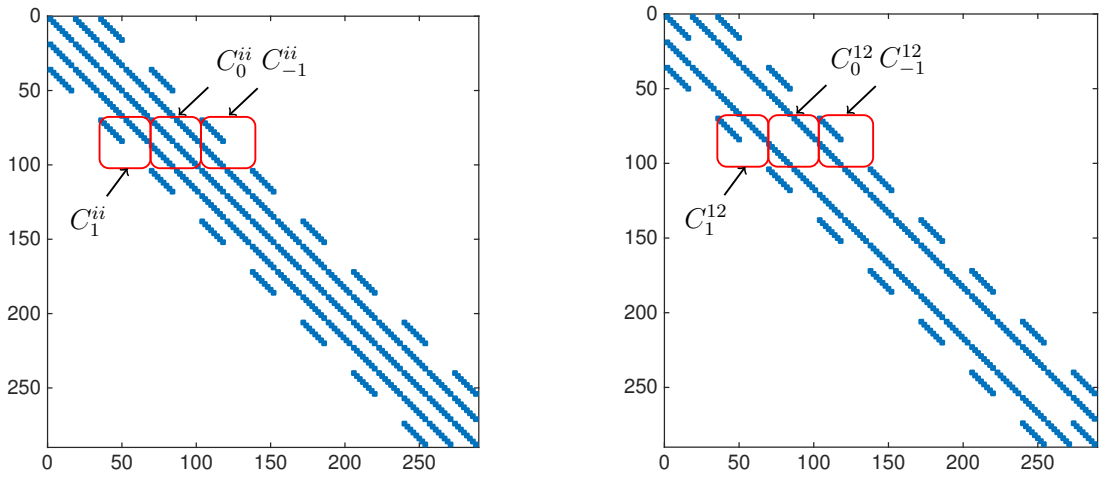
with $f^{C_l^{ij}}$ *being a* 4×4 *matrix-valued symbol, as expected ([31]). Consequently, using a tensor argument, recalling that unitary similarity transformation do not change eigenvalues and singular values, and taking into account the external structure in (51), it follows that*

$$(I_n \otimes \Pi) K_{ij} (I_n \otimes \Pi^T) = T_{\frac{n}{2},\frac{n}{2}}(f^{K^{ij}}),$$

where

$$f^{K_{ij}}(\theta_1, \theta_2) = f^{C_0^{ij}}(\theta_1) + f^{C_1^{ij}}(\theta_1) e^{\hat{\imath}\theta_2} + f^{C_{-1}^{ij}}(\theta_1) e^{-\hat{\imath}\theta_2}. \tag{55}$$

Taking into consideration (54) and (55), we derive the formal expression of the symbol $f^{K_{ij}}(\theta_1, \theta_2)$. We construct first the symbol of the blocks C_k^{ij}, $i, j = 1, 2$, $k = -1, 0, 1$ that constitute the structure of the blocks K_{ij}. From Figure 17 we see that each block C_i^{11} is



(a) K_{ii}, $i = 1, 2$ (b) K_{12}

Figure 17: Q2Q1: The sparsity pattern of the blocks of K

block tridiagonal and its structure translates to the symbols presented in (57).

$$C_1^{11} = \begin{bmatrix} A_2^{11} & A_1^{11} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \to f^{C_1^{11}}(\theta_1) = \frac{\mu}{90} \begin{bmatrix} f_2^{11}(\theta_1) & f_1^{11}(\theta_1) \\ \mathbf{0} & \mathbf{0} \end{bmatrix}, \tag{56}$$

$$C_0^{11} = \begin{bmatrix} A_0^{11} & A_{-1}^{11} \\ \widehat{A}_1^{11} & \widehat{A}_0^{11} \end{bmatrix} \to f^{C_0^{11}}(\theta_1) = \frac{\mu}{90} \begin{bmatrix} f_0^{11}(\theta_1) & f_{-1}^{11}(\theta_1) \\ \widehat{f}_1^{11}(\theta_1) & \widehat{f}_0^{11}(\theta_1) \end{bmatrix}, \tag{57}$$

$$C_{-1}^{11} = \begin{bmatrix} A_{-2}^{11} & \mathbf{0} \\ \widehat{A}_{-1}^{11} & \mathbf{0} \end{bmatrix} \to f^{C_{-1}^{11}}(\theta_1) = \frac{\mu}{90} \begin{bmatrix} f_{-2}^{11}(\theta_1) & \mathbf{0} \\ \widehat{f}_{-1}^{11}(\theta_1) & \mathbf{0} \end{bmatrix}. \tag{58}$$

Here $\mathbf{0}$ denotes a zero block matrix of proper size. We observe also that for the considered square mesh we have $\widehat{f}_1^{11}(\theta_1) = \widehat{f}_{-1}^{11}(\theta_1) = f_1^{11}(\theta_1) = f_{-1}^{11}(\theta_1)$ and $f_2^{11}(\theta_1) = f_{-2}^{11}(\theta_1)$ where

$$\begin{aligned}
f_2^{11}(\theta_1) &= \begin{bmatrix} -20 & 18 \\ 18 & -16 \end{bmatrix} + \begin{bmatrix} -3 & 18 \\ 0 & 0 \end{bmatrix} e^{\hat{\imath}\theta_1} + \begin{bmatrix} -3 & 0 \\ 18 & 0 \end{bmatrix} e^{-\hat{\imath}\theta_1} \\
f_1^{11}(\theta_1) &= \begin{bmatrix} -8 & -48 \\ -48 & -64 \end{bmatrix} + \begin{bmatrix} 12 & -48 \\ 0 & 0 \end{bmatrix} e^{\hat{\imath}\theta_1} + \begin{bmatrix} 12 & 0 \\ -48 & 0 \end{bmatrix} e^{-\hat{\imath}\theta_1} \\
f_0^{11}(\theta_1) &= \begin{bmatrix} 336 & -100 \\ -100 & 480 \end{bmatrix} + \begin{bmatrix} 2 & -100 \\ 0 & 0 \end{bmatrix} e^{\hat{\imath}\theta_1} + \begin{bmatrix} 2 & 0 \\ -100 & 0 \end{bmatrix} e^{-\hat{\imath}\theta_1} \\
\widehat{f}_0^{11}(\theta_1) &= \begin{bmatrix} 576 & -224 \\ -224 & 768 \end{bmatrix} + \begin{bmatrix} 16 & -224 \\ 0 & 0 \end{bmatrix} e^{\hat{\imath}\theta_1} + \begin{bmatrix} 16 & 0 \\ -224 & 0 \end{bmatrix} e^{-\hat{\imath}\theta_1}.
\end{aligned} \tag{59}$$

Figure 18 shows the match between sampling of the constructed symbol $f^{K_{11}}(\theta_1, \theta_2)$ compared with the eigenvalues of the block K_{11}.

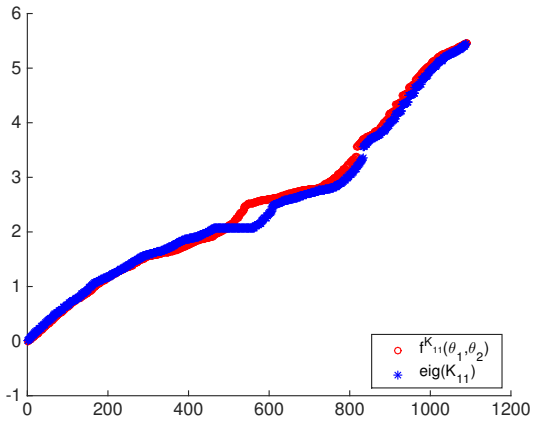


Figure 18: Q1isoQ1: Spectrum of K_{11} vs its sampled symbol $f^{K_{11}}(\theta_1, \theta_2)$

