Numerische Mathematik (2021) 149:933–971
https://doi.org/10.1007/s00211-021-01247-y

Numerische Mathematik

# A matrix-theoretic spectral analysis of incompressible Navier–Stokes staggered DG approximations and a related spectrally based preconditioning approach

**M. Mazza**[1] · **M. Semplice**[2] · **S. Serra-Capizzano**[1] · **E. Travaglia**[3]

Received: 10 May 2021 / Revised: 8 September 2021 / Accepted: 2 October 2021 /
Published online: 15 November 2021


**Abstract**
The incompressible Navier–Stokes equations are solved in a channel, using a Discontinuous Galerkin method over staggered grids. We study the structure and the spectral features of the matrices of the linear systems arising from the discretization. They are of block type, each block showing Toeplitz-like, band, and tensor structure at the same time. After introducing new tools to study Toeplitz-like matrix sequences with rectangular symbols, a quite complete spectral analysis is presented, with the target of designing and analyzing fast iterative solvers for the associated large linear systems. Promising numerical results are presented, commented, and critically discussed for elongated two- and three-dimensional geometries.

**Mathematics Subject Classification** 65F08 · 65N30 · 15B05 · 15A18

✉ E. Travaglia
elena.travaglia@unito.it

M. Mazza
mariarosa.mazza@uninsubria.it

M. Semplice
matteo.semplice@uninsubria.it

S. Serra-Capizzano
s.serracapizzano@uninsubria.it

[1] Dipartimento di Scienza Umane e dell'Innovazione per il Territorio, Università dell'Insubria, Via Valleggio, 11, 22100 Como, Italy

[2] Dipartimento di Scienza e Alta Tecnologia, Università dell'Insubria, Via C. Alberto, 8, Torino, Italy

[3] Dipartimento di Matematica, Università di Torino, Via C. Alberto, 8, Torino, Italy

## 1 Introduction

The discretization of incompressible Navier–Stokes equations leads to large saddle point linear systems, whose matrices take the block form $\left[\begin{smallmatrix} N & G \\ D & E \end{smallmatrix}\right]$, where $N$ contains the contribution of the mass and viscous terms, the (typically rectangular) blocks $G$ and $D$ represent the pressure gradient and velocity divergence terms and $E$ is either zero or a small stabilization term. The efficient solution of the linear system can be achieved by considering the Schur complement, with matrix $S = E - DN^{-1}G$. The matrix $S$ cannot be easily assembled due to the $N^{-1}$ term and thus the solution of the Schur complement system should leverage on iterative methods and preconditioners that do not rely on an assembled matrix, and hence they cannot include standard circulant preconditioners and fast transform algebra preconditioners (see [9,13,23,28, 31] and references therein). Their design thus requires a precise knowledge of the spectrum of $S$.

Here we focus on the smallest nontrivial dimensional setting, i.e. the case of a flow between two infinite (not necessarily straight nor parallel) plates and the flow in a pipe with variable cross-section. These flows are of particular interest in industrial, but also biomedical applications. When the geometry does not change abruptly, the flow is laminar and no recirculation is observed. A full three-dimensional solver, with a refined mesh in all three spatial directions would lead to very large problems to be solved, which seem disproportioned to the need of resolving an essentially one-dimensional flow. These cases are instead amenable for quasi-1D discretizations, in which the pipe is described attaching a cross-section to each point of a 1D object. Such a solver is surely faster than a fully three-dimensional one and much more amenable for tasks such as digital twins of industrial machinery, optimization problems for the design phase, fast creation of databases for Proper Orthogonal Decomposition (POD) methods, etc.

The only quasi-1D approach that is known to us are the Transversally Enriched Pipe Element Method of [26] and the discretization methods at the base of the hierarchical model reduction techniques of [22]. Both of them compute a three-dimensional flow in a domain that is discretized only along the axial coordinate, i.e. the elements are sections of the whole pipe of length $\Delta$x. The finite element bases are obtained by Cartesian product of different discretizations in the longitudinal and in the transversal directions. This approach is quite useful for example in industrial applications, where one often needs a fast solver to be used in shape-optimization problems.

In this work we study a further simplification of the model, in which the transversal velocity components are neglected and only the longitudinal velocity is considered. In particular we consider the incompressible Navier–Stokes equations

$$\rho\left(\frac{\partial \mathbf{u}}{\partial t} + \nabla \cdot F_c\right) = -\nabla p + \nabla \cdot (\mu\, \nabla \mathbf{u}) \tag{1a}$$

$$\nabla \cdot \mathbf{u} = 0, \tag{1b}$$

where $\mathbf{x} = (x, y, z)$ is the vector of spatial coordinates and $t$ denotes the time, $p$ is the physical pressure and $\rho$ is the constant fluid density and $\mu$ is the viscosity which is a constant function if we consider a newtonian fluid. $F_c = \mathbf{u} \otimes \mathbf{u}$ is the flux tensor

of the nonlinear convective terms, $\mathbf{u} = (u, v, w)$ is the velocity vector where $u$ is the component parallel to the pipe axis, while $v$ and $w$ are the transversal ones.

We consider as domain a pipe with a variable cross-section and since it has a length much greater than the section, we neglect the transverse velocities, i.e. we assume $v = w = 0$ (and consequently also $\partial_y p = \partial_z p = 0$), but we consider the dependence on the three spatial variables of the longitudinal component, i.e. $u = u(x, y, z)$. The discretization is then performed with Discontinous Galerkin (DG) methods on a staggered grid arrangement, i.e. velocity elements are dual to the main grid of the pressure elements, similarly to [15,36,37], leading to a saddle point problem for the longitudinal velocity and the pressure variables.

Having in mind the efficient solution of the arising linear systems, in this paper we focus on the spectral study of the coefficient matrix as well as of its blocks and Schur complement. More specifically, we first recognize that all the matrix coefficient blocks show a block Generalized Locally Toeplitz (GLT) structure and that, as such, can be equipped with a symbol. Second, we leverage on the symbols of the blocks to retrieve the symbol of the Schur complement and the symbol of the coefficient matrix itself. We stress that in order to accomplish these goals, we introduce some new spectral tools that ease the symbol computation when rectangular matrices are involved, and we provide an alternative to the somehow cumbersome approach followed in [12] of embedding into square matrices first and then of projecting by downsampling matrices borrowed from the context of multigrid methods for Toeplitz matrices as in [16,31] and references therein.

In this setting we design a block circulant preconditioner for the Schur complement that provides a constant number of iterations as the matrix-size increases and that, once nested into a Krylov-type solver for the original coefficient matrix, brings to lower CPU timings when compared with the least squares commutators proposed in [35]. Our results confirm the potential of a symbol-based approach in the context of Discontinuous Galerkin discretizations on staggered grids as foreseen in [13].

The paper is organized as follows. In Sect. 2 we describe in details the discretization of the quasi-1D incompressible Navier-Stokes model; in Sect. 3 we both recall the Toeplitz and GLT technology and we introduce some new spectral tools that will be used in Sect. 4 to perform the spectral analysis of the matrix of the saddle point problem. This leads to the proposal of an efficient optimal preconditioner for our system, which is tested in the numerical Sect. 5.

## 2 Discretization

We consider the incompressible Navier–Stokes equations (1) in an elongated pipe-like domain, with a variable cross-section. An example is depicted in Fig. 1. We impose a no-slip condition at the solid boundaries; at the outlet boundary we fix a null pressure, while at the inlet we impose Dirichlet data with a given velocity profile.

The channel is discretized only along its longitudinal dimension, so each cell is a section of the entire pipe of length $\Delta$x, (see Fig. 1). We denote the cells in this grid by $\Omega_1, \ldots, \Omega_n$. The discrete pressure is defined on this grid, while for the velocity we use a dual grid, whose first and last element have length equal to one half of the other

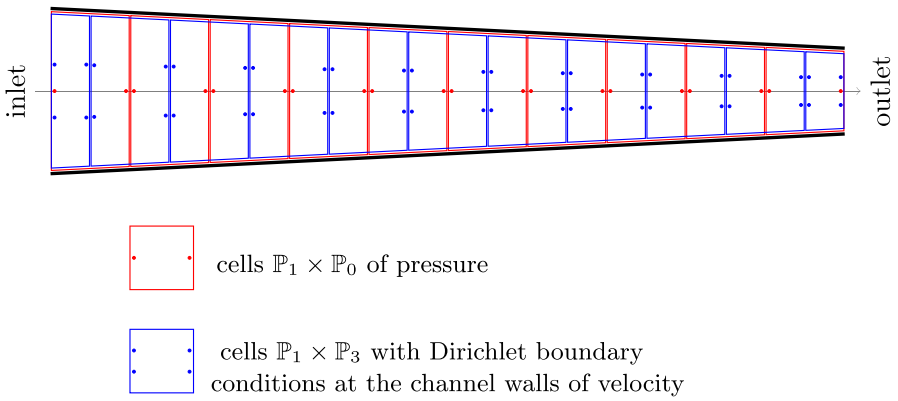



**Fig. 1** Illustration of the staggered grid arrangement in a nozzle for $n_x = 1$ and $n_y = 3$

cells. This type of staggered grid has been employed for example in [15,36,37]. We denote the cells of the dual grid by $\Omega_1^*, \ldots, \Omega_{n+1}^*$ and point out that each $\Omega_j$ has a nontrivial intersection only with $\Omega_j^*$ and $\Omega_{j+1}^*$ for $j = 1 \ldots n$.

For ease of presentation, we concentrate mainly on the two-dimensional case and denote the width of the channel at the position $x$ by $d(x)$. The longitudinal velocity $u = u(x, y)$, in each cell of the dual grid, is approximated by a $\mathbb{P}_{n_x} \otimes \mathbb{P}_{n_y}$ polynomial defined as the tensor product of the one dimensional polynomial of degree $n_x$ in the longitudinal direction and $n_y$ in the transverse one. In order to do this, we construct a polynomial basis on the standard reference elements, $\Omega_{ref} = [0, 1]^2$, using the Lagrange interpolation polynomials with equispaced nodes. Taking into account the no-slip boundary condition applied at the channel walls, there are $n_u := (n_x + 1) \times (n_y - 1)$ effective degrees of freedom for $u$ in each cell (blue dots in Fig. 1). We stress that in order to satisfy the no-slip boundary conditions one should take $n_y \geq 2$.

In the same way the pressure is approximated in each cell of the primal grid by a $\mathbb{P}_{n_x} \otimes \mathbb{P}_0$ polynomial, i.e. the pressure is constant in the transversal direction. For this reason, there are only $n_p := (n_x + 1)$ degrees of freedom for $p$ in each cell (red dots in Fig. 1). In general we are interested in a low degree $n_x$ but high degrees $n_y$, which are needed to compensate for the lack of mesh discretization in the transversal direction, and of course a mild but generic dependence of $d$ upon $x$.

To obtain a DG discretization on the staggered cell arrangements, we first integrate the momentum equation (1a) multiplied by a generic shape function $\psi$ for the velocity over a cell of the dual grid, $\Omega_i^*$, for $i = 1 \ldots n + 1$,

$$\int_{\Omega_i^*} \psi \, \rho \left( \frac{\partial u}{\partial t} + \nabla \cdot F_c \right) d\mathbf{x} = - \int_{\Omega_i^*} \psi \, \nabla p \, d\mathbf{x} + \int_{\Omega_i^*} \psi \, \nabla \cdot (\mu \, \nabla u) d\mathbf{x}. \quad (2a)$$

We then integrate the continuity equation (1b), multiplied by a generic shape function $\theta$ for the pressure, over a cell of the primal grid, $\Omega_j$ for $j = 1 \ldots n$

$$\int_{\Omega_j} \theta \, \nabla \cdot u \, d\mathbf{x} = 0, \tag{2b}$$

where $d\mathbf{x} = dxdy$.

Integrating by parts the viscous term in (2a), we must take into account that velocity at intercell boundaries is discontinuous and it is necessary to penalize the jumps in order to achieve a stable discretization. In the following we will adopt the standard notations for the average $\{\cdot\}$ and the jump $[\![\cdot]\!]$ of a vector function $\phi$, defined as

$$\{\phi\} = \tfrac{1}{2}\left(\phi_i + \phi_{i+1}\right), \qquad [\![\phi]\!] = \phi_i \cdot \mathbf{n}_i + \phi_{i+1} \cdot \mathbf{n}_{i+1},$$

where $\phi_i$ denotes the function value in cell $i$ and $\mathbf{n}_i$ the outgoing normal. We associate the bilinear form with the following viscous term

$$\begin{aligned} B(u,\psi) = \int_{\Omega_i^*} \mu \nabla u \cdot \nabla \psi \, d\mathbf{x} + \epsilon \int_{\partial\Omega_i^*} \mu [\![u]\!] \cdot \{\nabla \psi\} \, d\Gamma - \\ - \int_{\partial\Omega_i^*} \{\nabla u\} \cdot [\![\psi]\!] \, d\Gamma + \int_{\partial\Omega_i^*} \alpha\mu [\![u]\!][\![\psi]\!] \, d\Gamma, \end{aligned} \tag{3}$$

where $\alpha = \frac{\alpha_0}{\Delta x}$ is the penalization [1]. Changing the sign of $\epsilon$ we obtain symmetric (SIP) [42] and non-symmetric Interior Penalty (NIP) method [30]. In the first case the velocity jump term for the mean of the test function is subtracted in the bilinear form, so $\epsilon = -1$, while in the second method it is added. Following to [1], the bilinear form $B$ is coercive $\forall \alpha_0 > 0$ in the NIP case and for $\alpha_0 > \hat{\alpha} > 0$, for some $\hat{\alpha}$ in the SIP case. The estimation of $\hat{\alpha}$ is in general a nontrivial task, but the advantage of SIP is that the resulting matrix is symmetric and positive definite. Due to the advantage properties of SIP we discretize the viscosity term with this method and for all the test in this article we choose $\alpha_0 = 1$.

The integrand of the pressure term in (2a) contains a discontinuity since the pressure is defined on the primal grid and is thus not continuous on the dual velocity cells. The pressure integral is then split as follows:

$$\int_{\Omega_i^*} \psi \, \nabla p \, d\mathbf{x} = \int_{\Omega_i^* \cap \Omega_{i-1}} \psi \, \nabla p \, d\mathbf{x} + \int_{\Omega_i^* \cap \Omega_i} \psi \, \nabla p \, d\mathbf{x} + \int_{\Gamma_i} \psi \, (p_i - p_{i-1}) d\Gamma, \tag{4}$$

where $p_{i-1}$ and $p_i$ denote the discrete pressure in the cells $\Omega_{i-1}$ and $\Omega_i$ respectively and $\Gamma_i$ is the interface between $\Omega_{i-1}$ and $\Omega_i$, which is located in the middle of $\Omega_i^*$.

A similar difficulty appears in (2b), since the discrete velocity is discontinuous on pressure elements, and this is circumvented by computing the divergence term as

$$\begin{aligned} \int_{\Omega_j} \theta \, \nabla \cdot u \, d\mathbf{x} = \int_{\Omega_j \cap \Omega_j^*} \theta \, \nabla \cdot u \, d\mathbf{x} \\ + \int_{\Omega_j \cap \Omega_{j+1}^*} \theta \, \nabla \cdot u \, d\mathbf{x} + \int_{\Gamma_j^*} \theta \, (u_{j+1} - u_j) d\Gamma. \end{aligned} \tag{5}$$

Here above, $\Gamma_j$ denotes the interface between $\Omega_j^*$ and $\Omega_{j+1}^*$, which is located in the middle of $\Omega_j$.

Further, for stability, a penalty term must be added to the discretized continuity equation (2b) due to the choice of a discontinuous approximation for pressure [24]. Equation (2b) is thus modified adding the term

$$\int_{\Gamma_j} \alpha [\![p]\!][\![\theta]\!] \, d\Gamma \tag{6}$$

where the penalization constant is $\alpha = \Delta \mathrm{x}$. Without this additional term, pressure oscillations that grow as $\Delta \mathrm{x} \to 0$ would appear at the cell interfaces of the main grid.

The left hand side of (2a) gives rise to a mass matrix term and to a convective term that depends nonlinearly on $u$. By considering in (2) an implicit discretization for all terms except for the nonlinear convective term, one obtains a linear system for the velocity and pressure unknowns at time $t^{n+1}$ that has the following block structure

$$\mathcal{A}\mathbf{x} = \mathbf{f} \iff \begin{bmatrix} N & G \\ D & E \end{bmatrix} \begin{pmatrix} u \\ p \end{pmatrix} = \begin{pmatrix} b_u(u) \\ 0 \end{pmatrix}. \tag{7}$$

In the previous formula, $N = M + L$ is a square matrix formed by $L$ and $M$ that discretize the Laplacian and the mass operator, respectively. The dimension of the blocks depends on the number cells and on the degrees of freedom chosen to discretize the velocity and the pressure. In particular, the size of $M$ and $L$ is $(n+1)(n_u \times n_u)$ and the elements of the matrices are $\mathcal{O}(1)$ and $\mathcal{O}(\Delta \mathrm{x})$, respectively. $G$ is a rectangular tall matrix of size $(n+1)(n_u \times n_p)$ corresponding to the gradient operator (4), whose elements are $\mathcal{O}(\Delta \mathrm{t})$; $D$, coming from (5), is the transpose of $G$, up to a scaling factor, and its elements are $\mathcal{O}(1)$. Finally, $E$ is a square matrix of size $n$ $(n_p \times n_p)$ containing the penalty term (6), with elements of size $\mathcal{O}(\Delta \mathrm{x})$. In fact, technically speaking, it is worth mentioning that when the considered matrices have either a band or a sparsity structure with $\mathcal{O}(1)$ nonzero elements per row, their spectral norm and their $l^\infty$ induced matrix norm are both bounded from above by an absolute constant independent of the matrix size times the maximal modulus of the nonzero entries.

