Università degli Studi dell'Insubria
Dipartimento di Scienza e Alta Tecnologia

Dottorato di Ricerca in Informatica e Matematica del Calcolo

Toeplitz and Block-Toeplitz Structures with Variants: from the Spectral Analysis to Preconditioning and Multigrid Methods Using a Symbol Approach

Ph.D. thesis of

Paola Ferrari

Advisors: Prof. Marco Donatelli
Prof. Stefano Serra–Capizzano

XXXIII Cycle

Academic year 2019/2020

A mathematical problem should be difficult in order to entice us,
yet not completely inaccessible, lest it mock at our efforts.

David Hilbert

List of Papers

This thesis is based on the following papers.

- P. Ferrari, I. Furci, S. Hon, M. A. Mursaleen, S. Serra-Capizzano, *The eigenvalue distribution of special 2-by-2 block matrix sequences, with applications to the case of symmetrized Toeplitz structures.* **SIAM J. Matrix Anal. Appl**. 40(3), 1066–1086, 2019.

  All authors contributed equally to this work. In particular, the author of this thesis contributed to the theoretical part and had the responsibility of the numerical section.

- P. Ferrari, R.I. Rahla, C. Tablino Possio, S. Belhaj, S. and Serra–Capizzano. *Multigrid for $\mathbb{Q}_k$ Finite Element Matrices using a (block) Toeplitz symbol approach.* **Mathematics**. 8(5), 2020.

  The author of this thesis has given significant contributions to the paper both from theoretical and algorithmic points of view and, hence, the alphabetical order was not respected in listing the authors.

- P. Ferrari, N. Barakitis, S. Serra-Capizzano, *Spectral distribution of analytic functions of Toeplitz matrix-sequences.* **Numer. Linear Algebra Appl.** e2332, 2020.

  The author of this thesis exploited the preliminary idea of the co-authors to develop the proof of the main theorem, implemented the most significant numerical examples, and collected all the results in article form. Hence, the alphabetical order was not respected in listing the authors.

- M. Donatelli, P. Ferrari, I. Furci, D. Sesana, S. Serra-Capizzano, *Multigrid methods for block-circulant and block-Toeplitz large linear systems: algorithmic proposals and two-grid optimality analysis.* **Numer. Linear Algebra Appl.** (Accepted).

  The authors contributed equally to this work. In particular, the author of this thesis gave significant contributions for proving the main result and proposed novel ideas for the numerical section.

- P. Benedusi, P. Ferrari, C. Garoni, R. Krause, S. Serra-Capizzano, *Fast Solver for the Space-Time IgA-DG Discretization of the Anisotropic Diffusion Equation*, Technical Report 2019-011, Department of Information Technology, Uppsala University, 2019.

  The authors contributed equally to this work. In particular, the author of this thesis gave a substantial contribution in implementing the proposed algorithms.

- M. Bolten, M. Donatelli, P. Ferrari, and I. Furci, *A symbol based analysis for multigrid methods for block-circulant and block-toeplitz systems.* (Under preparation).

  The authors contributed equally to this work.

Contents

Introduction and Motivation

The principles behind the definition of circulant and Toeplitz structures constitute a significant example on how a fascinating and elegant mathematical theory can be useful for the solution of real problems resulting in science and engineering. According to this, the aim of the current thesis is of dual nature: it seeks to expand the well-known theoretical knowledge on such matrix structures and to apply the results to real-life problems.

The importance of circulant and Toeplitz operators is indeed a consequence of the connection with a variety of problems in Physics, Probability Theory, Cryptology, Engineering and Applied Sciences. In general, the intrinsic nature of these problems is continuous. In fact, a great deal of applications requires the solution of a Partial Differential Equation (PDE). However, this kind of continuous equation often does not admit an analytical solution, which needs to be approximated by means of a numerical method.

Several numerical methods that perform these kinds of approximations consist in creating a sequence of discrete problems and computing the corresponding numerical solutions, which converge to a quantity that permits to reconstruct the solution of the original PDE. If the PDE and the numerical method are both linear, the computation of the numerical solution reduces to solving a sequence of linear systems with increasing dimensions [5, 33, 64].