In a similar way we construct the symbol of $f^{C_k^{12}}, k = -1, 0, 1$, namely,

$$C_1^{12} = \begin{bmatrix} A_2^{12} & A_1^{12} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \to f^{C_1^{12}}(\theta_1) = \frac{\mu}{180} \times \begin{bmatrix} f_2^{12}(\theta_1) & f_1^{12}(\theta_1) \\ \mathbf{0} & \mathbf{0} \end{bmatrix},$$
$$C_0^{12} = \begin{bmatrix} \mathbf{0} & A_{-1}^{12} \\ \widehat{A}_1^{12} & \mathbf{0} \end{bmatrix} \to f^{C_0^{12}}(\theta_1) = \frac{\mu}{180} \times \begin{bmatrix} \mathbf{0} & f_{-1}^{12}(\theta_1) \\ \widehat{f}_1^{12}(\theta_1) & \mathbf{0} \end{bmatrix},$$
$$C_{-1}^{12} = \begin{bmatrix} A_{-2}^{12} & \mathbf{0} \\ \widehat{A}_{-1}^{12} & \mathbf{0} \end{bmatrix} \to f^{C_{-1}^{12}}(\theta_1) = \frac{\mu}{180} \times \begin{bmatrix} f_{-2}^{12}(\theta_1) & \mathbf{0} \\ \widehat{f}_{-1}^{12}(\theta_1) & \mathbf{0} \end{bmatrix},$$

where $\widehat{f}_1^{12}(\theta_1) = f_1^{12}(\theta_1) = -\widehat{f}_{-1}^{12}(\theta_1) = -f_{-1}^{12}(\theta_1), \quad f_2^{12}(\theta_1) = -f_{-2}^{12}(\theta_1)$ and

$$f_2^{12}(\theta_1) = \begin{bmatrix} 0 & 20 \\ -20 & 0 \end{bmatrix} + \begin{bmatrix} 5 & -20 \\ 0 & 0 \end{bmatrix} e^{\hat{\imath}\theta_1} + \begin{bmatrix} -5 & 0 \\ 20 & 0 \end{bmatrix} e^{-\hat{\imath}\theta_1},$$
$$f_1^{12}(\theta_1) = \begin{bmatrix} 0 & -80 \\ 80 & 0 \end{bmatrix} + \begin{bmatrix} -20 & 80 \\ 0 & 0 \end{bmatrix} e^{\hat{\imath}\theta_1} + \begin{bmatrix} 20 & 0 \\ -80 & 0 \end{bmatrix} e^{-\hat{\imath}\theta_1}.$$

Figure 19 illustrates how well a sampling of the constructed symbol $f^{K_{12}}(\theta_1, \theta_2)$ matches with the eigenvalues of the block K_{12}.
Finally, we construct the symbols $f^{C_k^{22}}, k = -1, 0, 1$,

$$C_1^{22} = \begin{bmatrix} A_2^{22} & A_1^{22} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \to f^{C_1^{22}}(\theta_1, \theta_2) = \frac{\mu}{90} \times \begin{bmatrix} f_2^{22}(\theta_1) & f_1^{22}(\theta_1) \\ \mathbf{0} & \mathbf{0} \end{bmatrix},$$
$$C_0^{22} = \begin{bmatrix} A_0^{22} & A_{-1}^{22} \\ \widehat{A}_1^{22} & \widehat{A}_0^{22} \end{bmatrix} \to f^{C_0^{22}}(\theta_1, \theta_2) = \frac{\mu}{90} \times \begin{bmatrix} f_0^{22}(\theta_1) & f_{-1}^{22}(\theta_1) \\ \widehat{f}_1^{22}(\theta_1) & \widehat{f}_0^{22}(\theta_1) \end{bmatrix},$$
$$C_{-1}^{22} = \begin{bmatrix} A_{-2}^{22} & \mathbf{0} \\ \widehat{A}_{-1}^{22} & \mathbf{0} \end{bmatrix} \to f^{C_{-1}^{22}}(\theta_1, \theta_2) = \frac{\mu}{90} \times \begin{bmatrix} f_{-2}^{22}(\theta_1) & \mathbf{0} \\ \widehat{f}_{-1}^{22}(\theta_1) & \mathbf{0} \end{bmatrix},$$

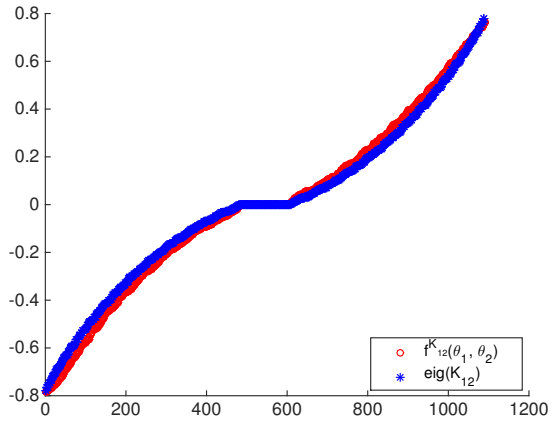


Figure 19: Q2Q1: The eigenvalues of K_{12} vs sampling of $f^{K_{12}}(\theta_1, \theta_2)$

where $\widehat{f}_1^{22}(\theta_1) = \widehat{f}_{-1}^{22}(\theta_1) = f_1^{22}(\theta_1) = f_{-1}^{22}(\theta_1)$, $f_2^{22}(\theta_1) = f_{-2}^{22}(\theta_1)$ and

$$
\begin{aligned}
f_2^{22}(\theta_1) &= \begin{bmatrix} 2 & 12 \\ 12 & 16 \end{bmatrix} + \begin{bmatrix} -3 & 12 \\ 0 & 0 \end{bmatrix} e^{\hat{\imath}\theta_1} + \begin{bmatrix} -3 & 0 \\ 12 & 0 \end{bmatrix} e^{-\hat{\imath}\theta_1}, \\
f_1^{22}(\theta_1) &= \begin{bmatrix} -100 & -48 \\ -48 & -224 \end{bmatrix} + \begin{bmatrix} 18 & -48 \\ 0 & 0 \end{bmatrix} e^{\hat{\imath}\theta_1} + \begin{bmatrix} 18 & 0 \\ -48 & 0 \end{bmatrix} e^{-\hat{\imath}\theta_1}, \\
f_0^{22}(\theta_1) &= \begin{bmatrix} 336 & -8 \\ -8 & 576 \end{bmatrix} + \begin{bmatrix} -20 & -8 \\ 0 & 0 \end{bmatrix} e^{\hat{\imath}\theta_1} + \begin{bmatrix} -20 & 0 \\ -8 & 0 \end{bmatrix} e^{-\hat{\imath}\theta_1}, \\
\widehat{f}_0^{22}(\theta_1) &= \begin{bmatrix} 480 & -64 \\ -64 & 768 \end{bmatrix} + \begin{bmatrix} -16 & -64 \\ 0 & 0 \end{bmatrix} e^{\hat{\imath}\theta_1} + \begin{bmatrix} -16 & 0 \\ -64 & 0 \end{bmatrix} e^{-\hat{\imath}\theta_1}.
\end{aligned}
$$

In Figure 20 we illustrate the correspondence between a sample of the constructed symbol $f^{K_{22}}(\theta_1, \theta_2)$ and the eigenvalues of the block K_{22}.

We proceed with computing the symbol of the matrices B_1 and B_2. As described in Section 4.2, we construct the symbol of larger square matrices $\widetilde{B}^1$ and $\widetilde{B}_2$. The actual blocks B_1 and B_2 result from a downsampling the square matrices $\widetilde{B}^1$ and $\widetilde{B}_2$ using a restriction matrix. Figure 21 depicts the sparsity pattern of B_1^T and B_2^T. Note that the pattern is described as a tensor product of two smaller matrices

$$
B_1^T = \frac{\mu h}{18}\, T \otimes R, \qquad B_2^T = \frac{\mu h}{18}\, - R \otimes T.
$$