In the right hand side, $b_u(u)$ is the discretization of the nonlinear convective terms with a classical explicit TVD Runge–Kutta method and Rusanov fluxes, as in [36]. Boundary conditions for a prescribed velocity profile at the inlet are inserted in the system in place of the first rows of $N$, $G$ and $b(u)$; we impose an outlet pressure by prescribing the stress modifying the last rows of the same blocks.

The time step $\Delta \mathrm{t}$ is restricted by a CFL-type restriction for DG schemes depending only on the fluid velocity. In the following analysis, we thus assume that $\frac{\Delta \mathrm{t}}{\Delta \mathrm{x}} = c = \mathcal{O}(1)$.

## 3 Preliminaries

Here we first formalize the definition of block Toeplitz and circulant sequences associated to a matrix-valued Lebesgue integrable function (see Sect. 3.1). Moreover, in Sect. 3.2 we introduce a class of matrix sequences containing block Toeplitz sequences known as the block Generalized Locally Toeplitz (GLT) class [6,18,19]. Some new spectral tools on certain non-square symbol sequences are introduced in Sect. 3.3 and will be used to derive the spectral properties of $\mathcal{A}$ in (7) as well as of its blocks and its Schur complement.

### 3.1 Block Toeplitz and circulant matrices

Let us denote by $L^1([-\pi,\pi],s)$ the space of $s\times s$ matrix-valued functions $f:[-\pi,\pi]\to\mathbb{C}^{s\times s}$, $f=[f_{ij}]_{i,j=1}^{s}$ with $f_{ij}\in L^1([-\pi,\pi])$, $i,j=1,\ldots,s$. In Definition 1 we introduce the notion of Toeplitz and circulant matrix sequences generated by $f$.

**Definition 1** Let $f\in L^1([-\pi,\pi],s)$ and let $t_j$ be its Fourier coefficients

$$t_j:=\frac{1}{2\pi}\int_{-\pi}^{\pi}f(\theta)\mathrm{e}^{-\mathbf{i}j\theta}\,\mathrm{d}\theta\in\mathbb{C}^{s\times s},$$

where the integrals are computed component-wise. Then, the $n$-th $s\times s$-block Toeplitz matrix associated with $f$ is the matrix of order $\widehat{n}=s\cdot n$ given by

$$T_n(f)=\left[t_{i-k}\right]_{i,k=1}^{n}.$$

Similarly, the $n$-th $s\times s$-block circulant matrix associated with $f$ is the following $\widehat{n}\times\widehat{n}$ matrix

$$C_n(f)=\left[t_{(i-k)\bmod n}\right]_{i,k=1}^{n}.$$

The sets $\{T_n(f)\}_n$ and $\{C_n(f)\}_n$ are called the *families of $s\times s$-block Toeplitz and circulant matrices generated by $f$*, respectively. The function $f$ is referred to as the *generating function* either of $\{T_n(f)\}_n$ or $\{C_n(f)\}_n$.

It is useful for our later studies to extend the definition of block-Toeplitz sequence also to the case where the symbol is a rectangular matrix-valued function.

**Definition 2** Let $f:[-\pi,\pi]\to\mathbb{C}^{s\times q}$, with $s\neq q$, and such that $f_{ij}\in L^1([-\pi,\pi])$ for $i=1,\ldots,s$ and $j=1,\ldots,q$. Then, given $n\in\mathbb{N}$, we denote by $T_n(f)$ the $s\cdot n\times q\cdot n$ matrix whose entries are $T_n(f)=[t_{i-k}]_{i,k=1}^{n}$, with $t_j\in\mathbb{C}^{s\times q}$ the Fourier coefficients of $f$.

The generating function $f$ provides a description of the spectrum of $T_n(f)$, for $n$ large enough in the sense of the following definition.

**Definition 3** Let $f : [a,b] \to \mathbb{C}^{s\times s}$ be a measurable matrix-valued function with eigenvalues $\lambda_i(f)$ and singular values $\sigma_i(f)$, $i = 1,\dots,s$. Assume that $\{A_n\}_n$ is a sequence of matrices such that $\dim(A_n) = d_n \to \infty$, as $n \to \infty$ and with eigenvalues $\lambda_j(A_n)$ and singular values $\sigma_j(A_n)$, $j = 1,\dots,d_n$.

- We say that $\{A_n\}_n$ is *distributed as $f$ over $[a,b]$ in the sense of the eigenvalues,* and we write $\{A_n\}_n \sim_\lambda (f,[a,b])$, if

$$\lim_{n\to\infty} \frac{1}{d_n}\sum_{j=1}^{d_n} F(\lambda_j(A_n)) = \frac{1}{b-a}\int_a^b \frac{\sum_{i=1}^{s} F(\lambda_i(f(t)))}{s}\mathrm{d}t, \tag{8}$$

  for every continuous function $F$ with compact support. In this case, we say that $f$ is the *spectral symbol* of $\{A_n\}_n$.
- We say that $\{A_n\}_n$ is *distributed as $f$ over $[a,b]$ in the sense of the singular values,* and we write $\{A_n\}_n \sim_\sigma (f,[a,b])$, if

$$\lim_{n\to\infty} \frac{1}{d_n}\sum_{j=1}^{d_n} F(\sigma_j(A_n)) = \frac{1}{b-a}\int_a^b \frac{\sum_{i=1}^{s} F(\sigma_i(f(t)))}{s}\mathrm{d}t, \tag{9}$$

  for every continuous function $F$ with compact support.

Throughout the paper, when the domain can be easily inferred from the context, we replace the notation $\{A_n\}_n \sim_{\lambda,\sigma} (f,[a,b])$ with $\{A_n\}_n \sim_{\lambda,\sigma} f$.

***Remark 4*** If $f$ is smooth enough, an informal interpretation of the limit relation (8) (resp. (9)) is that when $n$ is sufficiently large, then $d_n/s$ eigenvalues (resp. singular values) of $A_n$ can be approximated by a sampling of $\lambda_1(f)$ (resp. $\sigma_1(f)$) on a uniform equispaced grid of the domain $[a,b]$, and so on until the last $d_n/s$ eigenvalues (resp. singular values), which can be approximated by an equispaced sampling of $\lambda_s(f)$ (resp. $\sigma_s(f)$) in the domain.

For Toeplitz matrix sequences, the following theorem due to Tilli holds, which generalizes previous researches along the last 100 years by Szegő, Widom, Avram, Parter, Tyrtyshnikov, Zamarashkin (see [6,8,18,40] and references therein).

**Theorem 5** (see [38]) *Let $f \in L^1([-\pi,\pi],s)$, then $\{T_n(f)\}_n \sim_\sigma (f,[-\pi,\pi])$. If $f$ is a Hermitian matrix-valued function, then $\{T_n(f)\}_n \sim_\lambda (f,[-\pi,\pi])$.*

Since rectangular matrices always admit a singular value decomposition, Eq. (9) can also be extended to rectangular matrix sequences. Throughout we denote by $A_{m_1,m_2,s,q} \in \mathbb{C}^{s\cdot m_1\times q\cdot m_2}$ the rectangular matrix that has $m_1$ blocks of $s$ rows and $m_2$ blocks of $q$ columns. As a special case, with $[T_n(f)]_{m_1,m_2,s,q}$, $m_1,m_2 \le n$ we denote the 'leading principal' submatrix of $T_n(f)$ of size $s\cdot m_1 \times q\cdot m_2$. Moreover, if $f \in \mathbb{C}^{s\times q}$ then we omit the subscripts $s,q$ since they are implicitly clear from the size of the symbol.

**Definition 6** Given a measurable function $f : [a, b] \to \mathbb{C}^{s\times q}$, with $s \neq q$ and a matrix sequence $\{A_{m_1,m_2,s,q}\}_n$, with $A_n \in \mathbb{C}^{s\cdot m_1\times q\cdot m_2}$, $m_1 \sim m_2$, $m_1, m_2 \to \infty$ as $n \to \infty$ then we say that $\{A_{m_1,m_2,s,q}\}_n \sim_\sigma (f, [a, b])$ iff

$$\lim_{n\to\infty} \frac{1}{s\cdot m_1 \wedge q\cdot m_2} \sum_{j=1}^{s\cdot m_1\wedge q\cdot m_2} F(\sigma_j(A_{m_1,m_2,s,q})) = \frac{1}{b-a}\int_a^b \frac{\sum_{i=1}^{s\wedge q} F(\sigma_i(f(t)))}{s\wedge q}\mathrm{d}t,$$

with $x \wedge y := \min\{x, y\}$, for every continuous function $F$ with compact support.

***Remark 7*** Based on Definition 6 the first part of Theorem 5 extends also to rectangular block Toeplitz matrices in the sense of Definition 2 (see [38]) as well as to sequences whose $n$-th matrix is $A_{m_1,m_2,s,q} = [T_n(f)]_{m_1,m_2}$, $f \in \mathbb{C}^{s\times q}$, with $m_1, m_2 \leq n$, $m_1 \sim m_2$, $m_1, m_2 \to \infty$ as $n \to \infty$.

The following theorem is a useful tool for computing the spectral distribution of a sequence of Hermitian matrices. For the related proof, see [27, Theorem 4.3]. Here, the conjugate transpose of the matrix $X$ is denoted by $X^*$.

**Theorem 8** (see [27, Theorem 4.3]) *Let $\{A_n\}_n$ be a sequence of matrices, with $A_n$ Hermitian of size $d_n$, and let $\{P_n\}_n$ be a sequence such that $P_n \in \mathbb{C}^{d_n\times\delta_n}$, $P_n^* P_n = I_{\delta_n}$, $\delta_n \leq d_n$ and $\delta_n/d_n \to 1$ as $n \to \infty$. Then $\{A_n\}_n \sim_\lambda f$ if and only if $\{P_n^* A_n P_n\}_n \sim_\lambda f$.*

The following result allows us to determine the spectral distribution of a Hermitian matrix sequence plus a correction, not necessarily of Hermitian nature.

**Theorem 9** (see [7, Theorem 1]) *Let $\{X_n\}_n$ and $\{Y_n\}_n$ be two matrix sequences, with $X_n, Y_n \in \mathbb{C}^{d_n\times d_n}$, and assume that*

(a) *$X_n$ is Hermitian for all $n$ and $\{X_n\}_n \sim_\lambda f$;*
(b) *$\|Y_n\|_F = o(\sqrt{d_n})$ as $n \to \infty$, with $\|\cdot\|_F$ the Frobenius norm.*

*Then, $\{X_n + Y_n\}_n \sim_\lambda f$.*

For a given matrix $X \in \mathbb{C}^{m\times m}$, let us denote by $\|X\|_1$ the trace norm defined by $\|X\|_1 := \sum_{j=1}^m \sigma_j(X)$, where $\sigma_j(X)$ are the $m$ singular values of $X$.

**Corollary 10** (see [7, Corollary 2]) *Let $\{X_n\}_n$ and $\{Y_n\}_n$ be two matrix sequences, with $X_n, Y_n \in \mathbb{C}^{d_n\times d_n}$, and assume that (a) in Theorem 9 is satisfied. Moreover, assume that any of the following two conditions is met:*

- *$\|Y_n\|_1 = o(\sqrt{d_n})$;*
- *$\|Y_n\| = o(1)$, with $\|\cdot\|$ being the spectral norm.*

*Then, $\{X_n + Y_n\}_n \sim_\lambda f$.*

We end this subsection by reporting the key features of the block circulant matrices, also in connection with the generating function. We refer to [21], but the result is quite classical and can be found in many other references.

**Theorem 11** (see [21] and references therein) *Let $f \in L^1([-\pi,\pi], s)$ be a matrix-valued function with $s \geq 1$ and let $\{t_j\}_{j\in\mathbb{Z}}$, $t_j \in \mathbb{C}^{s\times s}$ be its Fourier coefficients. Then, the following (block-Schur) decomposition of $C_n(f)$ holds:*

$$C_n(f) = (F_n \otimes I_s) D_n(f) (F_n \otimes I_s)^*, \tag{10}$$

*where*

$$D_n(f) = \mathrm{diag}_{0\leq r\leq n-1} \left(S_n(f)\left(\theta_r\right)\right), \quad \theta_r = \frac{2\pi r}{n}, \quad F_n = \frac{1}{\sqrt{n}} \left(\mathrm{e}^{-\mathbf{i}j\theta_r}\right)_{j,r=0}^{n-1} \tag{11}$$

*with $S_n(f)(\cdot)$ the $n$-th Fourier sum of $f$ given by*

$$S_n(f)(\theta) = \sum_{j=0}^{n-1} t_j \mathrm{e}^{\mathbf{i}j\theta}. \tag{12}$$

*Moreover, the eigenvalues of $C_n(f)$ are given by the evaluations of $\lambda_t(S_n(f)(\theta))$, $t = 1,\ldots,s$, if $s \geq 2$ or of $S_n(f)(\theta)$ if $s = 1$ at the grid points $\theta_r$.*

***Remark 12*** If $f$ is a trigonometric polynomial of fixed degree (with respect to $n$), then it is worth noticing that $S_n(f)(\cdot) = f(\cdot)$ for $n$ large enough: more precisely, $n$ should be larger than the double of the degree. Therefore, in such a setting, the eigenvalues of $C_n(f)$ are either the evaluations of $f$ at the grid points if $s = 1$ or the evaluations of $\lambda_t(f(\cdot))$, $t = 1,\ldots,s$, at the very same grid points.

We recall that every matrix/vector operation with circulant matrices has cost $O(\widehat{n}\log\widehat{n})$ with moderate multiplicative constants: in particular, this is true for the matrix-vector product, for the solution of a linear system, for the computation of the blocks $S_n(f)(\theta_r)$ and consequently of the eigenvalues (see e.g. [41]).

### 3.2 Block generalized locally Toeplitz matrix sequences

In the sequel, we introduce the block GLT class, a $*$-algebra of matrix sequences containing block Toeplitz matrix sequences. The formal definition of block GLT matrix sequences is rather technical and can be found in the scalar unilevel, scalar multilevel, block unilevel, block multilevel in the following books and revue papers [5,6,18, 20], respectively. The construction is involved and needs a whole coherent set of definitions and mathematical objects. However, in the writing of the books and the reviews, the authors realized that the mathematical construction is equivalent to a set of operative axioms that can be used conveniently, in practice, for deciding if a given matrix sequence is of GLT type and for computing the related symbol. Therefore, we just give and briefly report and discuss four of these axioms of the block GLT class, which are sufficient for studying the spectral features of $\mathcal{A}$ as well as of its blocks and its Schur complement. The current formulation is taken from [6].

Throughout, we use the following notation

$$\{A_n\}_n \sim_{\mathrm{GLT}} \kappa(x,\theta), \qquad \kappa : [0,1]\times[-\pi,\pi] \to \mathbb{C}^{s\times s},$$

to say that the sequence $\{A_n\}_n$ is a $s \times s$-block GLT sequence with GLT symbol $\kappa(x, \theta)$.

Here we list four main features of block GLT sequences.

**GLT1** Let $\{A_n\}_n \sim_{\text{GLT}} \kappa$ with $\kappa : G \to \mathbb{C}^{s\times s}$, $G = [0, 1] \times [-\pi, \pi]$, then $\{A_n\}_n \sim_\sigma (\kappa, G)$. If the matrices $A_n$ are Hermitian, then it also holds that $\{A_n\}_n \sim_\lambda (\kappa, G)$.

**GLT2** The set of block GLT sequences forms a $*$-algebra, i.e., it is closed under linear combinations, products, conjugation, but also inversion when the symbol is invertible a.e. In formulae, let $\{A_n\}_n \sim_{\text{GLT}} \kappa_1$ and $\{B_n\}_n \sim_{\text{GLT}} \kappa_2$, then

- $\{\alpha A_n + \beta B_n\}_n \sim_{\text{GLT}} \alpha\kappa_1 + \beta\kappa_2, \quad \alpha, \beta \in \mathbb{C}$;
- $\{A_n B_n\}_n \sim_{\text{GLT}} \kappa_1\kappa_2$;
- $\{A_n^*\}_n \sim_{\text{GLT}} \kappa_1^*$;
- $\{A_n^{-1}\}_n \sim_{\text{GLT}} \kappa_1^{-1}$ provided that $\kappa_1$ is invertible a.e.