It often happens that the sequence of the coefficient matrices of these systems is a sequence of structured matrices with a certain type of either time or space invariance. This is the reason why the studies on circulant and Toeplitz matrices and of all the structures constructed with them have maintained high popularity over the years [23, 35, 114, 133, 140].

From a theoretical point of view, dealing with circulant matrices does not require significant computational efforts, since they constitute a class for which most matrix-theoretic questions may be resolved in closed form [35, 64, 100]. Indeed, the circulant matrices form an algebra and in particular they are simultaneously diagonalized by the discrete Fourier matrix. The latter theoretical aspect has two consequences in Numerical Analysis. On one hand, linear equations with circulant coefficient matrices do not present computational difficulties, since they may be quickly – with respect to the matrix-size – solved using only few fast Fourier transforms [32, 138]. On the other hand, this computational advantage causes them to be frequently chosen in several contexts, for instance when it is needed to approximate the inverse of a Toeplitz matrix, and in addition their built-in periodicity makes them suitable for many applications [29–31, 52, 123]. Moreover, circulant matrices are a subclass of Toeplitz operators and hence they are part of an additional fascinating context. Indeed, the study of the properties of Toeplitz matrices by means of their generating functions represents an exceptional example of interplay between matrix theory and function theory. In fact, there exists a correspondence

between the analytic properties of the generating functions and the algebraic properties of the associated Toeplitz matrices [113, 115, 140]. However what is even more significant is that, under certain hypotheses, the generating function can be an elegant and efficient tool that provides an asymptotic approximation of the singular values and eigenvalues of Toeplitz matrices [36, 41, 42, 115, 140], which is effective also for moderate dimensions. Typically, in the unilevel setting, the approximation error is proportional to the inverse of the matrix-size.

In order to better explain the meaning of the aforementioned approximation features, we have to deal with the concept of asymptotic distributions. The informal meaning of eigenvalue and singular value distribution of a given matrix-sequence $\{A_n\}_n$, which is rigorously presented in **Chapter I**, is that, for n sufficiently large, a uniform sampling of a given function f – which is called the spectral symbol – provides an approximation of the eigenvalues of A_n and, analogously, a uniform sampling of $|f|$ – which in this case is called the singular value symbol – over its domain gives an approximation of the singular values of A_n.

As we already suggested, for Toeplitz matrix-sequences the candidate asymptotic symbol is the generating function, however this conjecture is verified only under specific hypotheses. The broad studies that have been carried out in the past few decades provide us with a clear outline on the topic. Indeed, Szegő in [67] showed that the eigenvalues of the Toeplitz matrix $T_n[f]$ generated by real-valued $f \in L^\infty([-\pi,\pi])$ are asymptotically distributed as f. Moreover, Avram and Parter [6, 103] proved that the singular values of $T_n[f]$ are distributed as $|f|$ for a complex-valued $f \in L^\infty([-\pi,\pi])$. Tyrtyshnikov [134, 135, 140] later extended the spectral and singular value theorems to Toeplitz matrices $T_n[f]$ generated by functions $f \in L^1([-\pi,\pi])$.

In this well-structured framework, there is one feature missing: if a Toeplitz matrix is not Hermitian, in general we cannot discover its spectral properties by studying the generating function. Since the knowledge of the spectral information is crucial in the design and in the convergence analysis of fast solution methods for Toeplitz systems, as we will explain in more detail later, it might be convenient to develop a strategy that permits us to transform non-Hermitian linear systems into a form for which a spectral analysis is easier.

Indeed, under the hypothesis that the Toeplitz matrix $T_n[f]$ possesses real entries, a smart symmetrization procedure can be applied. Namely, as suggested in [104], we can premultiply $T_n[f]$ by the anti-identity matrix $Y_n \in \mathbb{R}^{n\times n}$ in order to study the symmetrized matrix $Y_nT_n[f]$.