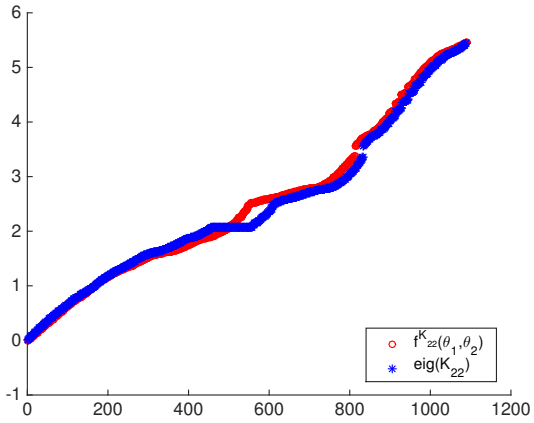
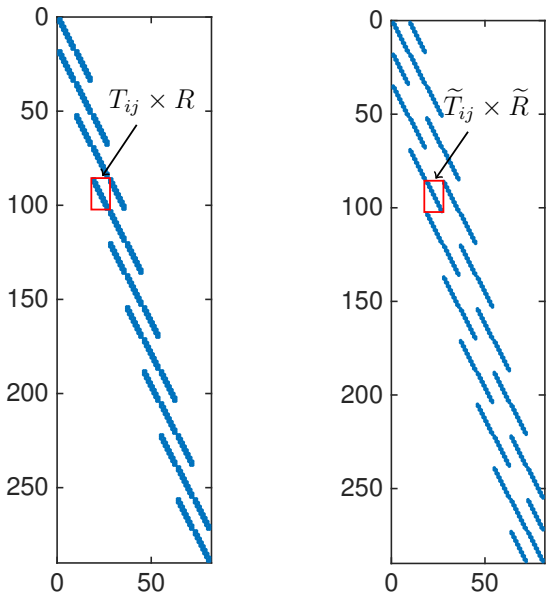


Figure 20: Q2Q1: The eigenvalues of K_{22} vs sampling of $f^{K_{22}}(\theta_1, \theta_2)$



(a) $(B_1)^T$ (b) $(B_2)^T$

Figure 21: Q2Q1: The sparsity pattern of the blocks of B^T

Now we define $\widetilde{B}_1 = \widetilde{T} \otimes \widetilde{R}$, $\widetilde{B}_2 = -\widetilde{R} \otimes \widetilde{T}$, where $\widetilde{T} = TH_b$, $\widetilde{R} = RH_b$ and

$$\widetilde{T} = \begin{bmatrix} 1 & 2 & & & & \\ 2 & 2 & 2 & & & \\ & 2 & 2 & 2 & & \\ & & \ddots & \ddots & \ddots & \\ & & & 2 & 2 & 2 \\ & & & & 2 & 1 \end{bmatrix}, \quad H_b = \begin{bmatrix} 1 & 0 & & & \\ 0 & 0 & & & \\ 0 & 1 & 0 & & \\ & 0 & 0 & & \\ & 0 & 1 & & \\ & & & \ddots & \end{bmatrix}, \tag{60}$$

$$\widetilde{R} = \begin{bmatrix} -5 & -4 & -1 & & & & & & \\ 4 & 0 & -4 & -1 & & & & & \\ 1 & 4 & 0 & -4 & -1 & & & & \\ & 1 & 4 & 0 & -4 & -1 & & & \\ & & \ddots & \ddots & \ddots & \ddots & \ddots & & \\ & & & 1 & 4 & 0 & -4 & -1 & \\ & & & & 1 & 4 & 0 & -4 & \\ & & & & & 1 & 4 & 5 & \end{bmatrix}. \tag{61}$$

From (60) and (61) the symbols of $\widetilde{B}_1$ and $\widetilde{B}_2$

$$f^{\widetilde{B}_1}(\theta_1, \theta_2) = \tfrac{\mu h}{18} i\phi(\theta_1)\psi(\theta_2), \qquad f^{\widetilde{B}_2}(\theta_1, \theta_2) = -\tfrac{\mu h}{18} i\psi(\theta_1)\phi(\theta_2),$$

where $\phi(\theta) = 2 + 4\cos(\theta)$ and $\psi(\theta) = 8\sin(\theta) + 2\sin(2\theta)$.

Notice that in accordance with Remark 4.2 both symbols are separable thanks to the rank-one structure of the related stencils. Furthermore, the function $\psi(\cdot)$ is a sine function, because it is related to a odd derivative (the first derivative), and $\phi(\cdot)$ is a cosine function, because it is related to a even derivative (the zero order derivative).

4.3.1 The symbol of the Schur complement

Next we compute the symbol of the exact Schur complement $\mathcal{S}$ of $\mathcal{A}$. To this end we sample the computed symbols in an appropriate way, but differently from the case of Q1isoQ1 elements, see Appendix **A** and Appendix **B** for an explanation. Taking into consideration (29), as the basic ingredient to construct the symbol of the Schur complement and using Proposition 5.8 in Appendix **B**, the symbol of $\mathcal{S}$ becomes

$$f^{\mathcal{S}}(\theta_1, \theta_2) = f^{M}(\theta_1, \theta_2) + G_{11}(\theta_1, \theta_2), \tag{62}$$

where the formal expression of G is given by

$$\begin{aligned} G(\theta_1, \theta_2) &= \left[g_{\widetilde{B}_1} [f^{K_{11}}]^{-1} g_{\widetilde{B}_1^T} + f^R [f^{\mathcal{S}_K}]^{-1} [f^R]^T \right](\theta_1, \theta_2), \\ f^R(\theta_1, \theta_2) &= g_{\widetilde{B}_2}(\theta_1, \theta_2) - g_{\widetilde{B}_1}(\theta_1, \theta_2)[f^{K_{11}}]^{-1}(\theta_1, \theta_2) f^{K_{12}}(\theta_1, \theta_2), \\ f^{\mathcal{S}_K}(\theta_1, \theta_2) &= f^{K_{22}}(\theta_1, \theta_2) - f^{K_{21}}(\theta_1, \theta_2)[f^{K_{11}]^{-1}}(\theta_1, \theta_2) f^{K_{12}}(\theta_1, \theta_2), \end{aligned} \tag{63}$$

with $0 < \theta_1, \theta_2 \leq \pi$ and where

$$g_{\widetilde{B}_j}(\theta_1, \theta_2) = \alpha_j^{[2]}(\theta_1) \otimes \beta_j^{[2]}(\theta_2), \qquad j = 1, 2.$$

The proof is found in Proposition 5.2 in Appendix **A**. We recall that $f^{\widetilde{B}_j}(\theta_1, \theta_2) = \alpha_j(\theta_1)\beta_j(\theta_2)$, $j = 1, 2$ and we observe that $\alpha_j(\theta)$ is related to $\alpha_j^{[2]}(\theta)$ and $\beta_j(\theta)$ is related to $\beta_j^{[2]}(\theta)$ according to the notation introduced in Remark 1.4. Finally we note that $g_{\widetilde{B}_1}$ and $g_{\widetilde{B}_2}$ are the conjugate transpose of $g_{\widetilde{B}_1^T}$ and $g_{\widetilde{B}_2^T}$.

Figure 22 shows the symbol of the exact Schur complement compared with a sampling of its symbol.

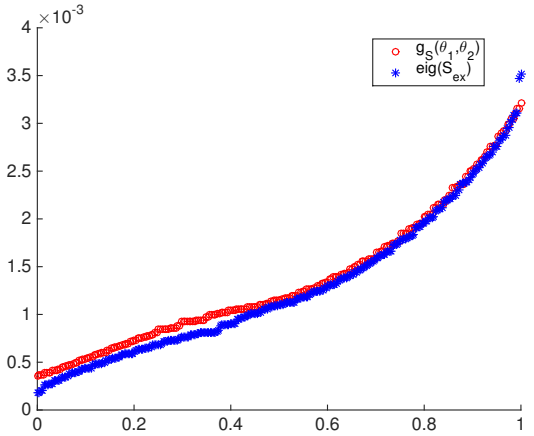


Figure 22: Q2Q1: The eigenvalues of the symmetric $\mathcal{S}$ vs sampling of its symbol

5 Conclusions and open problems

In this work we consider large linear systems of algebraic equations arising from the Finite Element approximation of coupled partial differential equations. As a case study we focus on the linear elasticity equations, formulated in a saddle point form in order to allow for modeling of purely incompressible materials. Using the notion of the so-called *spectral symbol* in the Generalized Locally Toeplitz (GLT) setting, we identify the GLT symbol (in the Weyl sense) of the sequence of matrices $\{\mathcal{A}_n\}$ approximating the elasticity equations. Further, by exploiting the property that the GLT class defines an algebra of matrix sequences and the fact that Schur complements are obtained via elementary operation on the blocks of $\mathcal{A}_n$, we derive the symbols g_s of the associated sequences of Schur complements $\{\mathcal{S}_n\}$. According to the GLT theory, the eigenvalues of $\mathcal{S}_n$ for large n are described by a sampling of g_s on a uniform grid of its domain of definition.