**GLT 3** Any sequence of block Toeplitz matrices $\{T_n(f)\}_n$ generated by a function $f \in L^1([-\pi, \pi], s)$ is a $s \times s$-block GLT sequence with symbol $\kappa(x, \theta) = f(\theta)$.

**GLT4** Let $\{A_n\}_n \sim_\sigma 0$. We say that $\{A_n\}_n$ is a *zero-distributed matrix sequence*. Note that for any $s > 1$ $\{A_n\}_n \sim_\sigma O_s$, with $O_s$ the $s \times s$ null matrix, is equivalent to $\{A_n\}_n \sim_\sigma 0$. Every zero-distributed matrix sequence is a block GLT sequence with symbol $O_s$ and viceversa, i.e., $\{A_n\}_n \sim_\sigma 0 \iff \{A_n\}_n \sim_{\text{GLT}} O_s$.

According to Definition 3, in the presence of a zero-distributed sequence the singular values of the $n$-th matrix (weakly) cluster around 0. This is formalized in the following result [18].

**Proposition 13** *Let $\{A_n\}_n$ be a matrix sequence with $A_n$ of size $d_n$ with $d_n \to \infty$, as $n \to \infty$. Then $\{A_n\}_n \sim_\sigma 0$ if and only if there exist two matrix sequences $\{R_n\}_n$ and $\{E_n\}_n$ such that $A_n = R_n + E_n$, and*

$$\lim_{n\to\infty} \frac{rank(R_n)}{d_n} = 0, \qquad \lim_{n\to\infty} \|E_n\| = 0.$$

*The matrix $R_n$ is called rank-correction and the matrix $E_n$ is called norm-correction.*

Regarding the low rank-correction versus relatively small norm-correction splitting, it should be noted that it represents an important theoretical tool for the analysis of spectral and singular-value distributions, as emphasized by Eugene Tyrtyshnikov in a very successful and seminal paper [39]. However, its use started for different reasons in the analysis of efficient preconditioners, especially for structured matrices of Toeplitz type, and the main name in this respect is that of Raymond Chan (see [9] and references therein). Subsequently, these tools have evolved into the more sophisticate notion of approximating class of sequences (see [18] and references therein), thanks to Tilli and to the third author.

### 3.3 Some new spectral tools

In this subsection we introduce some new spectral tools that will be used in Sect. 4. Specifically, we propose an alternative to the cumbersome embedding in larger square matrices followed in [12] for spectrally studying Schur complement matrix sequences composed of rectangular matrices.

The following theorem concerns the spectral behavior of matrix sequences whose $n$-th matrix is a product of a square block Toeplitz matrix by a rectangular one.

**Theorem 14** *Let $f : [-\pi, \pi] \to \mathbb{C}^{s\times s}$ and let $g : [-\pi, \pi] \to \mathbb{C}^{s\times q}$, $h : [-\pi, \pi] \to \mathbb{C}^{q\times s}$ with $q < s$. Then*

$$\{T_n(f)T_n(g)\}_n \sim_\sigma (f \cdot g, [-\pi, \pi]), \tag{13}$$

*and*

$$\{T_n(h)T_n(f)\}_n \sim_\sigma (h \cdot f, [-\pi, \pi]). \tag{14}$$

***Proof*** We only prove relation (13), since the same argument easily brings to (14) as well. Let us define $g_{\text{ex}} : [-\pi, \pi] \to \mathbb{C}^{s\times s}$ obtained completing $g$ with $s - q$ null columns. By **GLT3** and **GLT2** we know that

$$\{T_n(f)T_n(g_{\text{ex}})\}_n \sim_\sigma (f \cdot g_{\text{ex}}, [-\pi, \pi]). \tag{15}$$

Let us now explicitly write (15) according to Definition 3

$$\lim_{n\to\infty} \frac{1}{sn} \sum_{j=1}^{sn} F(\sigma_j(T_n(f)T_n(g_{\text{ex}}))) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{\sum_{i=1}^{s} F(\sigma_i(f(t)g_{\text{ex}}(t)))}{s} \mathrm{d}t.$$

The left-hand side of the previous equation can be rewritten as follows

$$\begin{aligned}
&\lim_{n\to\infty} \frac{1}{sn} \sum_{j=1}^{sn} F(\sigma_j(T_n(f)T_n(g_{\text{ex}}))) \\
&= \lim_{n\to\infty} \frac{1}{sn} \left[ \sum_{j=1}^{qn} F(\sigma_j(T_n(f)T_n(g_{\text{ex}}))) + \sum_{qn+1}^{sn} F(0) \right] \\
&= \lim_{n\to\infty} \frac{1}{sn} \sum_{j=1}^{qn} F(\sigma_j(T_n(f)T_n(g))) + \frac{(s-q)}{s} F(0),
\end{aligned}$$

while manipulating the right-hand side we obtain

$$\begin{aligned}
&\frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{\sum_{i=1}^{s}F(\sigma_i(f(t)g_{\mathrm{ex}}(t)))}{s}\mathrm{d}t\\
&=\frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{\sum_{i=1}^{q}F(\sigma_i(f(t)g_{\mathrm{ex}}(t)))+\sum_{i=q+1}^{s}F(0)}{s}\mathrm{d}t\\
&=\frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{\sum_{i=1}^{q}F(\sigma_i(f(t)g(t)))+(s-q)F(0)}{s}\mathrm{d}t\\
&=\frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{\sum_{i=1}^{q}F(\sigma_i(f(t)g(t)))}{s}\mathrm{d}t+\frac{(s-q)}{s}F(0).
\end{aligned}$$

Therefore we arrive at

$$\lim_{n\to\infty}\frac{1}{sn}\sum_{j=1}^{qn}F(\sigma_j(T_n(f)T_n(g)))=\frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{\sum_{i=1}^{q}F(\sigma_i(f(t)g(t)))}{s}\mathrm{d}t.$$

which proves (13), once multiplied by $\frac{s}{q}$.

□

***Remark 15*** Theorem 14 can easily be extended to the case where also $T_n(f)$ is a properly sized rectangular block Toeplitz matrix. In particular, when $f\cdot g$ (or $h\cdot f$) results in a Hermitian square matrix-valued function then the distribution also holds in the sense of the eigenvalues.

Along the same lines of the previous theorem the following result holds. We notice that Theorem 14 and Theorem 16 are special cases of a more general theory which connects GLT sequences having symbols with different matrix sizes: the considered general study is contained in the work [4].

**Theorem 16** *Let $g:[-\pi,\pi]\to\mathbb{C}^{s\times s}$ be Hermitian positive definite almost everywhere and let $f:[-\pi,\pi]\to\mathbb{C}^{q\times s}$ with $q<s$. Then*

$$\{T_n(f)T_n^{-1}(g)T_n(f^*)\}_n\sim_\sigma(f\cdot g^{-1}\cdot f^*,[-\pi,\pi]),$$

*and*

$$\{T_n(f)T_n^{-1}(g)T_n(f^*)\}_n\sim_\lambda(f\cdot g^{-1}\cdot f^*,[-\pi,\pi]).$$

The following result will be used in combination with Theorem 8 to obtain the spectral symbol of the whole coefficient matrix sequence appearing in (7). The idea of computing the symbol by similarity via a permutation transform to a Toeplitz is not new and in fact it can be found already in [12,21], in different and even more general contexts.

**Theorem 17** *Let*

$$A_n=\begin{bmatrix}T_n(\mathfrak{f}_{11}) & T_n(\mathfrak{f}_{12})\\ T_n(\mathfrak{f}_{21}) & T_n(\mathfrak{f}_{22})\end{bmatrix}$$

*with* $\mathfrak{f}_{11} : [-\pi, \pi] \to \mathbb{C}^{k\times k}$, $\mathfrak{f}_{12} : [-\pi, \pi] \to \mathbb{C}^{k\times q}$, $\mathfrak{f}_{21} : [-\pi, \pi] \to \mathbb{C}^{q\times k}$, $\mathfrak{f}_{22} : [-\pi, \pi] \to \mathbb{C}^{q\times q}$, $k, q \in \mathbb{N}$. *Then there exists a permutation matrix* $\Pi$ *such that* $A_n = \Pi T_n(\mathfrak{f})\Pi^T$ *with*

$$\mathfrak{f} = \begin{bmatrix} \mathfrak{f}_{11} & \mathfrak{f}_{12} \\ \mathfrak{f}_{21} & \mathfrak{f}_{22} \end{bmatrix}.$$

*Hence* $A_n$ *and* $T_n(\mathfrak{f})$ *share the same eigenvalues and the same singular values and consequently* $\{A_n\}_n$ *and* $\{T_n(\mathfrak{f})\}_n$ *enjoy the same distribution features.*

***Proof*** Let $I_{kn+qn}$ be the identity matrix of size $kn+qn$ and let us define the following sets of indexes $H = \{1, \ldots, kn+qn\}$ and $J = \{k+1, \ldots, k+q, 2k+q+1, \ldots, 2k+2q, 3k+2q+1, \ldots, 3k+3q, \ldots, nk+(n-1)q+1, \ldots, nk+nq\}$. Let $\Pi$ be the $(kn+qn)\times(kn+qn)$-matrix whose first $kn$ rows are defined as the rows of $I_{kn+qn}$ that correspond to the indexes in $H\setminus J$ and the remaining as the rows of $I_{kn+qn}$ that correspond to the indexes in $J$. The thesis easily follows observing that $\Pi$ is the permutation matrix that relates $A_n$ and $T_n(\mathfrak{f})$.

Thus $A_n$ and $T_n(\mathfrak{f})$ are similar because $\Pi^T$ is the inverse of $\Pi$ and as consequence both matrices $A_n$ and $T_n(\mathfrak{f})$ share the same eigenvalues. Furthermore both $\Pi$ and $\Pi^T$ are unitary and consequently by the singular value decomposition the two matrices $A_n$ and $T_n(\mathfrak{f})$ share the same singular values. Finally it is transparent that one of the matrix sequences (between $\{A_n\}_n$ and $\{T_n(\mathfrak{f})\}_n$) has a distribution if and only the other has the very same distribution. □

## 4 Spectral analysis

This section concerns the spectral study of the matrix $\mathcal{A}$ in (7) together with its blocks and Schur complement. In the following, we consider the case of $d(x) = d$ (constant width); we choose at first the smallest nontrivial case which is $n_x = 1$ and $n_y = 3$ ($n_u = (n_x+1)(n_y-1) = 4$ and $n_p = (n_x+1) = 2$) and then comment on the general case.

### 4.1 Spectral study of the blocks of $\mathcal{A}$

We start by spectrally analyzing the four blocks that compose the matrix $\mathcal{A}$.

*Laplacian and mass operator.* The (1, 1) block $N$ of $\mathcal{A}$ in (7) is a sum of two terms: the Laplacian matrix $L$ and the mass matrix $M$ that are respectively obtained by testing the PDE term $\nabla\cdot(\mu\,\nabla u)$ and the term $\partial_t u$ with the basis functions for velocity.

The matrix $L$ is organized in blocks of rows each of size $n_u = 4$ which corresponds to the number of test functions per cell (associated with the blue degrees of freedom in Fig. 2); in each row there are at most twelve nonzeros elements (associated with all the degrees of freedom in Fig. 2). Using SIP in (3) and excluding the boundary conditions, we can write

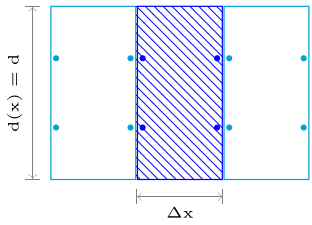



**Fig. 2** Illustration of the stencil that refers to the mass and Laplacian matrix

$$L_{n+1} = \frac{27}{70} d\mu c U_{n+1}$$

with

$$U_{n+1} = \text{tridiag} \left[\begin{array}{cccc|cccc|cccc} -\frac{1}{2} & \frac{1}{16} & 0 & 0 & 1 & -\frac{1}{8} & 0 & 0 & -\frac{1}{2} & \frac{1}{16} & 0 & 0 \\ \frac{1}{16} & -\frac{1}{2} & 0 & 0 & -\frac{1}{8} & 1 & 0 & 0 & \frac{1}{16} & -\frac{1}{2} & 0 & 0 \\ 0 & 0 & -\frac{1}{2} & \frac{1}{16} & 0 & 0 & 1 & -\frac{1}{8} & 0 & 0 & -\frac{1}{2} & \frac{1}{16} \\ 0 & 0 & \frac{1}{16} & -\frac{1}{2} & 0 & 0 & -\frac{1}{8} & 1 & 0 & 0 & \frac{1}{16} & -\frac{1}{2} \end{array}\right] + \mathcal{O}(\Delta \mathrm{x}^2),$$

where $\mu$ is the viscosity, $c = \frac{\Delta \mathrm{t}}{\Delta \mathrm{x}}$, and $n+1$ is the number of velocity cells. In fluid dynamics, it is natural to choose a timestep proportional to the grid size (and inversely proportional to the fluid velocity), and thus we assume that $c = \mathcal{O}(1)$.

It is then clear that $L_{n+1}$ is a $4 \times 4$-block Toeplitz matrix of size $\widehat{n} = 4 \cdot (n+1)$. As a consequence, we can obtain insights on its spectrum studying the symbol associated to $\{L_{n+1}\}_n$. With this aim, let us define

$$X = \begin{bmatrix} \frac{1}{2} & -\frac{1}{16} \\ -\frac{1}{16} & \frac{1}{2} \end{bmatrix},$$

and $l_1, l_0, l_{-1}$ as follows

$$l_1 = \left[\begin{array}{c|c} -X & 0 \\ \hline 0 & -X \end{array}\right], \qquad l_0 = \left[\begin{array}{c|c} 2X & 0 \\ \hline 0 & 2X \end{array}\right], \qquad l_{-1} = \left[\begin{array}{c|c} -X & 0 \\ \hline 0 & -X \end{array}\right].$$

Since we are assuming that $c = \mathcal{O}(1)$ the symbol associated to $\{L_{n+1}\}_n$ is the function $\mathscr{L} : [-\pi, \pi] \to \mathbb{C}^{4\times 4}$ defined as

$$\mathscr{L}(\theta) = \frac{27}{70} d\mu c (l_0 + l_1 e^{\mathbf{i}\theta} + l_{-1} e^{-\mathbf{i}\theta}) = \frac{27}{70} d\mu c \begin{bmatrix} (2 - 2\cos\theta) & 0 \\ 0 & (2 - 2\cos\theta) \end{bmatrix} \otimes X.$$

Recalling Theorem 5 and **GLT3**, we conclude that

$$\{L_{n+1}\}_n \sim_{\text{GLT},\sigma,\lambda} (\mathscr{L}, [-\pi, \pi]). \tag{16}$$

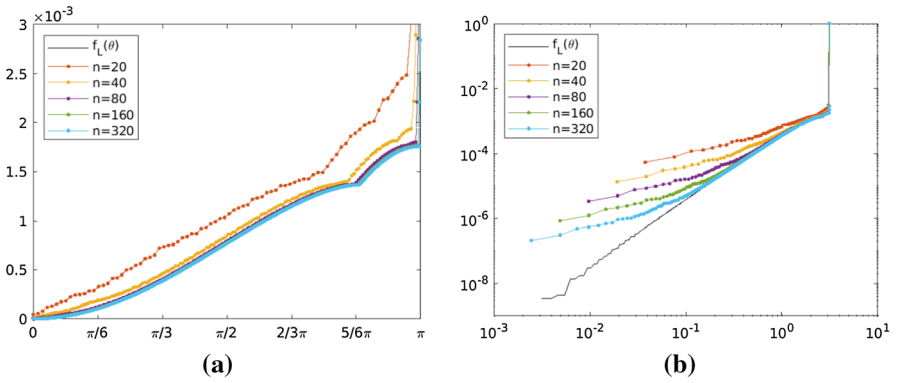



**Fig. 3** **a** The spectrum of $L_{n+1}$ with different number of cells versus sampling of the eigenvalue functions of the symbol $\mathscr{L}(\theta)$; **b** is the same picture, but in bilogarithmic scale

***Remark 18*** We have assumed that $L_{n+1}$ does not contain the boundary conditions, but if we let them come into play, then the spectral distribution would remain unchanged. Indeed, the matrix that corresponds to the Laplacian operator can be expressed as the sum $L_{n+1} + R_{n+1}$ with $R_{n+1}$ a rank-correction. Since the boundary conditions imply a correction in a constant number of entries and since the absolute values of such corrections are uniformly bounded with respect to the matrix size, it easily follows that $\|R_{n+1}\| = \mathcal{O}(1)$ and hence Theorem 9 can be applied.