One of the main contributions of this thesis is to give a spectral distribution result for sequences of the form $\{Y_nT_n[f]\}_n$ [53]. In particular, in **Chapter II** we show that the generating function f of $T_n[f]$ plays a fundamental role: we prove a result which informally means that roughly half of the eigenvalues of $Y_nT_n[f]$ are positive and they are approximated by a uniform sampling of $|f|$ and roughly half of the eigenvalues are negative and they are approximated by a uniform sampling of $-|f|$. Moreover, the proof is based on a new tool, which analyses the eigenvalue distribution of special 2-by-2 block matrix-sequences and has a general character, and, therefore, can be potentially used in different contexts.

A second goal of this thesis is to extend the latter setting, providing asymptotic distribution results for the analogous symmetrization of the sequence $\{h(T_n[f])\}_n$, where h is an analytic function [52]. In particular, we consider a function f in $L^\infty([-\pi,\pi])$ with real Fourier coefficients and an analytic function h with convergence radius r such that $\|f\|_\infty < r$. Under these hypotheses, we prove that the matrix-sequence $\{h(T_n[f])\}_n$ is distributed in the singular value

sense as $h \circ f$, which is a result with intrinsic significance. We exploit this property further to investigate the spectral distribution of the symmetrized sequence $\{Y_n h(T_n[f])\}_n$ and show that its spectral symbol is given by

$$\phi_{|h\circ f|}(\vartheta) = \begin{cases} |h \circ f(\vartheta)|, & \vartheta \in [0, 2\pi], \\ -|h \circ f(-\vartheta)|, & \vartheta \in [-2\pi, 0), \end{cases}$$

which informally means that, for a sufficiently large n, roughly half of the eigenvalues of $Y_n h(T_n[f])$ are approximated by a uniform sampling of $|h \circ f|$ and roughly half of the eigenvalues are approximated by a uniform sampling of $-|h \circ f|$. The proof of this result is based on the properties of the Generalized Locally Toeplitz (GLT) sequences, which form a particular class of matrix-sequences to which Toeplitz matrix-sequences with Lebesgue integrable generating functions belong [12, 13, 62, 63].

As we already mentioned, the study on Toeplitz-related sequences is crucial in many applications, when it is required to solve particular linear systems with structured coefficient matrices [16, 58]. In some cases, direct solution methods represent the best choice for their robustness and predictable behaviour. However, it often happens that the size of the linear systems increases as we seek more accuracy in the approximation of the solution of the problem. Thus, the computational complexity of the algorithm is a fundamental aspect in the development of feasible solution methods. If the bandwidth of the matrix is sufficiently small, Gaussian elimination is a reasonable choice [73]. However, if the Toeplitz matrices are not sparse or possess a multilevel structure, which often happens if the initial problem is a multidimensional PDE, many direct solvers – such as the standard Gaussian elimination – do not exploit the structure of the matrices and the computational cost could be not affordable even with high-performance computers. To overcome this, in the past decades many solution algorithms of iterative nature have been employed for the solution of linear systems with Toeplitz coefficient matrices [66, 70, 99, 108].

Some of the most successful iterative procedures that have been developed involve two key ingredients: Krylov subspace methods and preconditioning. Krylov subspace methods are a class of iterative solvers for a system of linear equations. Among them, it is worth citing the Conjugate Gradient (CG) method, developed by Hestenes and Stiefel [71] in 1952, the Minimal Residual (MINRES) method, designed by Paige and Saunders [102] in 1975, and the Generalized Minimal Residual (GMRES) method, conceived by Saad and Schultz [109] in 1986. On the other hand, preconditioning involves the alteration of the original linear system in order to accelerate the computation of the approximated solution.

From a theoretical point of view, in order to develop an efficient preconditioner for a linear system $Ax = b$, one strategy is to look for a matrix P such that the chosen Krylov subspace method converges faster for a linear system with coefficient matrix $P^{-1}A$. In practice, in this first case, in order to find the solution of the original system it is necessary to solve a linear system with coefficient matrix P. A second approach is to construct an approximate inverse $\tilde{P}$ of A as a preconditioner, which requires to perform matrix-vector multiplications with matrix $\tilde{P}$. Advanced preconditioning strategies do not involve the construction of a matrix, they consist in the development of a procedure whose on a vector has the same role as the matrix-vector multiplication with matrix $\tilde{P}$. Since the goal of preconditioning is to speed up the convergence of the chosen method, it is evident that the operations of solving a linear system with mat-

rix P, multiplying a vector by $\tilde{P}$ or applying the preconditioning procedure should not be as computationally expensive as the solution of the initial system.