We derive the symbols of $\mathcal{A}_n$ and $\mathcal{S}_n$ for two stable pairs of FE discretization spaces, Q1isoQ1 and Q2Q1. In the case Q1isoQ1 we also derive the corresponding symbols for

the case where the PDE problem includes an advection term and the corresponding system matrix, and respectively, the Schur complement matrix are nonsymmetric.

As the underlying matrices originate from a system of PDEs and, in addition, the FEM discretization is done to satisfy certain stability conditions, the resulting matrices have a nontrivial block structure. To explain how to apply the GTL technology in such a case, we extend the already existing results for stationary problems with novel theoretical findings.

All numerical experiments show that, for the studied discrete problems, the sampling of the symbol agrees very well with the computed spectrum even for a relatively small-sized matrices.

As a further step, subject of a forthcoming research, we plan to employ such a spectral information for suggesting and for studying proper preconditioned Krylov techniques, in particular, the element-wise Schur approximation, both in the case of linear elasticity and Navier Stokes problems.

Acknowledgements

The work of the third author is partly supported by Donation **KAW 2013.0341** from the Knut & Alice Wallenberg Foundation in collaboration with the Royal Swedish Academy of Sciences, supporting Swedish research in mathematics.

References

[1] A. Aricó, M. Donatelli, and S. Serra-Capizzano. V-cycle optimal convergence for certain (multilevel) structured linear systems. *SIAM Journal on Matrix Analysis and Applications*, 26(1):186–214, 2004.

[2] O. Axelsson. *Iterative solution methods*. Cambridge University Press, 1996.

[3] O. Axelsson. Finite difference methods. *Encyclopedia of Computational Mechanics*, 2004.

[4] O. Axelsson, R. Blaheta, and M. Neytcheva. Preconditioning of boundary value problems using elementwise schur complements. *SIAM Journal on Matrix Analysis and Applications*, 31(2):767–789, 2009.

[5] O. Axelsson and G. Lindskog. On the rate of convergence of the preconditioned conjugate gradient method. *Numerische Mathematik*, 48(5):499–523, 1986.

[6] O. Axelsson and M. Neytcheva. Preconditioning methods for linear systems arising in constrained optimization problems. *Numerical Linear Algebra with Applications*, 10(1-2):3–31, 2003.

[7] O. Axelsson and M. Neytcheva. A general approach to analyse preconditioners for two-by-two block matrices. *Numerical Linear Algebra with Applications*, 20(5):723–742, 2013.

[8] O. Axelsson and A. Padiy. On a robust and scalable linear elasticity solver based on a saddle point formulation. *International Journal for Numerical Methods in Engineering*, 44(6):801–818, 1999.

[9] E. Bängtsson and B. Lund. A comparison between two solution techniques to solve the equations of glacially induced deformation of an elastic earth. *International Journal for Numerical Methods in Engineering*, 75(4):479–502, 2008.

[10] B. Beckermann and A. B. J. Kuijlaars. Superlinear convergence of conjugate gradients. *SIAM J. Numer. Anal.*, 39(1):300–329, Jan. 2001.

[11] B. Beckermann and S. SerraCapizzano. On the asymptotic spectrum of finite element matrix sequences. *SIAM Journal on Numerical Analysis*, 45(2):746–769, 2007.

[12] M. Benzi, G. H. Golub, and J. Liesen. Numerical solution of saddle point problems. *Acta Numerica*, 14:1–137, 5 2005.

[13] R. Bhatia. *Matrix analysis*, volume 169. Springer Science & Business Media, 1997.

[14] A. Böttcher and B. Silbermann. *Introduction to large truncated Toeplitz matrices*. Springer Science & Business Media, 1999.

[15] D. Braess. *Finite elements: Theory, fast solvers, and applications in solid mechanics*. Cambridge University Press, 2007.

[16] P. G. Ciarlet. *The finite element method for elliptic problems*. Elsevier, 1978.

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

[18] I. Daubechies et al. *Ten lectures on wavelets*, volume 61. SIAM, 1992.

[19] P. J. Davis. *Circulant matrices*. American Mathematical Soc., 1979.

[20] M. Donatelli, C. Garoni, C. Manni, S. Serra-Capizzano, and H. Speleers. Robust and optimal multi-iterative techniques for iga galerkin linear systems. *Computer Methods in Applied Mechanics and Engineering*, 284(0):230 – 264, 2015. Isogeometric Analysis Special Issue.

[21] M. Donatelli, C. Garoni, C. Manni, S. Serra-Capizzano, and H. Speleers. Robust and optimal multi-iterative techniques for iga collocation linear systems. *Computer Methods in Applied Mechanics and Engineering*, under revision.

[22] M. Donatelli, C. Garoni, C. Manni, S. Serra-Capizzano, and H. Speleers. Spectral analysis of matrices in collocation methods with b-splines. *Math. Comput.*, under revision. TW648; U. Leuven, june 2014.

[23] M. Donatelli, C. Garoni, M. Mazza, S. Serra-Capizzano, and D. Sesana. Spectral behavior of preconditioned non-hermitian multilevel block toeplitz matrices with matrix-valued symbol. *Applied Mathematics and Computation*, 245(0):158 – 173, 2014.

[24] M. Donatelli, M. Neytcheva, and S. Serra-Capizzano. Canonical eigenvalue distribution of multilevel block toeplitz sequences with non-hermitian symbols. In W. Arendt, J. A. Ball, J. Behrndt, K.-H. Förster, V. Mehrmann, and C. Trunk, editors, *Spectral Theory, Mathematical System Theory, Evolution Equations, Differential and Difference Equations*, volume 221 of *Operator Theory: Advances and Applications*, pages 269–291. Springer Basel, 2012.

[25] A. Dorostkar, M. Neytcheva, and B. Lund. On some block-preconditioners for saddle point systems and their cpu–gpu performance. Technical Report 2015-003, Uppsala University, Geophysics, 2015.

[26] N. Dyn and D. Levin. Subdivision schemes in geometric modelling. *Acta Numerica*, 11:73–144, 1 2002.

[27] G. Fiorentino and S. Serra. Multigrid methods for symmetric positive definite block toeplitz matrices with nonnegative generating functions. *SIAM Journal on Scientific Computing*, 17(5):1068–1081, 1996.

[28] G. Fiorentino and S. Serra-Capizzano. Multigrid methods for toeplitz matrices. *CALCOLO*, 28(3-4):283–305, 1991.

[29] M. Fortin and F. Brezzi. *Mixed and hybrid finite element methods*. New York: Springer-Verlag, 1991.

[30] C. Garoni, C. Manni, F. Pelosi, S. Serra-Capizzano, and H. Speleers. On the spectrum of stiffness matrices arising from isogeometric analysis. *Numer. Math.*, 127(4):751–799, Aug. 2014.

[31] C. Garoni, S. Serra-Capizzano, and D. Sesana. Spectral analysis and spectral symbol of d-variate $\mathbb{Q}_p$ Lagrangian FEM stiffness matrices. Technical Report 2014-021, Department of Information Technology, Uppsala University, Nov. 2014.

[32] C. Garoni, S. Serra-Capizzano, and D. Sesana. Tools for determining the asymptotic spectral distribution of non-hermitian perturbations of hermitian matrix-sequences and applications. *Integral Equations and Operator Theory*, 81(2):213–225, 2015.

[33] 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.

[34] U. Grenander and G. Szegö. *Toeplitz forms and their applications*. Second Edition, Chelsea, New York, 1984.

[35] B. Gustafsson, H. Kreiss, and J. Oliger. *Time-Dependent Problems and Difference Methods*. Pure and Applied Mathematics: A Wiley Series of Texts, Monographs and Tracts. Wiley, 2013.

[36] R. A. Horn and C. R. Johnson. *Topics in Matrix Analysis*. Cambridge University Press, 1991. Cambridge Books Online.