It is easy to compute the four eigenvalue functions of $\mathscr{L}(\theta)$, which are $\frac{27}{70}d\mu c2(1-\cos\theta)\left(\frac{1}{2} \pm \frac{1}{16}\right)$, each with multiplicity 2. Note that all eigenvalue functions vanish at $\theta = 0$ with a zero of second order. Recalling Remark 4, we expect that a sampling of the eigenvalues of $\mathscr{L}(\theta)$ provides an approximation of the spectrum of the discretized Laplacian operator. This is confirmed in Fig. 3, where we compare the Laplacian matrix, including the boundary conditions, with an equispaced sampling of the eigenvalue functions of $\mathscr{L}(\theta)$ in $[-\pi, \pi]$.

The mass matrix $M_{n+1}$ is block diagonal and has the form

$$M_{n+1} = \frac{9}{70} d\Delta \mathrm{x} \rho \,\mathrm{diag} \begin{bmatrix} 1 & -\frac{1}{8} & \frac{1}{2} & -\frac{1}{16} \\ -\frac{1}{8} & 1 & -\frac{1}{16} & \frac{1}{2} \\ \frac{1}{2} & -\frac{1}{16} & 1 & -\frac{1}{8} \\ -\frac{1}{16} & \frac{1}{2} & -\frac{1}{8} & 1 \end{bmatrix}.$$

As for $L_{n+1}$, also $M_{n+1}$ is a $4 \times 4$-block Toeplitz of size $\widehat{n} = 4 \cdot (n+1)$. In order to study its symbol we look at the scaled matrix sequence $\{\frac{1}{\Delta x} M_{n+1}\}_n$. The reason for such scaling is that the symbol is defined for sequences of Toeplitz matrices whose elements do not vary with their size. The symbol of the scaled mass-matrix sequence $\{\frac{1}{\Delta x} M_{n+1}\}_n$ can be written as

$$\mathscr{M}(\theta) = \frac{9}{70} d\rho \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} \otimes X$$

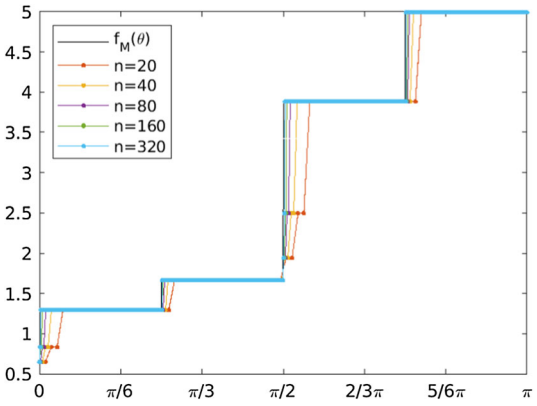



**Fig. 4** The eigenvalues of $\frac{1}{\Delta x} M_{n+1}$ matrix with different number of cells versus sampling of the eigenvalue functions of $\mathscr{M}(\theta)$

with $X$ as in (16) and again by Theorem 5 and **GLT3** we have

$$\left\{ \frac{1}{\Delta x} M_{n+1} \right\}_n \sim_{\text{GLT},\sigma,\lambda} (\mathscr{M}, [-\pi, \pi]). \tag{17}$$

Therefore, its eigenvalues are $\frac{9}{70} d\rho \, (2 \pm 1) \left( \frac{1}{2} \pm \frac{1}{16} \right)$.

In Fig. 4 we compare an equispaced sampling of the eigenvalues of $\mathscr{M}(\theta)$ with the spectrum of the mass matrix sequences and we see that the matching is getting better and better as the number of cells increases.

Since the (1, 1) block of $\mathcal{A}$ is given by the sum of $L_{n+1}$ and $M_{n+1}$, we are interested in the symbol of $\{N_{n+1} = L_{n+1} + M_{n+1}\}_n$. Let us first note that because of the presence of $\Delta$x in its definition, $M_{n+1}$ is a norm-correction of $L_{n+1}$ and that $N_{n+1}$ is real symmetric when boundary conditions are excluded. Then, by using Proposition 13, Eq. (16), and **GLT1-4** we have that

$$\{N_{n+1}\}_n \sim_{\text{GLT},\sigma,\lambda} (\mathscr{L}, [-\pi, \pi]). \tag{18}$$

Figure 5 checks numerically relation (18) by comparing the eigenvalues of $N_{n+1}$ modified by the boundary conditions (see Remark 18) with an equispaced sampling of the eigenvalue functions of $\mathscr{L}(\theta)$.

*Gradient operator.* The (1, 2) block $G$ of $\mathcal{A}$ in (7) is organized in blocks of rows, each of size $n_u = 4$ (blue degrees of freedom in Fig. 6); in each row there are $2n_p = 4$ nonzero elements (red degrees of freedom in Fig. 6), half of which are associated with the pressure cell intersecting the velocity cell in its left (respectively right) half.

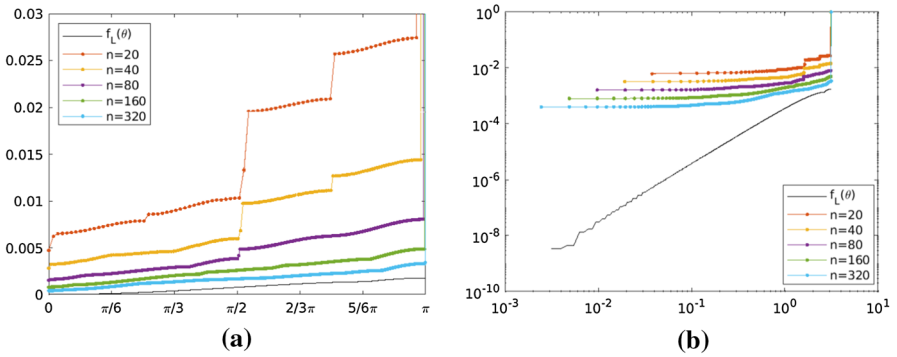



**Fig. 5** **a** The spectrum of $(M_{n+1} + L_{n+1})$ with different number of cells versus sampling of the eigenvalue functions of $\mathscr{L}(\theta)$ associated to the only matrix $L_{n+1}$; **b** is the same picture, but in bilogarithmic scale

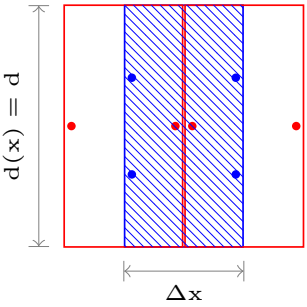


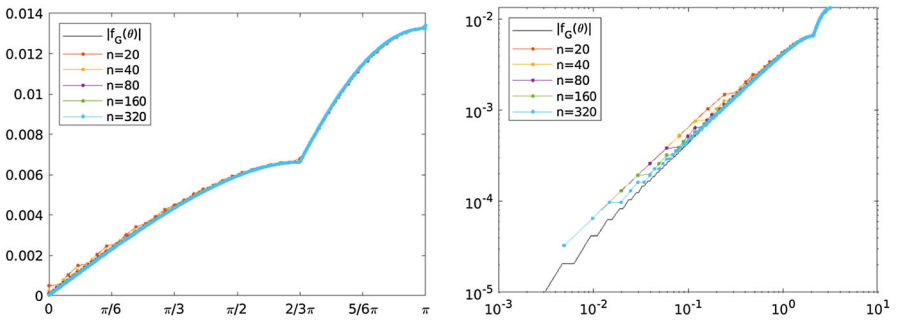

**Fig. 6** Illustration of the stencil that refers to the pressure gradient matrix $G_{n+1,n}$

Therefore the gradient matrix is a $4(n+1) \times 2n$ rectangular matrix that, excluding boundary conditions, can be written as

$$G_{n+1,n} = \frac{3}{64} d\Delta t \begin{bmatrix} g_0 & 0 & \cdots & \cdots & 0 \\ g_1 & g_0 & 0 & & \vdots \\ 0 & g_1 & g_0 & 0 & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & 0 & g_1 & g_0 \\ \vdots & & & 0 & g_1 \\ 0 & \cdots & \cdots & 0 & g_1 \end{bmatrix}$$

where $g_0 = \begin{bmatrix} 3 & 1 \\ 3 & 1 \\ 1 & 3 \\ 1 & 3 \end{bmatrix}$ and $g_1 = -g_0$.

Similarly to what has been done for the mass matrix sequence, due to the presence of $\Delta t$ in $G_{n+1,n}$, we focus on the symbol of the scaled sequence $\{\frac{1}{\Delta t} G_{n+1,n}\}_n$. Note that $\frac{1}{\Delta t} G_{n+1,n}$ is a submatrix of a $4 \times 2$-block rectangular Toeplitz, precisely $G_{n+1,n} =$




**(a)** **(b)**

**Fig. 7** **a** The singular values of $\frac{1}{\Delta t} G_{n+1,n}$ matrix with different number of cells versus sampling of the singular value functions of $\mathscr{G}(\theta)$; **b** is the same picture, but in bilogarithmic scale

$[T_n(\mathscr{G})]_{n+1,n}$ with $\mathscr{G} : [-\pi, \pi] \to \mathbb{C}^{4\times 2}$ defined by

$$\mathscr{G}(\theta) = \frac{3}{64} d \left( g_0 + g_1 e^{\mathbf{i}\theta} \right) = \frac{3}{64} d\, g_0 (1 - e^{\mathbf{i}\theta}) = -\mathbf{i} \frac{3}{32} d\, g_0\, e^{\mathbf{i}\frac{\theta}{2}} \sin\left(\frac{\theta}{2}\right),$$

and thanks to Remark 7 we deduce

$$\left\{ \frac{1}{\Delta t} G_{n+1,n} \right\}_n \sim_\sigma (\mathscr{G}, [-\pi, \pi]). \tag{19}$$

The singular value decomposition of $g_0$ is $U \Sigma V^T$ where

$$U = \frac{1}{2} \begin{bmatrix} -1 & -1 & -1 & -1 \\ -1 & -1 & 1 & 1 \\ -1 & 1 & 1 & -1 \\ -1 & 1 & -1 & 1 \end{bmatrix} \qquad V = \frac{\sqrt{2}}{2} \begin{bmatrix} -1 & -1 \\ -1 & 1 \end{bmatrix} \qquad \Sigma = 2\sqrt{2} \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}$$

and thus the singular value functions of the symbol $\mathscr{G}(\theta)$ are $-\frac{3}{8}\sqrt{2}\,\mathbf{i} e^{\mathbf{i}\frac{\theta}{2}} \sin\left(\frac{\theta}{2}\right)$ and $-\frac{3}{16}\sqrt{2}\,\mathbf{i} e^{\mathbf{i}\frac{\theta}{2}} \sin\left(\frac{\theta}{2}\right)$. Figure 7 shows the very good agreement of the spectrum of $\frac{1}{\Delta t} G_{n+1,n}$ with the sampling of the singular value functions of $\mathscr{G}(\theta)$ for different number of cells.

*Divergence operator.* The $(2, 1)$ block $D$ of the matrix $\mathcal{A}$ is organized in blocks of rows each of size $n_p = 2$ (red degrees of freedom in Fig. 8); in each row there are $2n_u = 8$ nonzero elements (blue degrees of freedom in Fig. 8), half of which are associated with the velocity cell intersecting the pressure cell in its left (respectively right) half.

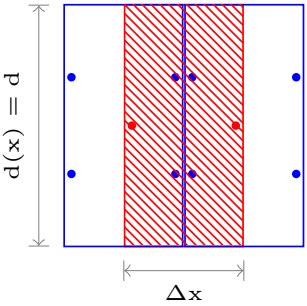



**Fig. 8** Illustration of the stencil that refers to the divergence matrix $D_{n,n+1}$

Similarly to the analysis of the gradient of the pressure, we can define $d_0 = \begin{bmatrix} 3 & 3 & 1 & 1 \\ 1 & 1 & 3 & 3 \end{bmatrix} = g_0^T$ and $d_{-1} = -d_0$, and we can write the divergence matrix as

$$D_{n,n+1} = \frac{3}{64} d \begin{bmatrix} d_0 & d_{-1} & 0 & \cdots & \cdots & \cdots & 0 \\ 0 & d_0 & d_{-1} & 0 & & & \vdots \\ \vdots & 0 & d_0 & d_{-1} & 0 & & \vdots \\ \vdots & & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & & 0 & d_0 & d_{-1} & 0 \\ 0 & \cdots & \cdots & \cdots & 0 & d_0 & d_{-1} \end{bmatrix}$$

Since the matrix $D_{n,n+1}$ is the transpose of $\frac{1}{\Delta t} G_{n+1,n}$, the generating function is

$$\mathscr{D}(\theta) = (\mathscr{G}(\theta))^* = \mathbf{i} \frac{3}{32} d \, g_0^T \, e^{-\mathbf{i}\frac{\theta}{2}} \, \sin\left(\tfrac{\theta}{2}\right)$$

which admits the same singular value functions of $\mathscr{G}(\theta)$. Therefore, by Remark 7 we find

$$\{D_{n,n+1}\}_n \sim_\sigma (\mathscr{D}, [-\pi, \pi]). \tag{20}$$

A comparison of the sampling of the singular values of $\mathscr{D}(\theta)$ with the singular values of $D_{n,n+1}$ is shown in Figs. 9 and 10.

***Remark 19*** If we analyse the product of the symbols for $D_{n,n+1}$ and $\frac{1}{\Delta t} G_{n+1,n}$, we obtain a $\mathbb{C}^{2\times 2}$-valued symbol:

$$\begin{aligned} \mathscr{D}(\theta)\mathscr{G}(\theta) = V \Sigma U^T U \Sigma V^T &= \begin{bmatrix} 5 & 3 \\ 3 & 5 \end{bmatrix} 4 \sin^2\left(\tfrac{\theta}{2}\right) \left(\frac{3}{32} d\right)^2 \\ &= \begin{bmatrix} 5 & 3 \\ 3 & 5 \end{bmatrix} 2(1 - \cos\theta) \left(\frac{3}{32} d\right)^2 \end{aligned}$$

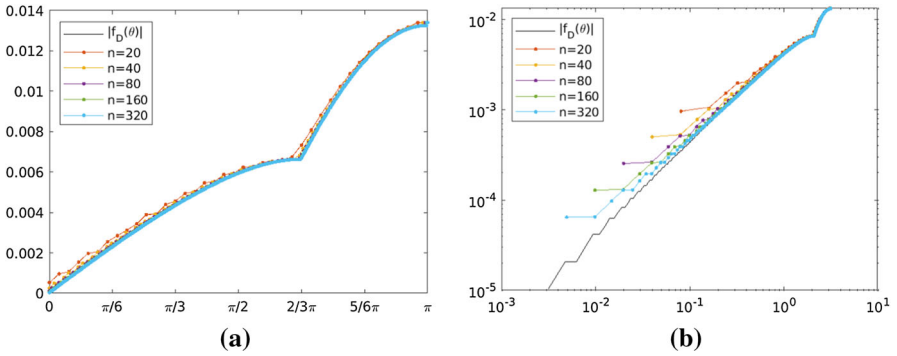



**Fig. 9** **a** The singular values of $D_{n,n+1}$ different number of cells versus sampling of the singular value functions of $\mathscr{G}(\theta)$; **b** is the same picture, but in bilogarithmic scale

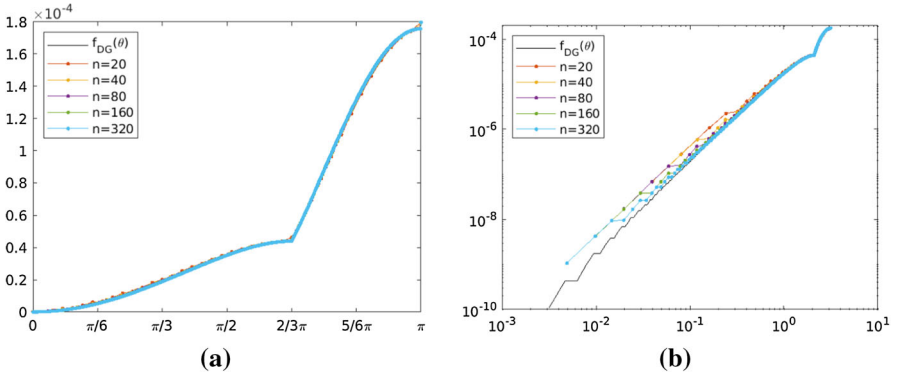



**Fig. 10** **a** The spectrum of the matrix product $\frac{1}{\Delta t} D_{n,n+1} G_{n+1,n}$ with different number of cells versus sampling of the eigenvalues of $\mathscr{D}(\theta)\mathscr{G}(\theta)$; **b** is the same picture, but in bilogarithmic scale

Its eigenvalue functions are $4(1-\cos\theta)\left(\frac{3}{64}d\right)^2$ and $16(1-\cos\theta)\left(\frac{3}{64}d\right)^2$. Notice that, since $D_{n,n+1} = [T_n(\mathscr{D})]_{n,n+1}$ and $\frac{1}{\Delta t} G_{n+1,n} = [T_n(\mathscr{G})]_{n+1,n}$, then $\frac{1}{\Delta t} D_{n,n+1} G_{n+1,n}$ is a principal submatrix of $T_n(\mathscr{D}) T_n(\mathscr{G})$. Therefore, thanks to Theorem 14 and Remark 15, $\mathscr{D}(\theta)\mathscr{G}(\theta)$ is the spectral symbol of $\{T_n(\mathscr{D}) T_n(\mathscr{G})\}_n$ and, by Theorem 8, it is also the symbol of $\{\frac{1}{\Delta t} D_{n,n+1} G_{n+1,n}\}_n$. As a consequence, we expect that a sampling of the eigenvalue functions of $\mathscr{D}(\theta)\mathscr{G}(\theta)$ provides an approximation of the spectrum of $\frac{1}{\Delta t} D_{n,n+1} G_{n+1,n}$. This is confirmed by Fig. 10.