As far as Toeplitz linear systems are concerned, a suitable preconditioner P might be sought in the class of circulant matrices. Indeed, many results on how circulant matrices approximate well Toeplitz matrices have been obtained [29–31]. Moreover, as we already pointed out, a linear system with a coefficient matrix in the circulant algebra can be quickly solved making use of a fast Fourier transform algorithm.

A significant feature of some Krylov methods such as the Conjugate Gradient and the Minimal Residual methods is that the convergence rate of the algorithm can be estimated using only the eigenvalues of the system matrix. Therefore, it is evident that the knowledge of the spectral distributions of the coefficient matrix-sequences for linear systems of increasing dimension is of critical importance in the design of a good preconditioner.

Combining the literature on circulant preconditioning and the aforementioned spectral results on symmetrized Toeplitz sequences, in this thesis we prove the effectiveness of preconditioning strategies for the matrix-sequences $\{Y_n T_n[f]\}_n$ and $\{Y_n h(T_n[f])\}_n$. Indeed, the final goal of our findings is to exploit the derived spectral clustering information on the preconditioned matrix-sequences in order to estimate the convergence rate of MINRES for the related preconditioned linear systems [52, 53].

Following the preconditioning strategy suggested by Pestana and Wathen in [104], given a circulant matrix C_n such that $\{C_n^{-1} T_n[f]\}_n$ is distributed in the singular value sense as the function 1, we propose as preconditioner for the symmetrized matrix $Y_n(T_n[f])$ the absolute value circulant matrix $|C_n|$. The latter is defined by

$$|C_n| \quad = F_n |\Lambda_n| F_n^H,$$

where F_n is the $n \times n$ Fourier matrix, and $|\Lambda_n|$ is the diagonal matrix in the eigendecomposition of C_n with all entries replaced by their magnitude.

Finally, we prove that the derived preconditioned matrix-sequence is distributed in the eigenvalue sense as

$$\phi_1(\vartheta) = \begin{cases} 1, & \vartheta \in [0, 2\pi], \\ -1, & \vartheta \in [-2\pi, 0), \end{cases}$$

under the mild assumption that f is sparsely vanishing. The latter implies that roughly half of the eigenvalues are clustered at 1 and roughly half of the eigenvalues are clustered at -1, which is a desirable property for the fast convergence of the preconditioned Minimal Residual method.

Along with the low – with respect to the matrix size – computational cost of performing one iteration, a crucial property of a preconditioned iterative method is its optimality, that is, the algorithm should have a convergence rate independent of the matrix size. While circulant preconditioning for Toeplitz linear systems often leads to optimal Krylov subspace iterations [30, 40, 114, 116], in the multilevel and multilevel block Toeplitz settings the performances of (multilevel block) circulant preconditioners deteriorate (see [101, 120, 124] and references therein). This is a reason why also the class of multigrid methods is of great interest in this context. In fact, multigrid methods achieve a fast convergence rate by constructing via consecutive projections a proper sequence of linear systems of decreasing dimensions.

In the case of circulant and Toeplitz matrix-sequences generated by a scalar-valued function, the convergence and optimality analysis of multigrid methods has been obtained in a compact and elegant form. This has been done firstly in the unilevel case in [27, 56, 82] and then in the multilevel case [57, 119, 127]. The cited works provide the convergence analysis of the two-grid method with proper choices of grid transfer operators, while the V-cycle analysis is present in more recent works [3, 4]. Following this approach, the importance of asymptotic distributions becomes evident once again: the grid transfer operators are defined exploiting the analytical properties of the symbols associated to the matrix-sequences for which the multigrid method is designed.