[37] C. Johnson. *Numerical solution of partial differential equations by the finite element method*. Cambridge University Press, 1988.

[38] J. Kraus. Algebraic multilevel preconditioning of finite element matrices using local schur complements. *Numerical Linear Algebra with Applications*, 13(1):49–70, 2006.

[39] J. Kraus. Additive schur complement approximation and application to multilevel preconditioning. *SIAM Journal on Scientific Computing*, 34(6):A2872–A2895, 2012.

[40] B. Lund and J. O. Näslund. Glacial isostatic adjustment: implications for glacially induced faulting and nuclear waste repositories. In C. B. Connor, N. A. Chapman, and L. J. Connor, editors, *Volcanic and Tectonic Hazard Assessment for Nuclear Facilities*, pages 142–155. Cambridge University Press, 2009. Cambridge Books Online.

[41] M. Neytcheva. On element-by-element schur complement approximations. *Linear Algebra and its Applications*, 434(11):2308 – 2324, 2011. Special Issue: Devoted to the 2nd {NASC} 08 Conference in Nanjing (NSC).

[42] M. Neytcheva and E. Bängtsson. Preconditioning of nonsymmetric saddle point systems as arising in modelling of viscoelastic problems. *Electron. Trans. Numer. Anal.*, 29:193–211, 2007/08.

[43] E. Ngondiep, S. Serra-Capizzano, and D. Sesana. Spectral features and asymptotic properties for g-circulants and g-toeplitz sequences. *SIAM Journal on Matrix Analysis and Applications*, 31(4):1663–1687, 2010.

[44] S. Parter. On the eigenvalues of certain generalisations of toeplitz matrices. *Archive for Rational Mechanics and Analysis*, 11(1):244–257, 1962.

[45] V. D. Prete, F. D. Benedetto, M. Donatelli, and S. Serra-Capizzano. Symbol approach in a signal-restoration problem involving block toeplitz matrices. *Journal of Computational and Applied Mathematics*, 272(0):399 – 416, 2014.

[46] W. Rudin. *Real and complex analysis*. McGraw-Hill Education, 1974.

[47] Y. Saad. A flexible inner-outer preconditioned gmres algorithm. *SIAM Journal on Scientific Computing*, 14(2):461–469, 1993.

[48] Y. Saad. *Iterative methods for sparse linear systems*. Siam, 2003.

[49] Y. Saad and M. H. Schultz. Gmres: A generalized minimal residual algorithm for solving nonsymmetric linear systems. *SIAM Journal on Scientific and Statistical Computing*, 7(3):856–869, 1986.

[50] S. Serra-Capizzano. An ergodic theorem for classes of preconditioned matrices. *Linear Algebra and its Applications*, 282:161 – 183, 1998.

[51] S. Serra-Capizzano. Locally x matrices, spectral distributions, preconditioning, and applications. *SIAM Journal on Matrix Analysis and Applications*, 21(4):1354–1388, 2000.

[52] S. Serra-Capizzano. A note on the asymptotic spectra of finite difference discretizations of second order elliptic partial differential equations. *ASIAN JOURNAL OF MATHEMATICS*, 4:499–514, 2000.

[53] S. Serra-Capizzano. Spectral behavior of matrix sequences and discretized boundary value problems. *Linear Algebra and its Applications*, 337:37 – 78, 2001.

[54] S. Serra-Capizzano. Convergence analysis of two-grid methods for elliptic toeplitz and pdes matrix-sequences. *Numerische Mathematik*, 92(3):433–465, 2002.

[55] S. Serra-Capizzano. Test functions, growth conditions and toeplitz matrices. *Proceedings of the Fourth International Conference on Functional Analysis and Approximation Theory*, 2:791–795, 2002.

[56] S. Serra-Capizzano. Generalized locally toeplitz sequences: spectral analysis and applications to discretized partial differential equations. *Linear Algebra and its Applications*, 366(0):371 – 402, 2003. Special issue on Structured Matrices: Analysis, Algorithms and Applications.

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

[58] S. Serra-Capizzano and C. T. Possio. Spectral and structural analysis of high precision finite difference matrices for elliptic operators. *Linear Algebra and its Applications*, 293:85 – 131, 1999.

[59] S. Serra-Capizzano and C. T. Possio. Analysis of preconditioning strategies for collocation linear systems. *Linear Algebra and its Applications*, 369(0):41 – 75, 2003.

[60] S. Serra-Capizzano and C. Tablino-Possio. Multigrid methods for multilevel circulant matrices. *SIAM Journal on Scientific Computing*, 26(1):55–85, 2004.

[61] G. Strang. A proposal for toeplitz matrix calculations. *Stud. Appl. Math.*, 74(2):171–176, Apr. 1986.

[62] J. Strikwerda. *Finite Difference Schemes and Partial Differential Equations, Second Edition*. Society for Industrial and Applied Mathematics, 2004.

[63] P. Tilli. Locally toeplitz sequences: spectral properties and applications. *Linear Algebra and its Applications*, 278:91 – 120, 1998.

[64] P. Tilli. A note on the spectral distribution of toeplitz matrices. *Linear and Multilinear Algebra*, 45(2-3):147–159, 1998.

[65] P. Tilli. Some results on complex toeplitz eigenvalues. *Journal of Mathematical Analysis and Applications*, 239(2):390 – 401, 1999.

[66] E. Tyrtyshnikov and N. Zamarashkin. Spectra of multilevel toeplitz matrices: Advanced theory via simple matrix relationships. *Linear Algebra and its Applications*, 270:15 – 27, 1998.

[67] E. E. Tyrtyshnikov. A unifying approach to some old and new theorems on distribution and clustering. *Linear Algebra and its Applications*, 232(0):1 – 43, 1996.

[68] A. van der Sluis and H. van der Vorst. The rate of convergence of conjugate gradients. *Numerische Mathematik*, 48(5):543–560, 1986.

[69] P. Wu. Deformation of an incompressible viscoelastic flat earth with powerlaw creep: a finite element approach. *Geophysical Journal International*, 108(1):35–51, 1992.

Appendix A. Formal derivation of the symbol of the Schur complement in the Q2Q1 case via GLT

Since the Schur complement of $\mathcal{A}$ is of the form

$$\mathcal{S} = \rho M + \begin{bmatrix} B_1 & B_2 \end{bmatrix} K^{-1} \begin{bmatrix} B_1^T \\ B_2^T \end{bmatrix} = \rho M + H^T \begin{bmatrix} \tilde{B}_1 & \tilde{B}_2 \end{bmatrix} K^{-1} \begin{bmatrix} \tilde{B}_1^T \\ \tilde{B}_2^T \end{bmatrix} H$$

when deriving its symbol we have first to compute the symbol of $\tilde{B}K^{-1}\tilde{B}^T$ and then we have to take into consideration the effects of H^T and H, on the symbol of $\tilde{B}K^{-1}\tilde{B}^T$, $H = H_b \otimes H_b$.

Here we only consider the Q2Q1 case, where the structure of the underlying matrices is less trivial. When writing the above formula in terms of the blocks of K (see (29)), we face a difficulty since $\tilde{B}_j^T$ is, up to a low-rank term (cf. Remark 4.1), a two-level Toeplitz matrix of the type $T_n(\alpha_j) \otimes T_n(\beta_j)$ with α_j, β_j scalar valued functions, while every block K_{ij} has a rather complicated structure and has a 4×4 symbol, cf. Remark 4.4.

Therefore, when applying **GLT2**, it is not immediately seen that the symbol is expressed as sums and products of the symbols of K_{ij} and $\tilde{B}_l, \tilde{B}_m^T$, as the considered symbols have different sizes.

The natural solution turns out to be to consider the simpler structure of the blocks $\tilde{B}_l, \tilde{B}_m^T$, as a special case of the structure of the K_{ij} blocks, and to deduce the associated symbol as a 4×4 matrix of functions. In view of the latter, we anticipate that the concepts introduced in Remark 1.4 are central. This issue is the subject of Lemma 5.1.