*Penalty term for pressure.* The (2, 2) block of matrix $\mathcal{A}$ is organized in blocks of rows, each of size $n_p = 2$ and it has the following form

$$E_n = d\Delta \mathrm{x}\, \mathrm{tridiag} \left[\begin{array}{cc|cc|cc} 0 & 1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 1 & 0 \end{array}\right],$$

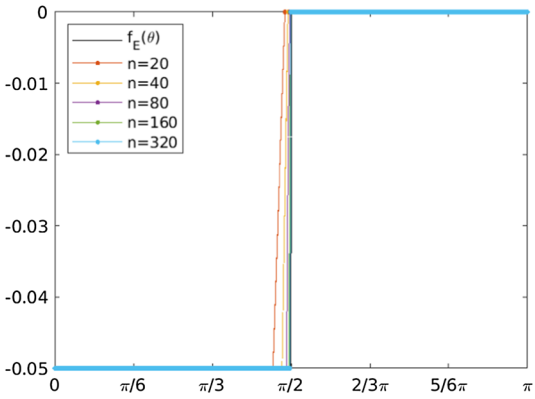



**Fig. 11** The spectrum of $\frac{1}{\Delta x} E_n$ with different number of cells versus sampling of the eigenvalue functions of $\mathscr{E}(\theta)$

where $n$ is the number of pressure cells (Fig. 11). The symbol associated to the scaled matrix sequence $\{\frac{1}{\Delta x} E_n\}_n$ is the function $\mathscr{E} : [-\pi, \pi] \to \mathbb{C}^{2\times 2}$ and can be written as

$$\mathscr{E}(\theta) = d \begin{bmatrix} -1 & e^{\mathbf{i}\theta} \\ e^{-\mathbf{i}\theta} & -1 \end{bmatrix}$$

and so its eigenvalues are 0 and $-2d$, while its eigenvectors are $\begin{pmatrix} e^{\mathbf{i}\theta} \\ \mathbf{i} \end{pmatrix}$ and $\begin{pmatrix} -e^{\mathbf{i}\theta} \\ \mathbf{i} \end{pmatrix}$. Since $E_n$ is real symmetric, by **GLT3** and **GLT1** we obtain

$$\left\{ \frac{1}{\Delta x} E_n \right\}_n \sim_{\mathrm{GLT},\sigma,\lambda} (\mathscr{E}, [-\pi, \pi]). \tag{21}$$

### 4.2 Spectral study of the Schur complement

We now study the spectral distribution of the Schur complement of $\mathcal{A}$. The formal expression of the Schur complement involves inversion of the $(1, 1)$ block of the matrix system and the multiplication by the $(1, 2)$ and $(2, 1)$ blocks that is: $S_n = E_n - D_{n,n+1} N_{n+1}^{-1} G_{n+1,n}$. To compute the symbol of the Schur complement sequence we need to compute the symbol of $\{(L_{n+1} + M_{n+1})^{-1}\}_n$. Thanks to relation (18) and to **GLT1-2** we have

$$\{(L_{n+1} + M_{n+1})^{-1}\}_n \sim_\lambda (\mathscr{L}^{-1}, [-\pi, \pi]) \tag{22}$$

with

$$\mathscr{L}^{-1}(\theta) = \frac{b}{1 - cos\theta} \begin{bmatrix} 8 & 1 & 0 & 0 \\ 1 & 8 & 0 & 0 \\ 0 & 0 & 8 & 1 \\ 0 & 0 & 1 & 8 \end{bmatrix}$$

where $b = \frac{560}{1701}\frac{1}{\mu dc}$. $\mathscr{L}^{-1}$ has two eigenvalue functions $\frac{9b}{1-cos\theta}$ and $\frac{7b}{1-cos\theta}$, each with multiplicity 2. Following (22), in Fig. 12 we compare the spectrum of $L_{n+1}^{-1}$ and of $(L_{n+1}+M_{n+1})^{-1}$ with a sampling of the eigenvalue functions of $\mathscr{L}^{-1}(\theta)$. In both cases, the spectra are well described by the sampling of the symbol eigenvalue functions.

At this point we can focus on the symbol of a properly scaled Schur complement sequence: $\{\frac{1}{\Delta t}S_n\}_n$. We know that $\frac{1}{\Delta t}S_n$ is a principal submatrix of

$$\tilde{S}_n := T_n\left(\frac{1}{c}\mathscr{E}\right) - T_n(\mathscr{D})T_n(\mathscr{L})^{-1}T_n(\mathscr{G}) + Z_n,$$

$Z_n$ being a correction-term. Since we are assuming that $c = \frac{\Delta t}{\Delta x} = \mathcal{O}(1)$ and since $\mathscr{L}(\theta)$ is an Hermitian positive definite matrix-valued function, by combining Theorem 16, and Eqs. (19), (20), (21) and (22) it holds that

$$\left\{T_n\left(\frac{1}{c}\mathscr{E}\right) - T_n(\mathscr{D})T_n(\mathscr{L})^{-1}T_n(\mathscr{G})\right\}_n \sim_{\sigma,\lambda} (\mathscr{S}, [-\pi,\pi])$$

where

$$\mathscr{S}(\theta) = \frac{1}{c}\mathscr{E}(\theta) - \mathscr{D}(\theta)\mathscr{L}^{-1}(\theta)\mathscr{G}(\theta) = \frac{d}{c}\begin{bmatrix} -1-5\frac{a}{\mu} & e^{\mathbf{i}\theta}-3\frac{a}{\mu} \\ e^{-\mathbf{i}\theta}-3\frac{a}{\mu} & -1-5\frac{a}{\mu}\end{bmatrix}$$

and $a = \frac{105}{2016}$. This combined with Theorem 9 guarantees that

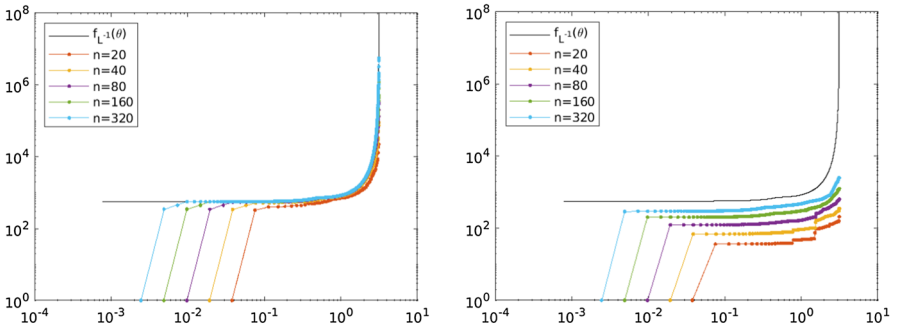
$$\left\{\tilde{S}_n\right\}_n \sim_\lambda (\mathscr{S}, [-\pi,\pi])$$

and consequently

$$\left\{\frac{1}{\Delta t}S_n\right\}_n \sim_\lambda (\mathscr{S}, [-\pi,\pi]). \tag{23}$$

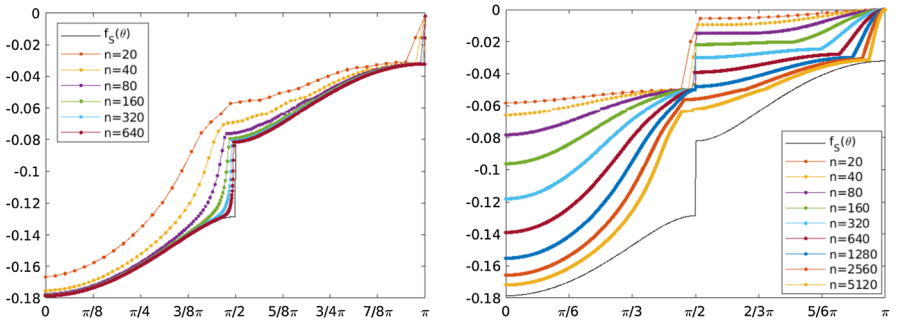
The eigenvalue functions of $\mathscr{S}(\theta)$ are $\frac{d}{c}\left(-1-5\frac{a}{\mu} \pm \sqrt{1+9\frac{a^2}{\mu^2}-6\frac{a}{\mu}cos\theta}\right)$. In Fig. 13 we compare a sampling of the eigenvalue functions of $\mathscr{S}(\theta)$ with the spectrum of $\frac{1}{\Delta t}S_n$ for different grid refinements. In the right panel, we consider the complete matrix $\mathcal{A}$ with $N_{n+1} = L_{n+1} + M_{n+1}$, while in the left panel we show the situation when replacing $N_{n+1}$ with $L_{n+1}$. Moreover, in Fig. 14 we compare the minimal eigenvalues of $-\frac{1}{\Delta t}S_n$ with functions of type $c\cdot\theta^\gamma$ and we see that for large $n$ the order $\gamma$ is approximately 2.

***Remark 20*** We stress that, thanks to the newly introduced Theorem 16, computing the symbol of the product $D_{n,n+1}N_{n+1}^{-1}G_{n+1,n}$ immediately follows by using standard spectral distribution tools as Theorem 9. The same result could be obtained following the much more involved approach used in [12]. Such approach asks to first extend




**(a)** $L_{n+1}^{-1}$ **(b)** $(L_{n+1} + M_{n+1})^{-1}$

**Fig. 12** The spectrum of $L_{n+1}^{-1}$ and $(L_{n+1} + M_{n+1})^{-1}$ vs the eigenvalue functions of $\mathscr{L}^{-1}(\theta)$




**(a)** $E_n - D_{n,n+1}L_{n+1}^{-1}G_{n+1,n}$ **(b)** $E_n - D_{n,n+1}(L+M)_{n+1}^{-1}G_{n+1,n}$

**Fig. 13** The spectrum of the matrix $\frac{1}{\Delta t}S_n$ with different number of cells versus sampling of the eigenvalue functions of the symbol $\mathscr{S}(\theta)$ In **a**, the (1,1) block contains only the $L_{n+1}$ term, while in **b** the block $N_{n+1}$ contains $L_{n+1} + M_{n+1}$

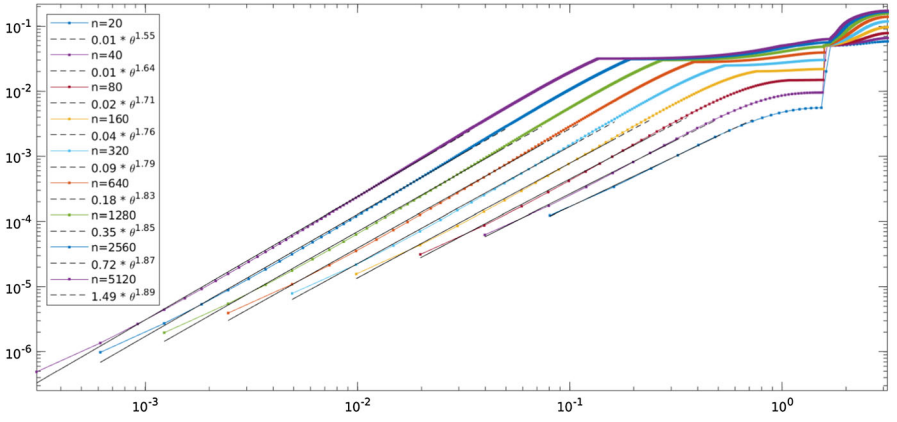



**Fig. 14** Smallest eigenvalues of $-\frac{1}{\Delta t}S_n$ and best fits with functions of the type $c \cdot \theta^{\gamma}$: for large $n$ the order $\gamma$ is, as expected, approximately 2

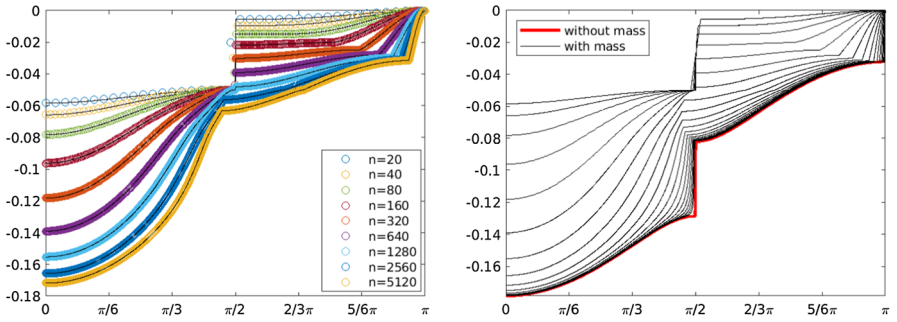



**Fig. 15** **a** The spectrum of the matrix $\frac{1}{\Delta t} S_n$ with different number of cells versus sampling of the eigenvalues of $\mathscr{S}_{\Delta x}(\theta)$, **b** visual convergence of the generating function $\mathscr{S}_{\Delta x}(\theta)$ (black lines) to $\mathscr{S}(\theta)$ (red line) as $\Delta x \to 0$ (color figure online)

the rectangular matrices $D_{n,n+1}$, $G_{n+1,n}$ to proper square block Toeplitz matrices, and then use the GLT machinery to compute the symbol of their product with $N_{n+1}^{-1}$. Finally, the symbol of the original product is recovered by projecting on the obtained matrix through ad hoc downsampling matrices and by leveraging the results on the symbol of projected Toeplitz matrices designed in the context of multigrid methods [33].

Aside from the symbol $\mathscr{S}(\theta)$, having in mind to build a preconditioner for the Schur matrix, we compute also the generating function of $\frac{1}{\Delta t} S_n$ for a fixed $n$, that is for a fixed $\Delta x$. Here we keep the contribution of the mass matrix in $N_{n+1}$, and, consequently, we introduce the dependence on the grid size. As a result, we get

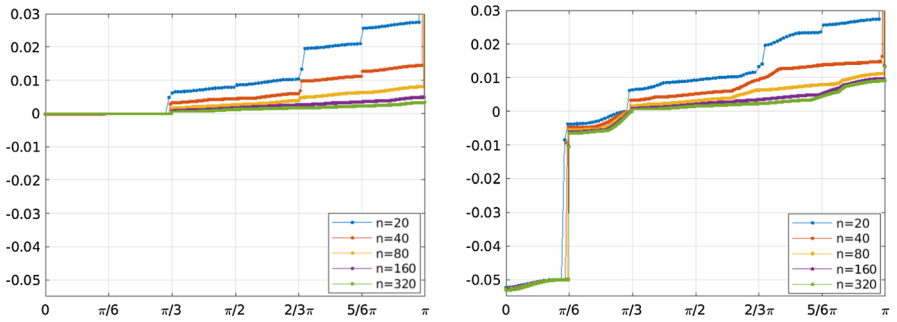
$$\mathscr{S}_{\Delta x}(\theta) = \frac{d}{c} \begin{bmatrix} -1 - (5a(\theta) - 3\Delta x\rho)b(\theta)c & e^{\mathbf{i}\theta} - (3a(\theta) - 5\Delta x\rho)b(\theta)c \\ e^{-\mathbf{i}\theta} - (3a(\theta) - 5\Delta x\rho)b(\theta)c & -1 - (5a(\theta) - 3\Delta x\rho)b(\theta)c \end{bmatrix} \quad (24)$$

with $a(\theta) = 6\,(1 - \cos\theta)\,\mu c + 2\Delta x\rho$ and $b(\theta) = \frac{15}{16} \frac{(1-\cos\theta)}{a(\theta)^2 - \Delta x^2 \rho^2}$. As shown in Fig. 15a, the sampling of the eigenvalue functions of $\mathscr{S}_{\Delta x}(\theta)$ perfectly matches the spectrum of the corresponding Schur matrix, and this paves the way to design a preconditioner that instead of $\mathscr{S}(\theta)$ involves $\mathscr{S}_{\Delta x}(\theta)$. Of course, in the limit when $\Delta x$ goes to zero, the symbol is equal to $\mathscr{S}(\theta)$. As a confirmation see Fig. 15b. The aim of this procedure is to obtain a good preconditioner, even when considering a coarse grid.