In the block-circulant and block-Toeplitz setting, that is, in the case where the matrix entries are small generic matrices instead of scalars, some algorithms have already been proposed regarding specific applications, but the attention to theoretical results is still marginal. Namely, when the generating function is matrix-valued and non-trivial, there is still a substantial lack of an effective projection proposal and of a rigorous convergence analysis. A further aim of this thesis is to fill this theoretical gap.

According to the classical Ruge and Stüben convergence analysis in [107], the two-grid convergence can be proven validating both a smoothing property and an approximation property. The first is easily generalizable in the block setting and we show how it mainly regards the choice of the specific relaxation parameter for the selected smoother [39]. Conversely, mimicking the proof for the approximation condition from the scalar structures is non-trivial, owing to the non-commutativity of the involved matrix-valued symbols.

In our analysis, we mainly focus on the crucial choice of conditions on the trigonometric polynomial used to construct the projector in order to ensure the optimal convergence rate of the two-grid method [39], since the generalization of the conditions present in the scalar setting is not sufficient for this purpose. Firstly, we assume the trigonometric polynomial that generates the block-circulant matrix used in the construction of the grid transfer operator to be unitarily diagonalizable at all points and to satisfy a specific commutativity condition. This approach provides us with the tools to define a class of grid transfer operators suitable for the achievement of the two-grid convergence. Then, we prove the approximation property in a more general case, observing that many multigrid methods, known in the literature, usually do not fit in the previous setting, having, for instance, a non-diagonalizable matrix-valued symbol. In both cases, we prove that the two-grid convergence rate is optimal, independent from the matrix size, in the case of positive definite block matrices with generic blocks [20, 39].

Furthermore, taking inspiration from the approach in [21], we propose a measure of the ill-conditioning of the symbol at the coarser levels in order to choose a robust grid transfer operator that yields to fast multigrid convergence for more than two grids.

To show the numerical validity of our theoretical results, we consider the case of large positive definite block linear systems stemming from quadrilateral Lagrangian Finite Element Methods (FEM) – denoted in the sequel as $\mathbb{Q}_s$ – applied to the Poisson problem [64]. An important step for the numerical approximation involves the solution of linear systems which possess a natural block (and multilevel block) Toeplitz structure, up to a low rank correction.

Firstly, we propose a classical multigrid strategy that follows a functional approach, that is, we define the prolongation operator as the inclusion operator between the coarser and finer

functional spaces [55]. We analyse the prolongation matrix as a cut block-Toeplitz matrix and we prove that its symbol satisfies the hypotheses for the optimality of the two-grid and V-cycle convergence rate. We perform an analogous analysis also to a second multigrid strategy, where we choose a linear interpolation prolongation operator.

The last projecting strategy that we present for the $\mathbb{Q}_s$ stiffness matrices has a more general interest, namely, it can be applied to every positive definite Toeplitz matrix-sequence with generating function $\mathbf{f}$ that is singular at exactly one point ϑ_0 in its domain. Indeed, for the $\mathbb{Q}_s$ stiffness matrices, the construction of the grid transfer operators depends on a suitable trigonometric polynomial, which is chosen based only on an algebraic analysis of the symbol $\mathbf{f}$ associated with the linear systems matrix-sequence. In particular, the class of grid transfer operators is generated by the following matrix-valued trigonometric polynomial $\mathbf{p}_z$:

$$\mathbf{p}_z(\vartheta) = F_s \begin{bmatrix} z(1+\cos\vartheta) & & & \\ & 1+\cos\vartheta & & \\ & & \ddots & \\ & & & 1+\cos\vartheta \end{bmatrix} F_s^H,$$

where s is the block-size. As we will explain, the choice of a function such as $1+\cos\vartheta$ on the diagonal is connected to the behaviour of the eigenvalues of $\mathbf{f}(\vartheta)$ varying ϑ. Moreover, the circulant structure of $\mathbf{p}_z(\vartheta)$ is a consequence of the study of the eigenvector associated to the null eigenvalue of $\mathbf{f}(\vartheta_0)$. In order to have an optimal two-grid convergence rate, the choice $z=1$ is feasible and has a general character. On the other hand, for the extension of the optimal convergence rate to more than two grids, we numerically study a better choice of the parameter z such that the aforementioned conditioning of the symbol at the coarser levels does not worsen.