Lemma 5.1. *Assume that X_N is a square matrix of size $N = n^2$, n even, such that*

$$X_N = T_n(\alpha) \otimes T_n(\beta)$$

with $\alpha, \beta \in L^1(Q)$, $Q = (-\pi, \pi)$. Then there exists a 4×4 matrix-valued generating function Φ belonging to $L^1(Q^2)$, such that

$$X_N = (I_n \otimes \Pi^T) T_{\frac{n}{2}\,\frac{n}{2}}(\Phi)(I_n \otimes \Pi) \tag{64}$$

with Π being the permutation matrix in Remark 4.4, with $\Phi(\theta_1, \theta_2)$ defined as

$$\Phi(\theta_1, \theta_2) = \alpha^{[2]}(\theta_2) \otimes \beta^{[2]}(\theta_1),$$

and where the relation between γ and $\gamma^{[2]}$, $\gamma \in \{\alpha, \beta\}$, refers to the notation introduced in Remark 1.4.

Proof. As a first step we recognize that the generic matrix X_N has the structure described in the relation (64) for some 4×4 matrix-valued generating function Φ belonging to $L^1(Q^2)$, cf. Remark 4.4. According to Definition 1.2, the matrix X_N is a block Toeplitz matrix of size $\frac{n}{2}$ with blocks A_j of size $2n$. In other words, X_N exhibits exactly the same structure as the blocks K_{ij}, even if every block K_{ij} has a simplified representation due to its tridiagonal character. Thus, we have

$$X_N = T_{\frac{n}{2}} = \begin{bmatrix} A_0 & A_{-1} & A_{-2} & \cdots & \cdots & A_{1-\frac{n}{2}} \\ A_1 & A_0 & A_{-1} & \ddots & & \vdots \\ A_2 & A_1 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & A_{-1} & A_{-2} \\ \vdots & & \ddots & A_1 & A_0 & A_{-1} \\ A_{\frac{n}{2}-1} & \cdots & \cdots & A_2 & A_1 & A_0 \end{bmatrix}, \tag{65}$$

where

$$A_l = \begin{bmatrix} T_{\frac{n}{2}}(\alpha^{11,l}) & T_{\frac{n}{2}}(\alpha^{12,l}) \\ T_{\frac{n}{2}}(\alpha^{21,l}) & T_{\frac{n}{2}}(\alpha^{22,l}) \end{bmatrix}$$

with A_l of size $2n$, and with $\alpha^{vw,l}$ being a 2×2 matrix-valued function, taking into account Definition 1.2 and Definition 1.3.

Next we treat the matrix $X_N = T_n(\alpha) \otimes T_n(\beta)$ as a member of the structure given in (65). To this end it is enough to prove the same statement separately for the two matrices

$$R_N = I_n \otimes T_n(\beta(\theta_1),\ L_N = T_n(\alpha(\theta_2) \otimes I_n,$$

since, by standard tensor algebra, $X_N = R_N L_N = L_N R_N$.

By a direct check, we see that R_N has the form reported in (65) with A_j null matrix for $j \neq 0$ and

$$A_0 = \begin{bmatrix} T_n(\beta) & \mathbf{0} \\ \mathbf{0} & T_n(\beta) \end{bmatrix}.$$

Finally, as n is even and, following the notation introduced in Remark 1.4, we have

$$A_0 = \begin{bmatrix} T_{\frac{n}{2}}(\beta^{[2]}) & \mathbf{0} \\ \mathbf{0} & T_{\frac{n}{2}}(\beta^{[2]}) \end{bmatrix}$$

so that the symbol of R_n is the 4×4 matrix-valued function $I_2 \otimes \beta^{[2]}(\theta_1)$. Now, denoting by α_j the Fourier coefficients of the univariate function $\alpha(\theta_2)$, we consider the matrix L_n which has the form

$$L_N = T_n = \begin{bmatrix} \alpha_0 I_n & \alpha_{-1} I_n & \alpha_{-2} I_n & \cdots & \cdots & \alpha_{1-n} I_n \\ \alpha_1 I_n & \alpha_0 I_n & \alpha_{-1} I_n & \ddots & & \vdots \\ \alpha_2 I_n & \alpha_1 I_n & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & A_{-1} & \alpha_{-2} I_n \\ \vdots & & \ddots & \alpha_1 I_n & A_0 & \alpha_{-1} I_n \\ \alpha_{n-1} I_n & \cdots & \cdots & \alpha_2 I_n & \alpha_1 I_n & \alpha_0 I_n \end{bmatrix}.$$

The latter can be viewed as a matrix of the form reported in (65), by defining

$$A_j = \begin{bmatrix} \alpha_{2j} I_n & \alpha_{2j-1} I_n \\ \alpha_{2j-1} I_n & \alpha_{2j} I_n \end{bmatrix} = \tilde{\alpha}_j \otimes I_n = (\tilde{\alpha}_j \otimes I_2) \otimes I_{\frac{n}{2}}$$

By noticing that the 2×2 block $\tilde{\alpha}_j$ is the j-th Fourier coefficient of the function $\alpha^{[2]}(\theta_2)$, cf. Remark 1.4, we infer that the symbol of L_N is the 4×4 matrix-valued function $\alpha^{[2]}(\theta_2) \otimes I_2$.

Finally, since the blocks of the matrices L_N and R_N commute, the global symbol of $X_N = R_N L_N = L_N R_N$ is exactly $\Phi(\theta_1, \theta_2) = \alpha^{[2]}(\theta_2) \otimes \beta^{[2]}(\theta_1)$. ∎

Now we are ready to derive the symbol $G = \tilde{B} K^{-1} \tilde{B}^T$ in the Q2Q1 case.

Proposition 5.2. *Taking into account the relevant notations from Section 4.3.1, we prove formula (63), namely,*

$$\begin{aligned} G(\theta_1, \theta_2) &= \left[g_{\widetilde{B}_1} [f^{K_{11}}]^{-1} g_{\widetilde{B}_1^T} + f^R [f^{\mathcal{S}_K}]^{-1} [f^R]^T \right] (\theta_1, \theta_2), \\ f^R(\theta_1, \theta_2) &= g_{\widetilde{B}_2}(\theta_1, \theta_2) - g_{\widetilde{B}_1}(\theta_1, \theta_2) [f^{K_{11}}]^{-1}(\theta_1, \theta_2) f^{K_{12}}(\theta_1, \theta_2), \\ f^{\mathcal{S}_K}(\theta_1, \theta_2) &= f^{K_{22}}(\theta_1, \theta_2) - f^{K_{21}}(\theta_1, \theta_2) [f^{K_{11}]^{-1}}(\theta_1, \theta_2) f^{K_{12}}(\theta_1, \theta_2), \end{aligned}$$

with $0 < \theta_1, \theta_2 \leq \pi$ *and* $g_{\widetilde{B}_j}(\theta_1, \theta_2) = \alpha_j^{[2]}(\theta_1) \otimes \beta_j^{[2]}(\theta_2),\ j = 1, 2.$

Proof. We start by looking carefully at formula (29), by taking into consideration the structure of the B blocks. Then we have

$$\mathcal{S} = \rho M + \begin{bmatrix} B_1 & B_2 \end{bmatrix} K^{-1} \begin{bmatrix} B_1^T \\ B_2^T \end{bmatrix} = \rho M + H^T \begin{bmatrix} \tilde{B}_1 & \tilde{B}_2 \end{bmatrix} K^{-1} \begin{bmatrix} \tilde{B}_1^T \\ \tilde{B}_2^T \end{bmatrix} H.$$

Therefore, using items **GLT1**, **GLT2**, **GLT3**, and observing that the involved blocks are of Toeplitz type up to low rank corrections (see Remark 4.1) and up to the very same permutation matrix, the symbol of $\tilde{B}K^{-1}\tilde{B}^T = \begin{bmatrix} \tilde{B}_1 & \tilde{B}_2 \end{bmatrix} K^{-1} \begin{bmatrix} \tilde{B}_1^T \\ \tilde{B}_2^T \end{bmatrix}$ is obtained as $v^*(f^K)^{-1}v$ with the vector v such that $v_1 = f^{\tilde{B}_1}$ and $v_2 = f^{\tilde{B}_2}$: setting $G = f^{\tilde{B}K^{-1}\tilde{B}^T}$ and taking into account Lemma 5.1, all the symbols in the expression of $v^*(f^K)^{-1}v$ are 4×4 matrix-valued so that the proposition follows. ∎

Appendix B. The role of the downsampling matrix: Q1isoQ1 vs Q2Q1

This appendix presents details and formal proofs on the role of the two-dimensional cutting matrix $H = H_b \otimes H_b$ applied from the left(transposed) and from the right to the matrix product $BK^{-1}B^T$. From algebraic point of view, the action of H is well understood, (see Figures 8-10). From a functional point of view the matter is studied in a multigrid context for the convergence analysis of the two-grid and of the V-cycle methods, in the case of matrix algebras [28, 27, 54, 60]. There, the matrix H is referred to as the 'cutting matrix'.