### 4.3 Spectral study of the coefficient matrix

The results obtained in Sects. 4.1–4.2 suggest to scale the coefficient matrix $\mathcal{A}$ by columns through the following matrix

$$V = \begin{bmatrix} I & 0 \\ 0 & \frac{1}{\Delta t} I \end{bmatrix},$$




**(a)** original $\mathcal{A}$ **(b)** scaled $\mathcal{A}V$

**Fig. 16** The spectrum of the coefficient matrix

that is to solve the system $\mathcal{A}_n\mathbf{x} = \mathbf{f}$, with $\mathcal{A}_n := \mathcal{A}V$ in place of system (7). As a result of the scaling, the blocks $\frac{1}{\Delta t}G_{n+1,n}$ and $\frac{1}{\Delta t}E_n$ of $\mathcal{A}_n$ have size $\mathcal{O}(1)$, similar to the size of $N_{n+1}$ and $D_{n,n+1}$, which remain unchanged. Moreover, the scaling improves the arrangement of the eigenvalues of $\mathcal{A}$ since the small negative eigenvalues are shifted towards negative values of larger modulus, as we can see in Fig. 16. Indeed, excluding the boundary conditions and due to the block-factorization

$$\mathcal{A}_n = WDW^T = \begin{bmatrix} I_{n+1} & 0 \\ D_{n,n+1}N_{n+1}^{-1} & I_n \end{bmatrix} \begin{bmatrix} N_{n+1} & 0 \\ 0 & \frac{1}{\Delta t}S_n \end{bmatrix} \begin{bmatrix} I_{n+1} & N_{n+1}^{-1}\frac{1}{\Delta t}G_{n+1,n} \\ 0 & I_n \end{bmatrix},$$

by the Sylvester inertia law we can infer that the signature of $\mathcal{A}_n$ is the same of the signature of the diagonal matrix formed by $N_{n+1}$ and $\frac{1}{\Delta t}S_n = \frac{1}{\Delta t}(E_n - D_{n,n+1}N_{n+1}^{-1}G_{n+1,n})$, which we know has negative eigenvalues distributed according to $\mathcal{S}(\theta)$.

In order to obtain the symbol of $\{\mathcal{A}_n\}_n$, when including also the boundary conditions, let us observe that $\mathcal{A}_n$ can be written as $\mathcal{A}_n = \tilde{\mathcal{A}}_n + \mathcal{Q}_n$, where $\mathcal{Q}_n$ is a correction term and $\tilde{\mathcal{A}}_n$ is a principal Hermitian submatrix (obtained removing the last 2 rows and the last 2 columns) of the matrix

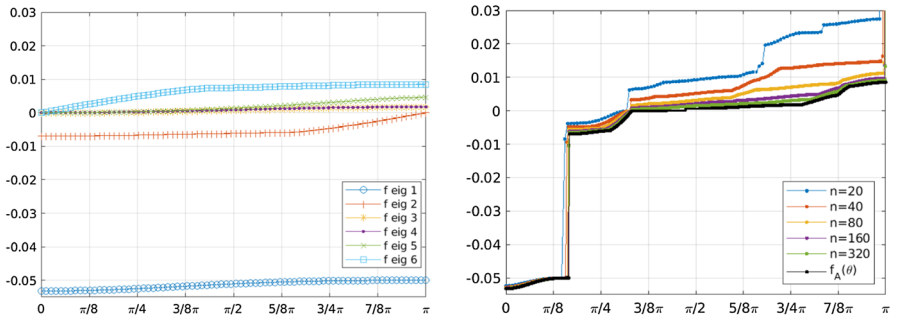
$$\begin{aligned} \mathcal{B}_n :&= \begin{bmatrix} T_n(\mathscr{L}) + \Delta x T_n(\mathscr{M}) & T_n(\mathscr{G}) \\ T_n(\mathscr{D}) & T_n(\frac{1}{c}\mathscr{E}) \end{bmatrix} \\ &= \begin{bmatrix} T_n(\mathscr{L}) & T_n(\mathscr{G}) \\ T_n(\mathscr{D}) & T_n(\frac{1}{c}\mathscr{E}) \end{bmatrix} + \Delta x \begin{bmatrix} T_n(\mathscr{M}) & O \\ O & O \end{bmatrix}. \end{aligned}$$

Now, by Theorem 17, the two involved matrices are similar that is

$$\mathcal{B}_n \sim T_n(\mathscr{F}) + \Delta x T_n(\mathscr{C})$$

with $\mathscr{F} := \begin{bmatrix} \mathscr{L} & \mathscr{G} \\ \mathscr{D} & \frac{1}{c}\mathscr{E} \end{bmatrix}$ and $\mathscr{C} := \begin{bmatrix} \mathscr{M} & 0 \\ 0 & 0 \end{bmatrix}$. Therefore,

$$\{\mathcal{B}_n\}_n \sim_\lambda (\mathscr{F}, [-\pi, \pi]),$$




**(a)** **(b)**

**Fig. 17** **a** A plot of the eigenvalue functions of $\mathscr{F}(\theta)$ made without knowing their analytical expression, **b** The spectrum of the scaled coefficient matrix $\mathcal{A}V$ with different number of cells versus the sampling of the eigenvalue functions of $\mathscr{F}(\theta)$

and this, thanks to Theorem 8, implies that

$$\{\tilde{\mathcal{A}}_n\}_n \sim_\lambda (\mathscr{F}, [-\pi, \pi]).$$

Finally, by following the same argument applied in the computation of the Schur complement symbol at the beginning of Sect. 4.2, by using again Theorem 9, we arrive at

$$\{\mathcal{A}_n\}_n \sim_\lambda (\mathscr{F}, [-\pi, \pi]).$$

In conclusion, the correction term $\mathcal{Q}_n$ does not affect the symbol of the matrix sequence $\{\mathcal{A}_n\}$ and the eigenvalues of $\{\mathcal{A}_n\}_n$ are distributed in the same way as the eigenvalues of $\{\tilde{\mathcal{A}}_n\}_n$. Since the symbol $\mathscr{F}$ is a $6 \times 6$ matrix-valued function, retrieving an analytical expression for its eigenvalue functions asks for some extra computation, but we can easily give a numerical representation of them which is sufficient for our aims simply following these three steps:

- evaluate the symbol $\mathscr{F}$ on an equispaced grid in $[0, \pi]$;
- for each obtained $6 \times 6$ matrix compute the spectrum;
- take all the smallest eigenvalues as a representation of $\lambda_1(\mathscr{F})$ and so on so forth till the largest eigenvalues as a representation of $\lambda_6(\mathscr{F})$.

Figure 17a has been realized following the previous steps. Notice that two eigenvalue functions of $\mathscr{F}$ show the same behavior and we suspect they indeed have the same analytical expression. Figure 17b compares the equispaced sampling of the eigenvalue functions with the actual eigenvalues of the coefficient matrix and highlights an improving matching as the matrix-size increases.

***Remark 21*** The eigenvalue structure in the general case of a variable cross-section $d = d(x)$ does not pose technical problems and in reality it is perfectly covered by the GLT theory: more specifically, we refer to item **GLT1** where the GLT symbol depends

on $(x, \theta) \in [0, 1] \times [-\pi, \pi]$ and where $x$ is in our context exactly the scaled physical variable of the coefficient $d = d(x)$.

The case of a variation of the degrees $n_x, n_y$ is more delicate to treat, since, in this setting, the size of the basic small blocks of the matrix is affected. This is the parameter $s$ defining the range $\mathbb{C}^{s\times s}$ of the symbol $\kappa$ in the GLT theory (see Sect. 3). Despite the theoretical difficulty of treating a varying parameter $s$ for a precise spectral analysis, as shown in the next section, the performances of our preconditioning techniques are satisfactory also in this tricky setting.

***Remark 22*** Our discretization can be extended to three-dimensional pipes by introducing tensor product shape functions in the transverse plane, using polynomial degrees $n_y$ and $n_z$ for the velocity. Leaving fixed $n_x = 1$ for the pressure variable, our theory should extend to this more general setting and yield a symbol for the $(1, 1)$ block of the coefficient matrix with values in $\mathbb{C}^{2(n_y-1)(n_z-1)\times 2(n_y-1)(n_z-1)}$, symbols for $(1, 2)$ and $(2, 1)$ blocks in $\mathbb{C}^{2(n_y-1)(n_z-1)\times 2}$ and $\mathbb{C}^{2\times 2(n_y-1)(n_z-1)}$ respectively. In any case, the symbol for $(2, 2)$ block and the Schur complement will still take values in $\mathbb{C}^{2\times 2}$ independently of $n_y$ and $n_z$. The size $2 \times 2$ for the symbol of the Schur complement is controlled by the choice of $n_x = 1$ for the pressure variable, and for larger $n_x$ the symbol of the Schur complement should take values in $\mathbb{C}^{(n_x+1)\times(n_x+1)}$.

## 5 Numerical experiments

In this section we focus on the solution of system (7) by leveraging the spectral findings in Sect. 4 and with the help of the PETSc [2,3] library. To ease the notation, here after we omit the subscripts for the blocks $N_{n+1}$, $G_{n+1,n}$, $D_{n,n+1}$, $E_n$ of $\mathcal{A}$. The main solver for $\mathcal{A}_n = \mathcal{A}V$, say $\mathcal{K}_{\mathcal{A}}$, is GMRES and the preconditioner of this Krylov solver is based on the Schur complement; more precisely, an application of the preconditioner consists in solving

$$\hat{S}\hat{p} = r_p - D\widetilde{N^{-1}}r_u \qquad \hat{u} = \widetilde{N^{-1}}\left(r_u - \frac{1}{\Delta t}Gr_p\right)$$

where the block vector $\left(\begin{smallmatrix} r_u \\ r_p \end{smallmatrix}\right)$ is the residual.

If the inversion of $N$ were exact and $\hat{S}$ were the exact Schur complement of $\mathcal{A}_n$, the main solver $\mathcal{K}_{\mathcal{A}}$ would of course be a direct method. Here above, instead, $\widetilde{N^{-1}}$ denotes the application of a suitable Krylov solver, say $\mathcal{K}_N$, to the linear operator $N$ and in our numerical experiments this was chosen as GMRES with a relative stopping tolerance $10^{-5}$ and ILU(0) preconditioner, since $N$ is a narrow-banded matrix. Further, the Schur complement is approximated by $\hat{S} = \frac{1}{\Delta t}(E - D\widetilde{N^{-1}}G)$. However, since the inverse of $N$ is approximated by the action of the solver $\mathcal{K}_N$, matrix $\hat{S}$ cannot be explicitly assembled, although its action on any vector can be computed with a call to $\mathcal{K}_N$.

The solution of the system with matrix $\hat{S}$ required in the preconditioner inside $\mathcal{K}_{\mathcal{A}}$ is then performed with a Krylov solver, say $\mathcal{K}_{\hat{S}}$. In $\mathcal{K}_{\hat{S}}$, the matrix-vector multiplication is performed as described above, while the preconditioner is the block circulant

preconditioner generated by $\mathscr{S}_{\Delta x}(\theta)$ given in (24), that is (see Theorem 11)

$$C_n(\mathscr{S}_{\Delta x}) = (F_n \otimes I_2) D_n(\mathscr{S}_{\Delta x})(F_n^* \otimes I_2)$$

with

$$D_n(\mathscr{S}_{\Delta x}) = \mathrm{diag}_{r=0,\ldots,n-1}(\mathscr{S}_{\Delta x}(\theta_r)), \quad F_n = \frac{1}{\sqrt{n}}\left[e^{-\mathbf{i}j\theta_r}\right]_{j,r=0}^{n-1}, \quad \theta_r = \frac{2\pi r}{n}.$$

More precisely, since $\mathscr{S}_{\Delta x}(\theta)$ has a unique zero eigenvalue at $\theta_0 = 0$, we use as preconditioner

$$\mathcal{C}_n := C_n(\mathscr{S}_{\Delta x}) + \frac{1}{(2n)^2}\mathbf{1}^T\mathbf{1} \otimes \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \tag{25}$$

with $\mathbf{1} = [1, \ldots, 1] \in \mathbb{R}^n$, that is we introduce a circulant rank-one correction aimed at avoiding singular matrices. We notice that $\{\mathcal{C}_n\}_n$ and the sequence of the Schur complements are GLT matrix sequences having the same symbol, i.e., $\mathscr{S}(\theta)$. Therefore, since $\mathscr{S}(\theta)$ is not singular by **GLT2** we infer that the sequence of the preconditioned matrices is a GLT with symbol 1. Given the one-level structure of the involved matrices, we expect that the related preconditioned Krylov solvers converge within a constant number of iterations independent of the matrix-size, just because the number of possible outliers is bounded from above by a constant independent of the mesh-size. Hence the global cost is given by $\mathcal{O}(n \log n)$ arithmetic operations when using the standard FFT based approach for treating the proposed block circulant preconditioner.

It is worth mentioning that the coefficient matrix, as well as all its blocks, are sparse matrices, then matrix-vector product with the original matrix has optimal cost of $\mathcal{O}(n)$ arithmetic operations. Reducing the cost of $\mathcal{O}(n \log n)$ of each preconditioned iteration to the optimal cost is possible by using specialized multigrid solvers designed ad hoc for circulant structures [33].

The full solver is applied with the help of the PETSc libraries, choosing `fgmres` as main solver, the `fieldsplit` preconditioner of `schur` type with `full` factorization, selecting `gmres` as the Krylov solver $\mathcal{K}_{\hat{S}}$ and providing a `shell` preconditioner that implements the solution of the circulant system (see [2]). The standard GMRES implementation in the PETSc library by default restarts after 30 iterations. We have not changed the default behavior since this is not affecting our computations: 30 iterations are never reached with our preconditioner, as we can see in subsequent tests. The circulant preconditioner is applied with the help of the FFTW3 library [17], observing that the action of the tensor product of a discrete Fourier matrix and $I_2$ corresponds to the computation of two FFT tranforms of length $n$ on strided subvectors. In our numerical tests, a relative stopping tolerance of $10^{-6}$ was chosen for $\mathcal{K}_{\hat{S}}$.

As comparison solver we consider another preconditioning technique that does not require to assemble the Schur complement, namely the Least Squares Commutators (LSC) of [14,35]. It is based on the idea that one can approximate the inverse of the

Schur complement, without considering the contribution of the block $E$, by

$$\bar{S}^{-1} = \frac{1}{\Delta t}\widetilde{(DG)^{-1}} DNG \widetilde{(DG)^{-1}}.$$

Matrix $\overline{S}$ is never assembled, but the action of $\bar{S}^{-1}$ is computed with the above formula, where we have indicated with $\widetilde{(DG)^{-1}}$ the application of a solver for the matrix $\frac{1}{\Delta t}DG$, which we denote with $\mathcal{K}_{DG}$. In our tests, we have chosen for $\mathcal{K}_{DG}$ a preconditioned conjugate gradient solver with relative stopping tolerance of $10^{-5}$, since, in the incompressible framework, the product $\frac{1}{\Delta t}DG$ is a Laplacian. To provide a circulant preconditioner for $\mathcal{K}_{DG}$, it is enough to consider the block circulant matrix generated by $\mathscr{D}(\theta)\mathscr{G}(\theta)$ defined as in Remark 19. Note that, for $\theta = 0$, $\mathscr{D}(\theta)\mathscr{G}(\theta)$ is the null matrix, therefore in order to avoid singular matrices we introduce a rank-two correction and define the whole preconditioner for the product $\frac{1}{\Delta t}DG$ as

$$\mathcal{P}_n := C_n(\mathscr{D}\mathscr{G}) + \frac{1}{(2n)^2}\mathbf{1}^T\mathbf{1} \otimes I_2 \tag{26}$$

again with $\mathbf{1} = [1, \ldots, 1] \in \mathbb{R}^n$.

For a complete Navier-Stokes simulation, the solver $\mathcal{K}_{\mathcal{A}}$ is applied at each iteration of the main non-linear Picard solver that computes a timestep. In all numerical tests, $\mathcal{K}_{\mathcal{A}}$ is a FGMRES solver with relative tolerance of $10^{-8}$.

It should be noted that the literature provides a quite limited theory regarding the FGMRES convergence. In particular, the considered method may give slow convergence or break down: however, in the present setting, the convergence behavior in terms of iteration count and CPU timing of the FGMRES has been very satisfactory and competitive with the more standard preconditioned Krylov techniques.

*Flow between parallel plates.* In the first test we consider a 2D domain with constant cross-section $d(x) = 0.025$ m. In inlet we impose a parabolic velocity profile with flow rate $5 \times 10^{-6}\text{m}^2/\text{s}$, while at the outlet we fix a null pressure. Of course there would be no need to use a numerical model to compute the solution in this particular geometry, since an exact solution is known, but we conduct this as a test to verify the performance of our solver. Using $n_x = 1$ and $n_y = 3$ this setting is exactly the one adopted in Sects. 3 and 4.