The last chapter of this thesis is dedicated to the design of a fast solution method for systems of linear equations with a more complicated structure. To this end, we consider the space-time discretization of the linear anisotropic diffusion equation [15], using an isogeometric analysis (IgA) approximation in space and a discontinuous Galerkin (DG) approximation in time [16]. The solution method for the resulting space-time linear system includes a newly proposed preconditioner for the Preconditioned GMRES (PGMRES) algorithm, which involves a few iterations of an appropriate multigrid method. Both the preconditioning and the multigrid strategy are designed following the leading concepts on the development of iterative solvers for structured matrices that constitute the basis of the whole thesis, that is, exploiting the spectral information on the derived matrices, which is known from the eigenvalue distribution results provided in [16]. Moreover, we pay particular attention to the development of an algorithm that can be parallelized and performs well on parallel computers, since this is an aspect that has been of great interest in recent years, due to the physical constraints for the clock frequency of processors. The numerical experiments confirm that our preconditioned solution method possesses good parallel scaling properties and is competitive in terms of robustness and run-time.

We conclude this general introductory part briefly describing the contents of the thesis chapter by chapter.

In **Chapter I** we set the notation used throughout the thesis and we provide the fundamental definitions and results that are preliminary to the subsequent chapters. In particular, we give the

definition of circulant and Toeplitz matrices – and their generalizations to the block and block multilevel cases – and we present their main structural and spectral features. Then, we formally introduce the concepts of asymptotic distributions and of approximating classes of sequences and provide the minimal notions for understanding the basics of the GLT theory, which is an essential tool in **Chapter III**. We also give an overview of iterative methods and report in more detail results for the preconditioned MINRES method and for multigrid methods.

In **Chapter II**, we consider the sequence of matrices $\{Y_nT_n[f]\}_n$, where $T_n[f]$ is the n-by-n Toeplitz matrix generated by a function f in $L^1([-\pi,\pi])$ and Y_n is the anti-identity matrix. Because of the unitary nature of Y_n, the singular values of $T_n[f]$ and $Y_nT_n[f]$ coincide. However, the eigenvalues are affected substantially by the action of Y_n. Under the assumption that the Fourier coefficients of f are real, we prove that $\{Y_nT_n[f]\}_n$ is distributed in the eigenvalue sense as

$$\phi_g(\vartheta) = \begin{cases} g(\vartheta), & \vartheta \in [0, 2\pi], \\ -g(-\vartheta), & \vartheta \in [-2\pi, 0), \end{cases}$$

with $g(\vartheta) = |f(\vartheta)|$. A generalization of this result to the block Toeplitz case is also shown. Next, we consider the circulant preconditioning introduced by J. Pestana and A. Wathen [104] and prove that the preconditioned matrix-sequence is distributed in the eigenvalue sense as ϕ_1 under mild assumptions on f. A number of numerical experiments is provided and critically discussed.

In **Chapter III**, we extend the results proven in **Chapter II** to matrix-sequences of the form $\{h(T_n[f])\}_n$, where h is an analytic function. In particular, we provide the singular value distribution of the sequence $\{h(T_n[f])\}_n$, the eigenvalue distribution of the sequence $\{Y_nh(T_n[f])\}_n$, and the conditions on f and h for these distributions to hold. The final goal of the chapter is to exploit our theoretical findings for the fast solution of linear systems stemming from some applications of interest. In particular, we provide efficient circulant preconditioning strategies for the matrix-sequence $\{Y_nh(T_n[f])\}_n$ in several settings. Starting from the case where the function h is simply a polynomial, we finally study the case of the exponential of a real nonsymmetric Toeplitz matrix stemming from computational finance, in particular, from the option pricing framework in jump-diffusion models, where a partial integro-differential equation (PIDE) needs to be solved.