Here, in order to prove (46), we give details on the functional aspect. First we consider the case Q1isoQ1 and then the case Q2Q1, by taking into account Remark 4.1.

Lemma 5.3. *Assume that X_N is a square matrix of size $N = n^2$, n even, such that there exists a scalar-valued generating function f belonging to $L^1(Q^2)$, $Q = (-\pi, \pi)$, for which*

$$X_N = T_{nn}(f).$$

Consider H_b to be the unilevel downsampling matrix of size $n \times \frac{n}{2}$ and define

$$Y_n = (H_b^T \otimes H_b^T) X_n (H_b \otimes H_b).$$

Then $\{Y_n\}$ is a GLT sequence with symbol

$$\widehat{f} = \frac{1}{4} \left(\sum_{l=0}^{1} \sum_{m=0}^{1} f\left(\frac{\theta_1}{2} + l\pi, \frac{\theta_2}{2} + m\pi \right) \right) \tag{66}$$

and, indeed,

$$Y_n = T_{\frac{n}{2}\frac{n}{2}}(\widehat{f}). \tag{67}$$

Proof. As a first step we take $\gamma(\theta)$ to be a univariate scalar-valued function and consider

$$T_n(\gamma) = \begin{bmatrix} a_0 & a_{-1} & a_{-2} & \cdots & \cdots & a_{1-n} \\ a_1 & a_0 & a_{-1} & \ddots & & \vdots \\ a_2 & a_1 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & a_{-1} & a_{-2} \\ \vdots & & \ddots & a_1 & a_0 & a_{-1} \\ a_{n-1} & \cdots & \cdots & a_2 & a_1 & a_0 \end{bmatrix},$$

where a_j are the Fourier coefficients of γ, $j \in \mathbb{Z}$, cf. Definition 1.3.

By direct computation we obtain

$$T_{\frac{n}{2}} = H^T T_n(\gamma) H = \begin{bmatrix} a_0 & a_{-2} & a_{-4} & \cdots & \cdots & a_{2-n} \\ a_2 & a_0 & a_{-2} & \ddots & & \vdots \\ a_4 & a_2 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & a_{-2} & a_{-4} \\ \vdots & & \ddots & a_2 & a_0 & a_{-2} \\ a_{n-2} & \cdots & \cdots & a_4 & a_2 & a_0 \end{bmatrix}, \tag{68}$$

and observe that only the even Fourier coefficients of γ remain but with a shifted position. Indeed, the coefficient a_{2j}, $|2j| \leq n-1$, instead of appearing in the $2j$-th diagonal, is found on the j-th diagonal. Such an algebraic operation has a functional counterpart that has been extensively exploited in a multigrid context (see [28, 27, 54, 60, 1]). Therefore, the matrix $T_{\frac{n}{2}}$ in (68) is such that

$$T_{\frac{n}{2}} = H_b^T T_n(\gamma) H_b = T_{\frac{n}{2}}(\widehat{\gamma}), \quad \widehat{\gamma}(\theta) = \frac{1}{2}\left(\gamma\left(\frac{\theta}{2}\right) + \gamma\left(\frac{\theta}{2} + \pi\right)\right). \tag{69}$$

Now we consider the two-level case, by taking a bivariate trigonometric polynomial f so that

$$f(\theta_1, \theta_2) = \sum_{i=0}^{q_1} \sum_{j=0}^{q_2} \phi_i(\theta_1)\psi_j(\theta_2).$$

Setting $r_{ij}(\theta_1, \theta_2) = \phi_i(\theta_1)\psi_j(\theta_2)$, by exploiting basic properties of the tensor calculus, and taking into account (69), we deduce the following chain of identities

$$\begin{aligned} (H_b^T \otimes H_b^T) T_{nn}(r_{ij})(H_b \otimes H_b) &= (H_b^T \otimes H_b^T)(T_n(\psi_j) \otimes T_n(\phi_i))(H_b \otimes H_b) \\ &= (H_b^T T_n(\psi_j) H_b) \otimes (H_b^T T_n(\phi_i) H_b) \\ &= T_{\frac{n}{2}}(\widehat{\psi}_j) \otimes T_{\underline{n}}(\widehat{\phi}_i) \\ &= T_{\frac{n}{2}\frac{n}{2}}(\widehat{r}_{ij}), \end{aligned}$$

with $\widehat{r}_{ij}(\theta_1,\theta_2)=\widehat{\phi}_i(\theta_1)\widehat{\psi}_j(\theta_2)$. We observe that

$$\widehat{r}_{ij}(\theta_1,\theta_2)=\frac{1}{4}\left(\sum_{l=0}^{1}\sum_{m=0}^{1} r_{ij}\left(\frac{\theta_1}{2}+l\pi,\frac{\theta_2}{2}+m\pi\right)\right),$$

which coincides with formula (66) in the specific case of a separable symbol.

Furthermore, the argument of linearity leads to the desired formula when $f(\theta_1,\theta_2)$ is a generic bivariate trigonometric polynomial, while a density argument in the $L^1(Q^2)$ Banach space equipped with its natural L^1 norm leads to the desired formula when $f(\theta_1,\theta_2)\in L^1(Q^2)$.

Finally, because of (67) and **GLT3**, $\{Y_n\}$ is a GLT sequence with symbol $\widehat{f}$. ∎

Lemma 5.4. *Assume that k is a given positive number and $X_N^{(1)},\ldots,X_N^{(k)}$ are square matrices of size $N=n^2$, n even, such that there exist scalar-valued generating functions $f^{(1)},\ldots,f^{(k)}$ belonging to $L^1(Q^2)$, $Q=(-\pi,\pi)$, for which*

$$X_N^{(j)}=T_{nn}(f^{(j)})+E_N^{(j)},$$

with $\{E_N^{(j)}\}\sim_\sigma 0$, $j=1,\ldots,k$.

Consider $Z_N=g(X_N^{(1)},\ldots,X_N^{(k)})$ where g is obtained via a finite number of elementary operations on the input matrices, involving only operations such as summation, multiplication, multiplication by a scalar and inversion (if the symbol associated to the inverted sequence of matrices has a set of zeros of at most zero Lebesgue measure). Then $\{Z_N\}$ is a GLT sequence with a symbol $g=g(f^{(1)},\ldots,f^{(k)})$.

Let H_b be the unilevel downsampling matrix of size $n\times\frac{n}{2}$ and define

$$Y_n=(H_b^T\otimes H_b^T)Z_N(H_b\otimes H_b).$$

Then $\{Y_n\}$ is a GLT sequence with symbol

$$\widehat{g}=\frac{1}{4}\left(\sum_{l=0}^{1}\sum_{m=0}^{1} g\left(\frac{\theta_1}{2}+l\pi,\frac{\theta_2}{2}+m\pi\right)\right).$$

Proof. Since $\{E_N^{(j)}\}\sim_\sigma 0$, it is also a GLT sequence with identically zero symbol, by **GLT3**. Again by **GLT3**, $\{T_{nn}(f^{(j)})\}$ is GLT sequence with symbol $f^{(j)}$, $j=1,\ldots,k$. Now $X_N^{(j)}=T_{nn}(f^{(j)})+E_N^{(j)}$ by the assumptions of the lemma and then, by invoking item **GLT2**, $\{X_N^{(j)}\}$ is GLT sequence with symbol $f^{(j)}$, $j=1,\ldots,k$. By the same argument contained in **GLT2** we deduce the first part of the thesis, i.e., $\{Z_N\}$ is a GLT sequence with symbol $g=g(f^{(1)},\ldots,f^{(k)})$.