The main solver $\mathcal{K}_{\mathcal{A}}$ converges in at most 2 iterations, while the number of iterations of $\mathcal{K}_{\hat{S}}$ stays constant as the number of cells grows which confirms that the block circulant preconditioner $\mathcal{C}_n$ in (25) is optimal, Table 1, even if the analysis and therefore the optimality have been proved only at the asymptotic level. For this example we also check the performances of the block circulant preconditioner $C_n(\mathscr{S})$ in $\mathcal{K}_{\hat{S}}$, that is, without taking into account the contribution of the mass matrix. Looking again at Table 1, we see that in this case the inner solver $\mathcal{K}_{\hat{S}}$ does not converge when the number of cells increases. The discrepancy in the performances of $C_n(\mathscr{S})$ compared with those of $\mathcal{C}_n$ is in line with the results in Fig. 15a that clearly show how good $\mathscr{S}_{\Delta x}$ matches the spectrum of the Schur complement compared with $\mathscr{S}$.

**Table 1** Iterations of the solvers in the 2D parallel plates test

| $n$ | $\mathcal{C}_n$ | | | $C_n(\mathscr{S})$ | | LSC with $\mathcal{P}_n$ | | | |
|---|---|---|---|---|---|---|---|---|---|
| | $\mathcal{K}_{\mathcal{A}}$ | $\mathcal{K}_{\hat{S}}$ | Time (s) | $\mathcal{K}_{\mathcal{A}}$ | $\mathcal{K}_{\hat{S}}$ | $\mathcal{K}_{\mathcal{A}}$ | $\mathcal{K}_{\bar{S}}$ | $\mathcal{K}_{DG}$ | Time (s) |
| 10 | 2 | 11–12 | $2.50 \times 10^{-2}$ | 2 | 15–16 | 2 | 2–10 | 5–6 | $2.08 \times 10^{-1}$ |
| 20 | 2 | 10–11 | $1.52 \times 10^{-1}$ | 2 | 20 | 2 | 4–12 | 5–6 | $1.58 \times 10^{0}$ |
| 40 | 2 | 9–11 | $2.97 \times 10^{-1}$ | 2 | 24 | 2 | 3–14 | 6–7 | $3.70 \times 10^{0}$ |
| 80 | 2 | 9–10 | $5.51 \times 10^{-1}$ | 2 | 31 | 2 | 3–14 | 5–7 | $7.85 \times 10^{0}$ |
| 160 | 2 | 8–9 | $1.42 \times 10^{0}$ | 2 | no conv. | 2 | 1–14 | 4–8 | $2.06 \times 10^{1}$ |
| 320 | 2 | 8–9 | $7.46 \times 10^{0}$ | 2 | no conv. | 3 | 1–21 | 6–8 | $2.42 \times 10^{2}$ |
| 640 | 2 | 7–9 | $4.94 \times 10^{1}$ | 2 | no conv. | 4 | 7–24 | 3–10 | $2.65 \times 10^{3}$ |
| 1280 | 2 | 7–8 | $3.68 \times 10^{2}$ | 2 | no conv. | 7 | 9–28 | 3–10 | $4.55 \times 10^{4}$ |

$\mathcal{K}_{\hat{S}}$ refers to our approach, while $\mathcal{K}_{\bar{S}}$ and $\mathcal{K}_{DG}$ refer to the LSC approach. The times are the total CPU time spent in the main Krylov solver $\mathcal{K}_{\mathcal{A}}$ and its sub-solvers

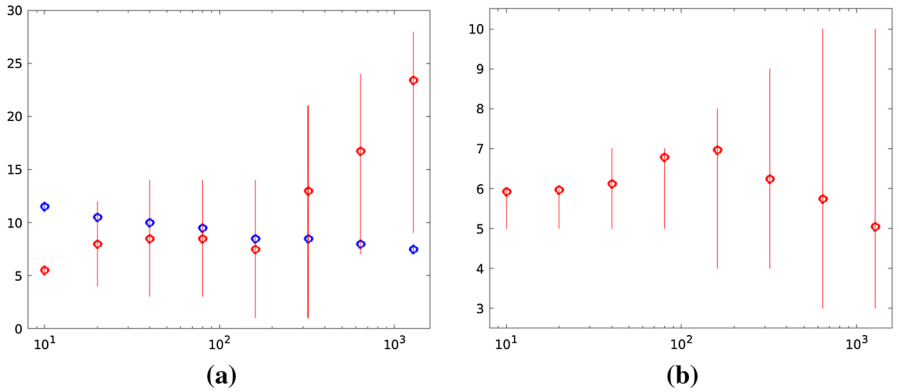



**Fig. 18** **a** The average number and the range of iterations of $\mathcal{K}_{\hat{S}}$ in blue and of $\mathcal{K}_{\bar{S}}$ in red; **b** The average number and the range of iterations of $\mathcal{K}_{DG}$ (color figure online)

Concerning the LSC approach, the number of iterations of $\mathcal{K}_{DG}$ does not grow significantly with $n$, indicating that the block circulant preconditioner $\mathcal{P}_n$ in (26) for $\frac{1}{\Delta t} DG$ is optimal, see also Fig. 18b. The full solver for $\mathcal{A}_n$, however, needs considerably more time to reach the required tolerance, for two reasons: (1) the number of iterations of $\mathcal{K}_{\hat{S}}$ in our approach is lower than those of $\mathcal{K}_{\bar{S}}$ in LSC (see Fig. 18a); (2) the LSC approach invokes the inner solver $\mathcal{K}_{DG}$ twice per each iteration of $\mathcal{K}_{\bar{S}}$, affecting the final computation time.

*Flow between converging plates.* In this second test we consider a 2D domain with variable cross-section, where $d(x)$ decreases linearly from 0.025 to 0.0125 m. To perform the simulations we impose the same boundary conditions as in the previous test and again take $n_x = 1, n_y = 3$. In Table 2 we compare the number of iterations computed by $\mathcal{K}_{\hat{S}}$ considering as preconditioners

**Table 2** Iterations of the solvers in the 2D converging plates test, i.e. with variable $d(x)$

| $n$ | $d(x)$ in $\mathcal{K}_{\hat{S}}$ | | | $d(x) = \bar{d}$ in $\mathcal{K}_{\hat{S}}$ | | |
|---|---|---|---|---|---|---|
| | Non linear solver | $\mathcal{K}_{\mathcal{A}}$ | $\mathcal{K}_{\hat{S}}$ | Non linear solver | $\mathcal{K}_{\mathcal{A}}$ | $\mathcal{K}_{\hat{S}}$ |
| 10 | 6 | 1–2 | 12–13 | 6 | 1–2 | 14–15 |
| 20 | 5 | 1–2 | 11–12 | 5 | 1–2 | 15–16 |
| 40 | 3 | 1–2 | 10–12 | 3 | 1–2 | 14–16 |
| 80 | 2 | 1–2 | 9–11 | 2 | 1–2 | 13–16 |
| 160 | 2 | 1–2 | 9–11 | 2 | 1–2 | 13–16 |
| 320 | 2 | 1–2 | 9–11 | 2 | 1–2 | 13–16 |
| 640 | 2 | 1–2 | 9–11 | 2 | 1–2 | 13–16 |
| 1280 | 2 | 1–2 | 9–11 | 2 | 1–2 | 13–17 |

In the left part, we use a diagonal scaling, defined through $d(x)$, of the block circulant preconditioner $\mathcal{C}_n$; on the right, we use $\mathcal{C}_n$ with $d = \bar{d}$, that is equal to the average of the cross-section along the pipe

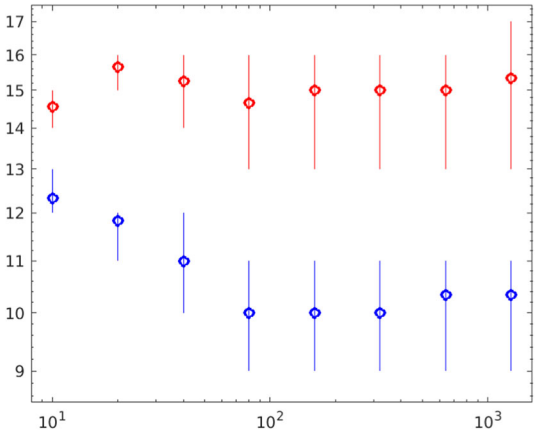



**Fig. 19** The average number and the range of iteration of $\mathcal{K}_{\hat{S}}$ for a 2D pipe with variable cross-section. The blue values are obtained employing as preconditioner in $\mathcal{K}_{\hat{S}}$ a diagonal scaling (defined through $d(x)$) of the block circulant preconditioner $\mathcal{C}_n$; the red values are obtained using $\mathcal{C}_n$ with $d = \bar{d}$, that is equal to the average of the cross-section along the pipe (color figure online)

1. $\mathcal{D}_n(\frac{1}{d}\mathcal{C}_n(\mathscr{S}_{\Delta x}) + \mathcal{R}_n)$, with $\mathcal{D}_n$ a diagonal matrix whose entries are an equispaced sampling of $d(x)$ on its domain (see Remark 21), and $\mathcal{R}_n = \frac{1}{(2n)^2}\mathbf{1}^T\mathbf{1} \otimes \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$;
2. $\mathcal{C}_n$ with $d = \bar{d}$, that is equal to the average of the cross-section along the pipe.

In the first case the $\mathcal{K}_{\hat{S}}$ converges in a number of iterations that does not increase significantly with $n$, showing its optimality. Approximating the channel width with a constant value instead, avoids the diagonal matrix multiplication in the preconditioner, but causes a slightly faster increase of the iteration counts for $\mathcal{K}_{\hat{S}}$, refer to Fig. 19.

*Using higher polynomial degree in the transversal direction.* In this test we analyse the efficiency of the preconditioner $\mathcal{C}_n$ in $\mathcal{K}_{\hat{S}}$ when considering different polynomial degrees $n_y$ in the transversal direction for the velocity, but fixed $n_x = 1$ for the pressure variable. In this setting, we expect symbols for (1,1) block of the coefficient matrix to take values in $\mathbb{C}^{2(n_y-1)\times 2(n_y-1)}$, those for (1, 2) and (2, 1) blocks in $\mathbb{C}^{2(n_y-1)\times 2}$ and $\mathbb{C}^{2\times 2(n_y-1)}$ respectively, while those for the (2, 2) block and the Schur complement

**Table 3** Range of iterations for the non-linear solver and $\mathcal{K}_{\hat{S}}$, in the converging plates test, with different polynomial degree in the transversal direction for the velocity

| $n$ | $n_y = 4$ | | $n_y = 5$ | | $n_y = 6$ (vecchio!) | |
|---|---|---|---|---|---|---|
| | Non linear solver | $\mathcal{K}_{\hat{S}}$ | Non linear solver | $\mathcal{K}_{\hat{S}}$ | Non linear solver | $\mathcal{K}_{\hat{S}}$ |
| 10 | 7 | 12 | 8 | 11–12 | 9 | 11–12 |
| 20 | 6 | 11–12 | 7 | 10–12 | 8 | 10–12 |
| 40 | 4 | 10–12 | 4 | 10–12 | 4 | 10–11 |
| 80 | 3 | 9–11 | 3 | 9–11 | 4 | 9–11 |
| 160 | 4 | 9–11 | 4 | 9–11 | 4 | 9–11 |
| 320 | 4 | 9–11 | 4 | 9–11 | 4 | 9–12 |
| 640 | 4 | 9–12 | 4 | 9–12 | 3 | 9–12 |
| 1280 | 3 | 9–12 | 3 | 9–12 | 3 | 9–12 |

will still take values in $\mathbb{C}^{2\times 2}$, irrespectively of $n_y$. On such basis, we can readily apply $\mathcal{C}_n$ in $\mathcal{K}_{\hat{S}}$ being sure that the sizes of all the involved matrices are consistent.

Taking again the converging plates case, we increase $n_y$ to 4, 5 and 6 and report the results in Table 3. We note that, despite the "looser" approximation in the preconditioner, the solver $\mathcal{K}_{\hat{S}}$ still converges in an almost constant number of iterations when $n$ increases. The number of iterations of $\mathcal{K}_A$ is always 2 and was thus not reported in the table. From this example we can infer that the symbol of the preconditioner for the Schur complement is not changing much as far as $n_x$ stays fixed to 1.

*3D square nozzle.* To perform a three-dimensional test, we consider a square pipe with width decreasing linearly from 0.025 to 0.0125 m, so that the square section area decreases quadratically from $6.25 \times 10^{-4}$ to $1.56 \times 10^{-4}$ m$^2$. At the inlet we fix a constant flow rate of $5 \times 10^{-6}$ m$^3$/s with a parabolic profile in both the transverse directions. We point out that, the product of two parabolic profiles in the $y$ and the $z$ directions would not be an exact solution even in a constant cross-section case, see [25]. The solution is computed using different combinations of transverse polynomial degrees $n_y$ and $n_z$ for the velocity, fixed $n_x = 1$ for the pressure variable.

Thanks to the matrix-sizes match pointed out in Remark 22, one could be tempted to directly apply in $\mathcal{K}_{\hat{S}}$ the preconditioner $\mathcal{C}_n$ derived for the two-dimensional case also in this three-dimensional setting. However such choice causes high iteration numbers and sometimes stagnation of the outer nonlinear solver (results not reported).

The reason for these poor performances may be understood by noticing that the two dimensional discretization represents a flow between infinite parallel plates at a distance $d(x)$. It is not surprising that using such a flow to precondition the computation in a three dimensional pipe is not optimal. More precisely, the two dimensional setting can be understood as choosing $n_z = 0$ in 3D. However, constant shape functions in the $z$ direction can not match the zero velocity boundary condition on the channel walls and only $n_z \geq 2$ would allow to satisfy them.

Fixing $n_y = 3, n_z = 2$ and following the same steps of Sect. 4, we have computed an ad hoc block circulant preconditioner for the three-dimensional case. For this special choice of $n_y$ and $n_z$ the symbols of the various matrices involved in the discretization

are matrix-valued with the same size as in Sect. 4, but now the generating function associated with the scaled Schur complement $\frac{1}{\Delta t} S_n$ shows a dependency on the cross-sectional area and is given by

$$\mathscr{S}_{\Delta x}(\theta) = \frac{Area}{c} \begin{bmatrix} -1 - (5a(\theta) - 3\Delta x\rho)b(\theta)c & e^{\mathbf{i}\theta} - (3a(\theta) - 5\Delta x\rho)b(\theta)c \\ e^{-\mathbf{i}\theta} - (3a(\theta) - 5\Delta x\rho)b(\theta)c & -1 - (5a(\theta) - 3\Delta x\rho)b(\theta)c \end{bmatrix}, \tag{27}$$

where $a(\theta) = 6\,(1 - \cos\theta)\,\mu c + 2\Delta x\rho$ and $b(\theta) = \frac{25}{96} \frac{(1-\cos\theta)}{a(\theta)^2 - \Delta x^2 \rho^2}$. This symbol is very similar to the one in (24), but the different constant in the function $b(\theta)$ reflects the presence of non trivial velocity shape functions in the $z$ direction.

Therefore, we use as preconditioner in $\mathcal{K}_{\hat{S}}$ the block circulant matrix generated by $\mathscr{S}_{\Delta x}(\theta)$ defined as in (27) properly shifted by a rank-one block circulant matrix and scaled by a diagonal matrix whose entries are given by a sampling of the function that defines the cross-sectional area of the pipe.

Table 4 shows the range of iterations for $\mathcal{K}_{\mathcal{A}}$ and $\mathcal{K}_{\hat{S}}$. In the left part we have applied the 3D block circulant preconditioner to the corresponding simulation with $n_y = 3$ and $n_z = 2$. As in the two-dimensional cases, the number of iterations of $\mathcal{K}_{\hat{S}}$ does not change significantly with $n$, in particular, already with 80 cells the range of iterations relative to the solver $\mathcal{K}_{\hat{S}}$ reaches optimality. Moreover, the nonlinear solver performs an higher number of iterations (compare with Table 2) for low $n$, but they reduce fast with the increasing resolution.

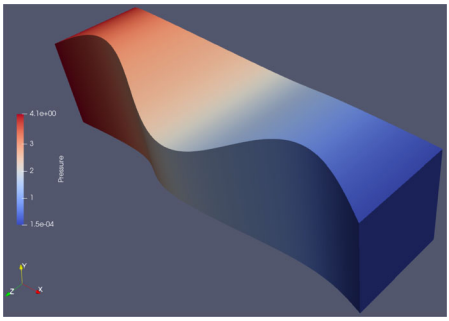
In the central and right part of the table we check the performance of the 3D block circulant preconditioner based on (27) in the discretizations for $n_y = n_z = 3$ and $n_y = n_z = 4$, respectively. As in the two-dimensional examples, for $n_y = n_z = 3$, the iteration numbers remain basically unchanged, despite the fact that the preconditioner is based on $\mathscr{S}_{\Delta x}(\theta)$ in (27) which corresponds to a different number of degrees of freedom. For $n_y = n_z = 4$ the number of iterations of $\mathcal{K}_{\hat{S}}$ are still quite the same, but the nonlinear solver has more problems in its convergence history: it requires an higher number of iterations at low resolution, compared to the other two cases, but it gets better when increasing $n$. This is suggesting that the actual generating function of the Schur complement for this case departs more from the one in (27) than for the case $n_y = n_z = 3$.

*3D pipe with generic geometry.* to highlight the potential of the circulant preconditioner and to confirm its effectiveness even in cases very different from the one in which it was computed, we consider a pipe in which the height and width vary as two out-of-phase sinusoidal functions (see Fig. 20), and are respectively $0.0125 + \sin(3\pi x/L)/200$ and $0.0125 - \sin(3\pi x/L)/200$. The length of the channel, as well as the inlet flow rate, are chosen as for the previous simulations, i.e. respectively equal to $L = 0.1$ m and $5 \times 10^{-6}\,\mathrm{m}^3/\mathrm{s}$. We performed the simulation with $n_y = n_z = 4$ to obtain a good representation of the velocity profile, which departs substantially from a parabolic profile.

In the right part of Fig. 20 we show the range of iterations of the solvers as a function of the number of cells. We can observe that the nonlinear solver performs much more iterations than in the previous tests until a sufficiently fine resolution is reached; on

**Table 4** Range of iterations for $\mathcal{K}_{\mathcal{A}}$ and $\mathcal{K}_{\hat{S}}$, in a 3D pipe with variable cross-section, with different polynomial degrees in the transversal directions for the velocity

| $n$ | $n_y = 3, n_z = 2$ | | | $n_y = 3, n_z = 3$ | | | $n_y = 4, n_z = 4$ | | |
|---|---|---|---|---|---|---|---|---|---|
| | Non linear solver | $\mathcal{K}_{\mathcal{A}}$ | $\mathcal{K}_{\hat{S}}$ | Non linear solver | $\mathcal{K}_{\mathcal{A}}$ | $\mathcal{K}_{\hat{S}}$ | Non linear solver | $\mathcal{K}_{\mathcal{A}}$ | $\mathcal{K}_{\hat{S}}$ |
| 10 | 13 | 1 | 10 | 13 | 1–2 | 10 | 27 | 1–2 | 11–12 |
| 20 | 8 | 1–2 | 10–12 | 8 | 1–2 | 10–12 | 34 | 1–2 | 11–13 |
| 40 | 3 | 1–2 | 10–12 | 3 | 1–2 | 11–12 | 37 | 1–2 | 11–13 |
| 80 | 3 | 1–2 | 11–12 | 3 | 1–2 | 11–12 | 19 | 1–2 | 11–13 |
| 160 | 2 | 2 | 11–12 | 2 | 2 | 11–12 | 4 | 1–2 | 11–13 |
| 320 | 2 | 2 | 11–12 | 2 | 2 | 11–12 | 3 | 1–2 | 11–14 |
| 640 | 2 | 2 | 11–13 | 2 | 2 | 11–13 | 2 | 2 | 11–14 |
| 1280 | 2 | 2 | 11–13 | 2 | 2 | 11–13 | 2 | 2 | 11–14 |




| $n$ | $n_y = 4$, $n_z = 4$ non linear solver | $\mathcal{K}_{\mathcal{A}}$ | $\mathcal{K}_{\hat{S}}$ |
|---|---|---|---|
| 20 | 36 | 1–2 | 9 – 11 |
| 40 | 38 | 1–2 | 9 – 10 |
| 80 | 26 | 1–2 | 10 – 11 |
| 160 | 15 | 1–2 | 10 – 12 |
| 320 | 10 | 1–2 | 10 – 12 |
| 640 | 7 | 2 | 10 – 12 |
| 1280 | 2 | 2 | 9 – 13 |

**Fig. 20** 3D pipe with generic geometry. Left: computed pressure. Right: range of iterations for $\mathcal{K}_{\mathcal{A}}$ and $\mathcal{K}_{\hat{S}}$

the other hand, the linear solvers still appear to be optimal and their iteration numbers are still very low.

## 6 Conclusion and perspectives

The incompressible Navier–Stokes equations have been solved in a pipe, using a Discontinuous Galerkin discretization over one-dimensional staggered grids. The approximation of the flow is achieved by discretization only along the pipe axis, leveraging only on high polynomial degrees in the transverse directions. The resulting linear systems have been studied both in terms of the associated matrix structure and in terms of the spectral features of the related coefficient matrices. In fact, the resulting matrices are of block type, each block shows Toeplitz-like, band, and tensor structure at the same time.

We have introduced a new technique to spectrally studying sequences of (possibly rectangular) Toeplitz and inverses of Toeplitz matrices. With the help of these theoretical findings we could describe the spectrum of the Schur complement matrices, without resorting to the technique of embedding the problem in a larger square one already present in the literature. Our target was to design and analyze fast iterative solvers for the associated large linear systems. At this stage we limited ourselves to the case of block circulant preconditioners in connection with Krylov solvers: the spectral clustering at 1 has been proven and the computational counterpart has been checked in terms of constant number of iterations and in terms of the whole arithmetic cost. A rich set of numerical experiments have been presented, commented, and critically discussed. In particular, despite the asymptotic nature of the proven results, the numerical experiments showed very good performances for the designed preconditioners. In fact the number of iterations for the main solver and for the inner Schur complement solver does not increase with the grid size starting from as few as 20 computational cells. This lucky behavior is not surprising after many research papers on the subject and under various conditions (local methods such as Finite Differences, Finite Elements, Discontinuous Galerkin, Isogeometric Analysis, Finite Volumes and different types of operators; see e.g. [5,10–13,18,20,21] and references therein). Even

if the GLT spectral theory is only (or at least essentially) of asymptotic nature, often it happens that the actual behavior of the related spectra stabilizes to the associated symbol already for moderate sizes. Moreover, in the preconditioning context a cluster around the symbol of radius 0.1 is often very good for convergence purposes of the related (preconditioned) Krylov solvers, even if an error of 0.1 cannot be considered very small in the classical theory of function approximation. Hence the combination of these two aspects makes the GLT tools already useful for moderate matrix dimensions.

For sure a quantitative theoretical analysis of the convergence speed of the discrete distributions to the symbol would be precious, but it is difficult in general, especially for more involved geometries, and at the moment it is out of the scope of the present work. It will represent an important part of the to-do list of future researches, to be addressed in the GLT setting.

Furthermore, the new spectral tools introduced in this paper might ease the analysis of some related problems. For example, the spectral analysis for more general variable coefficient 2D and 3D problems (dropping the hypothesis of elongated domain) appears achievable with the GLT theory, except for the case of variable degrees which is a real challenge. Also, more sophisticated solvers related to the Toeplitz technology, including multigrid type procedures and preconditioners can be studied for the solution of the arising saddle point problems. More in technical terms, few further directions of future research can be listed as follows:

- It is worthwhile observing that in a true multilevel setting (2D, 3D etc) there will be a potential deterioration of the quality of the clustering for the preconditioned matrix sequences, according to the theory developed by the third author, Tyrtyshnikov, Noutsos, Vassalos (see [29,32,34] and references therein). For overcoming the described difficulty, a successful approach is to combine preconditioned Krylov solvers and multigrid techniques, in order to neutralize the effect of the growing number of outliers when the dimensionality, i.e., the number of level increases; see e.g. [10–12].
- The presence of more general and even oscillating profiles $d(x)$ does not pose a real challenge, thanks to great flexibility of the GLT theory and this issue will be one of the first to be considered in future researches.
- As already remarked in the Introduction, the use of standard block circulant preconditioners (of Strang type, Frobenius optimal etc; see e.g. [9,23,28] and references therein) is not completely natural in the present setting, because structures like the Schur complement lose the Toeplitz character and therefore more sophisticate GLT ideas based on the spectral symbol become useful, in order to design efficient preconditioners.

All these open problems will be the subject of future investigations.

**Acknowledgements** All the authors are members of the INdAM research group GNCS. The work of the first author was partly supported by the GNCS-INdAM Young Researcher Project 2020 titled "Numerical methods for image restoration and cultural heritage deterioration".

## References


1. Arnold, D.N., Brezzi, F., Cockburn, B.: Unified analysis of discontinuous Galerkin methods for elliptic problems. SIAM J. Numer. Anal. **39**(5), 1749–1799 (2002)
2. Balay, S., Abhyankar, S., Adams, M.F., Brown, J., Brune, P., Buschelman, K., Dalcin, L., Eijkhout, V., Gropp, W.D., Karpeyev, D., Kaushik, D., Knepley, M.G., May, D.A., McInnes, L.C., Mills, R.T., Munson, T., Rupp, K., Sanan, P., Smith, B.F., Zampini, S., Zhang, H., Zhang, H.: PETSc users manual. Technical Report ANL-95/11-Revision 3.11, Argonne National Laboratory (2019)
3. Balay, S., Gropp, W.D., McInnes, L.C., Smith, B.F.: Efficient management of parallelism in object oriented numerical software libraries. In: Arge, E., Bruaset, A.M., Langtangen, H.P. (eds.) Modern Software Tools in Scientific Computing, pp. 163–202. Birkhäuser Press, New York (1997)
4. Barbarino, G., Garoni, C., Mazza, M., Serra-Capizzano, S.: Connecting GLT sequences with symbols of different matrix sizes (in preparation) (2021)
5. Barbarino, G., Garoni, C., Serra-Capizzano, S.: Block generalized locally toeplitz sequences: theory and applications in the multidimensional case. Electron. Trans. Numer. Anal. **53**, 113–216 (2020)
6. Barbarino, G., Garoni, C., Serra-Capizzano, S.: Block generalized locally toeplitz sequences: theory and applications in the unidimensional case. Electron. Trans. Numer. Anal. **53**, 28–112 (2020)
7. Barbarino, G., Serra-Capizzano, S.: Non-hermitian perturbations of Hermitian matrix-sequences and applications to the spectral analysis of the numerical approximation of partial differential equations. Numer. Linear Algebra Appl. **27**(3), 2286 (2020)
8. Böttcher, A., Silbermann, B.: Analysis of Toeplitz Operators. Springer, Berlin (2013)
9. Chan, R., Ng, M.: Conjugate gradient methods for Toeplitz systems. SIAM Rev. **38**(3), 427–482 (1996)
10. Donatelli, M., Garoni, C., Manni, C., Serra-Capizzano, S., Speleers, H.: Robust and optimal multi-iterative techniques for IgA Galerkin linear systems. Comput. Methods Appl. Mech. Eng. **284**, 1120–1146 (2015)
11. Donatelli, M., Garoni, C., Manni, C., Serra-Capizzano, S., Speleers, H.: Symbol-based multigrid methods for Galerkin B-spline isogeometric analysis. SIAM J. Numer. Anal. **55**, 31–62 (2017)
12. Dorostkar, A., Neytcheva, M., Serra-Capizzano, S.: Spectral analysis of coupled PDEs and of their Schur complements via Generalized Locally Toeplitz sequences in 2d. Comput. Methods Appl. Mech. Eng. **309**, 74–105 (2016)
13. Dumbser, M., Fambri, F., Furci, I., Mazza, M., Serra-Capizzano, S., Tavelli, M.: Staggered discontinuous Galerkin methods for the incompressible Navier–Stokes equations: spectral analysis and computational results. Numer. Linear Algebra Appl. **25**(5), 2151 (2018)
14. Elman, H., Howle, V., Shadid, J., Shuttleworth, R., Tuminaro, R.: Block preconditioners based on approximate commutators. SIAM J. Sci. Comput. **27**(5), 1651–1668 (2006)
15. Fambri, F., Dumbser, M.: Semi-implicit Discontinuous Galerkin methods for the incompressible Navier–Stokes equations on adaptive staggered Cartesian grids. Comput. Methods Appl. Mech. Eng. **324**, 170–203 (2017)
16. Fiorentino, G., Serra-Capizzano, S.: Multigrid methods for symmetric positive definite block Toeplitz matrices with nonnegative generating functions. SIAM J. Sci. Comput. **17**(5), 1068–1081 (1996)
17. Frigo, M., Johnson, S.G.: The design and implementation of FFTW3. Proc. IEEE **93**(2), 216–231 (2005)
18. Garoni, C., Serra-Capizzano, S.: Generalized locally Toeplitz sequences: theory and applications, vol. I. Springer, Cham (2017)
19. Garoni, C., Mazza, M., Serra-Capizzano, S.: Block generalized locally Toeplitz sequences: from the theory to the applications. Axioms **7**(3), 49 (2018)
20. Garoni, C., Serra-Capizzano, S.: Generalized locally Toeplitz sequences: a spectral analysis tool for discretized differential equations. In: Splines and PDEs: From Approximation Theory to Numerical Linear Algebra, pp. 161–236. Springer (2018)
21. Garoni, C., Serra-Capizzano, S., Sesana, D.: Spectral analysis and spectral symbol of $d$-variate $\mathbb{Q}_l$ lagrangian FEM stiffness matrices. SIAM J. Matrix Anal. Appl. **36**(3), 1100–1128 (2015)
22. Guzzetti, S., Perotto, S., Veneziani, A.: Hierarchical model reduction for incompressible fluids in pipes. Int. J. Numer. Methods Eng. **114**(5), 469–500 (2018)
23. Jin, X.: Developments and Applications of Block Toeplitz Iterative Solvers, vol. 2. Springer, Berlin (2003)
24. Kanschat, G.: Discotinuous Galerkin Methods for Viscous Incompressible Flow. Deutscher Universität Verlag (2007)

25. Liu, C.H., Lin, K.H., Mai, H.C., Lin, C.A.: Thermal boundary conditions for thermal lattice Boltzmann simulations. Comput. Math. Appl. **59**(7), 2178–2193 (2010)
26. Mansilla Alvarez, L., Blanco, P., Bulant, C., Dari, E., Veneziani, A., Feijóo, R.: Transversally enriched pipe element method (TEPEM): an effective numerical approach for blood flow modeling. Int. J. Numer. Methods Biomed. Eng. **33**(4), 2808 (2017)
27. Mazza, M., Ratnani, A., Serra-Capizzano, S.: Spectral analysis and spectral symbol for the 2d curl-curl (stabilized) operator with applications to the related iterative solutions. Math. Comput. **88**, 1155–1188 (2018)
28. Ng, M.: Iterative Methods for Toeplitz Systems. Oxford University Press, New York (2004)
29. Noutsos, D., Serra-Capizzano, S., Vassalos, P.: Matrix algebra preconditioners for multilevel Toeplitz systems do not insure optimal convergence rate. Theor. Comput. Sci. **315**(2–3), 557–579 (2004)
30. Rivière, B., Wheeler, M., Girault, V.: Improved energy estimates for interior penalty, constrained and discontinuous Galerkin methods for elliptic problems. Part I. Comput. Geosci. **3**, 337–360 (1999)
31. Serra-Capizzano, S.: Convergence analysis of two-grid methods for elliptic Toeplitz and PDEs matrix-sequences. Numer. Math. **92**(3), 433–465 (2002)
32. Serra-Capizzano, S.: Matrix algebra preconditioners for multilevel Toeplitz matrices are not superlinear. Linear Algebra Appl. **343**, 303–319 (2002)
33. Serra-Capizzano, S., Tablino-Possio, C.: Multigrid methods for multilevel circulant matrices. SIAM J. Sci. Comput. **26**(1), 55–85 (2004)
34. Serra-Capizzano, S., Tyrtyshnikov, E.: Any circulant-like preconditioner for multilevel matrices is not superlinear. SIAM J. Matrix Anal. Appl. **21**(2), 431–439 (2000)
35. Silvester, D., Elman, H., Kay, D., Wathen, A.: Efficient preconditioning of the linearized Navier–Stokes equations for incompressible flow. J. Computat. Appl. Math. **128**(1–2), 261–279 (2001)
36. Tavelli, M., Dumbser, M.: A staggered semi-implicit Discontinuous Galerkin method for the two dimensional incompressible Navier–Stokes equations. Appl. Math. Comput. **248**, 70–92 (2014)
37. Tavelli, M., Dumbser, M.: A staggered space-time Discontinuous Galerkin method for the incompressible Navier–Stokes equations on two-dimensional triangular meshes. Comput. Fluids **119**, 235–249 (2015)
38. Tilli, P.: A note on the spectral distribution of Toeplitz matrices. Linear Multilin. Algebra **45**(2–3), 147–159 (1998)
39. Tyrtyshnikov, E.: A unifying approach to some old and new theorems on distribution and clustering. Linear Algebra Appl. **232**, 1–43 (1996)
40. Tyrtyshnikov, E., Zamarashkin, N.: Spectra of multilevel Toeplitz matrices: advanced theory via simple matrix relationships. Linear Algebra Appl. **270**(1–3), 15–27 (1998)
41. Van Loan, C.: Computational Frameworks for the Fast Fourier Transform. SIAM, Philadelphia (1992)
42. Wheeler, M.: An elliptic collocation-finite element method with interior penalties. SIAM J. Numer. Anal. **15**(1), 152–161 (1978)