The rest of the lemma can be proven by following the very same steps as in Lemma 5.3 and so we omit the details. ∎

Proposition 5.5. *Taking into account the relevant notations from Section 4.2.1, we prove formula (46).*

Proof. Taking into account formula (45) and the explicit derivation of the symbol G in Section 4.2.1, we deduce that (46) is obtained as a special case of Lemma 5.4. ∎

We next prove formula (62) in the Q2Q1 case.

Lemma 5.6. *Assume that X_N is a square matrix of size $N = n^2$, n even, such that there exists a 4×4 matrix-valued generating function Φ belonging to $L^1(Q^2)$, $Q = (-\pi, \pi)$, for which*

$$X_N = (I_n \otimes \Pi^T) T_{\frac{n}{2}\frac{n}{2}}(\Phi)(I_n \otimes \Pi)$$

with Π being the permutation matrix in Remark 4.4. Let H_b be the unilevel cutting matrix of size $n \times \frac{n}{2}$ and define

$$Y_n = (H_b^T \otimes H_b^T) X_N (H_b \otimes H_b).$$

Then $\{\Pi^T X_N(\Phi)\Pi\}$ is a GLT sequence with symbol Φ and $\{Y_n\}$ is a GLT sequence with symbol Φ_{11}.

Proof. As already emphasized in Appendix **A**, in (65), the structure of X_N is such that

$$X_N = T_{\frac{n}{2}} = \begin{bmatrix} A_0 & A_{-1} & A_{-2} & \cdots & \cdots & A_{1-\frac{n}{2}} \\ A_1 & A_0 & A_{-1} & \ddots & & \vdots \\ A_2 & A_1 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & A_{-1} & A_{-2} \\ \vdots & & \ddots & A_1 & A_0 & A_{-1} \\ A_{\frac{n}{2}-1} & \cdots & \cdots & A_2 & A_1 & A_0 \end{bmatrix},$$

with

$$A_l = \begin{bmatrix} T_{\frac{n}{2}}(\alpha^{11,l}) & T_{\frac{n}{2}}(\alpha^{12,l}) \\ T_{\frac{n}{2}}(\alpha^{21,l}) & T_{\frac{n}{2}}(\alpha^{22,l}) \end{bmatrix}$$

and with A_l of size $2n$, with $\alpha^{vw,l}$ being a 2×2 matrix-valued function, taking into account Definition 1.2 and Definition 1.3.

Now consider $Y_n = (H_b^T \otimes H_b^T) X_N (H_b \otimes H_b)$: we make use of basic features and tensor calculus and we obtain

$$Y_n = (I_{\frac{n}{2}} \otimes H_b^T)(H_b^T \otimes I_n) X_N (H_b \otimes I_n)(I_{\frac{n}{2}} \otimes H_b).$$

The matrix $\tilde{X}_n = (H_b^T \otimes I_n) X_N (H_b \otimes I_n)$ has exactly the same structure as X_N but in each block A_l of size $2n$ the last n rows and the last n columns are deleted. Therefore

$$\tilde{X}_N = T_{\frac{n}{2}} = \begin{bmatrix} \tilde{A}_0 & \tilde{A}_{-1} & A_{-2} & \cdots & \cdots & \tilde{A}_{1-\frac{n}{2}} \\ \tilde{A}_1 & \tilde{A}_0 & \tilde{A}_{-1} & \ddots & & \vdots \\ \tilde{A}_2 & \tilde{A}_1 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \tilde{A}_{-1} & \tilde{A}_{-2} \\ \vdots & & \ddots & \tilde{A}_1 & \tilde{A}_0 & \tilde{A}_{-1} \\ \tilde{A}_{\frac{n}{2}-1} & \cdots & \cdots & \tilde{A}_2 & \tilde{A}_1 & \tilde{A}_0 \end{bmatrix},$$

with

$$\tilde{A}_l = T_{\frac{n}{2}}(\alpha^{11,l}).$$

Here we remind that $T_{\frac{n}{2}}(\alpha^{11,l})$ is square matrix of size n since each function $\alpha^{11,l}$ is 2×2 matrix-valued. Now we consider $Y_n = (I_{\frac{n}{2}} \otimes H_b^T)\tilde{X}_N(I_{\frac{n}{2}} \otimes H_b)$ which has the effect that every block $T_{\frac{n}{2}}(\alpha^{11,l})$ is transformed into the new block of size $\frac{n}{2}$ given by $H_b^T T_{\frac{n}{2}}(\alpha^{11,l})H_b = T_{\frac{n}{2}}([\alpha^{11,l}]_{11})$.

In conclusion, the matrix Y_n is a standard two-level Toeplitz matrix generated by a scalar-valued function, namely the component $(1,1)$ of the 4×4 matrix-valued generating function Φ, which completes the proof. ∎

Lemma 5.7. *Assume that k is a given positive number and $X_N^{(1)}, \ldots, X_N^{(k)}$ are square matrices of size $N = n^2$, n even, such that there exist 4×4 matrix-valued generating functions $\Phi^{(1)}, \ldots, \Phi^{(k)}$ belonging to $L^1(Q^2)$, $Q = (-\pi, \pi)$, for which*

$$X_N^{(j)} = \widehat{X}_N^{(j)} + E_N^{(j)}, \qquad \widehat{X}_N^{(j)} = (I_n \otimes \Pi^T)T_{\frac{n}{2}\frac{n}{2}}(\Phi^{(j)})(I_n \otimes \Pi),$$

with $\{E_N^{(j)}\} \sim_\sigma 0$, $j = 1, \ldots, k$, and with Π being the permutation matrix in Remark 4.4.

Consider $Z_N = g(X_N^{(1)}, \ldots, X_N^{(k)})$ where g is obtained via a finite number of elementary operations on the input matrices, involving only operations such as summation, multiplication, multiplication by a scalar and inversion (if the symbol associated to the inverted sequence of matrices is singular on a set of at most zero Lebesgue measure). Then $\{(I_n \otimes \Pi)Z_N(I_n \otimes \Pi^T)\}$ is a GLT sequence with symbol $g(\Phi^{(1)}, \ldots, \Phi^{(k)})$.

Let H_b to be the unilevel downsampling matrix of size $n \times \frac{n}{2}$ and define

$$Y_n = (H_b^T \otimes H_b^T)Z_N(H_b \otimes H_b).$$

Then $\{Y_n\}$ is a GLT sequence with symbol $[g(\Phi^{(1)}, \ldots, \Phi^{(k)})]_{11}$.

Proof. Since $\{E_N^{(j)}\} \sim_\sigma 0$, it is also a GLT sequence with identically zero symbol, by **GLT3**. Again by **GLT3**, $\{T_{\frac{n}{2}\frac{n}{2}}(\Phi^{(j)})\}$ is GLT sequence with symbol $\Phi^{(j)}$, $j = 1, \ldots, k$. Now $(I_n \otimes \Pi)X_N^{(j)}(I_n \otimes \Pi^T) = T_{nn}(f^{(j)}) + (I_n \otimes \Pi)E_N^{(j)}(I_n \otimes \Pi^T)$ by the assumptions of the lemma and then, by recalling that a similarity transformation by a permutation does change the singular values invoking **GLT2**, we infer that $\{(I_n \otimes \Pi)X_N^{(j)}(I_n \otimes \Pi^T)\}$ is GLT sequence with symbol $f^{(j)}$, $j = 1, \ldots, k$. By the same argument contained in **GLT2** we deduce the first part of the thesis, i.e., $\{(I_n \otimes \Pi)Z_N(I_n \otimes \Pi^T)\}$ is a GLT sequence with symbol $g = g(f^{(1)}, \ldots, f^{(k)})$.

The rest of the lemma can be proven exactly as in Lemma 5.6. ∎

Proposition 5.8. *Taking into account the relevant notations of Section 4.3.1, we prove formula (62).*

Proof. Taking into account the explicit derivation of the symbol G in Section 4.3.1 and Lemma 5.2, we deduce that (62) is obtained as a special case of Lemma 5.7. ∎