In **Chapter IV**, we propose a general two-grid convergence analysis, proving an optimal convergence rate independent of the matrix size, in the case of positive definite block-circulant matrices with generic blocks. The proof of the approximation property is not a straightforward generalization of the scalar case, we have to require additional conditions on the block symbol of the grid-transfer operator. In particular, we analyse a first case when the trigonometric polynomial that generates the block-circulant matrix used in the construction of the grid transfer operator is unitarily diagonalizable at all points and satisfies a specific commutativity condition. However, most of the known multigrid methods do not fit in this particular setting, which suggests that the hypotheses for the optimal two-grid convergence rate can be relaxed, namely, we prove the approximation property for a grid transfer operator with a block symbol that might be non-diagonalizable. In this case, it becomes clear that not only the eigenvalue functions of the symbol are crucial for the convergence of the method, but also the eigenvectors should be carefully analysed. Then, we provide a generalization of the convergence results to multilevel

block-circulant matrices, where the multilevel grid transfer operator possesses a tensor structure.

In **Chapter V**, we exploit the theoretical findings of **Chapter IV** to develop and analyse multigrid strategies for the solution of linear systems stemming from the $\mathbb{Q}_s$ Finite Elements approximation of elliptic partial differential equations with Dirichlet boundary conditions and where the operator is $\operatorname{div}(-a(\mathbf{x})\nabla\cdot)$, with a continuous and positive over $[0,1]^k$. Firstly, we propose a classical multigrid strategy that follows a functional approach, that is, we define the prolongation operator as the inclusion operator between the coarser and finer functional spaces. We analyse the prolongation matrix as a cut block-Toeplitz matrix and we prove that its symbol satisfies the hypotheses for the two-grid and V-cycle convergence and optimality. We perform an analogous analysis also for a second multigrid strategy, where we choose a linear interpolation prolongation operator. Finally, we present a third class of grid transfer operators, which has a different genesis. According to the analysis, we show how to exploit the properties of the eigenvalue functions to define a class of grid transfer operators that satisfy the theoretical conditions of **Chapter IV**. In this way, we explain how to choose the trigonometric polynomial that generates the block-Toeplitz matrix used in the construction of the grid transfer operator focusing only on algebraic considerations on the symbol of the linear system matrix-sequence. Even though we focus on the $\mathbb{Q}_s$ stiffness matrices, the presented procedure has a wider interest, since it might be applied to every matrix-sequence that falls into the theoretical setting. Results of numerical experiments that test all the considered methods are presented, both in one dimension and in higher dimension, showing an optimal behaviour in terms of the dependency on the matrix size and a robustness with respect to the dimensionality.

In **Chapter VI** we consider the space-time discretization of the (linear) anisotropic diffusion equation, using an isogeometric analysis (IgA) approximation in space and a discontinuous Galerkin (DG) approximation in time. Drawing inspiration from a former spectral analysis, we propose for the resulting space-time linear system a new preconditioner for the PGMRES algorithm, which involves a few iterations of an appropriate multigrid method. The performance of our preconditioned solution method is illustrated through numerical experiments, which show its competitiveness in terms of robustness, run-time and parallel scaling.

A conclusion chapter ends the present thesis including a list of open questions, perspectives, and future issues to be addressed in further researches.

Chapter I

Preliminary Definitions and Results

The present chapter introduces the notation used throughout the thesis and provides the fundamental definitions and results that are preliminary to the subsequent chapters.

First, we illustrate how to work with matrix-valued functions and their associated eigenvalue functions. Moreover, we formally introduce the concepts of approximating classes of sequences and of asymptotic distributions.

Then, we give the definition of circulant and Toeplitz matrices, we present their main structural and spectral features and their generalizations to the block and block multilevel cases. We also provide the minimal tools for understanding the basics of the Generalized Locally Toeplitz theory. Furthermore, we write a suitable definition of function of matrices and we study its meaning in the Toeplitz case.

In the final sections of the chapter we recall the fundamentals on iterative methods, paying particular attention to preconditioned Krylov solvers, especially the preconditioned MINRES method, and on multigrid methods, for which we follow an algebraic Ruge-Stüben approach.

I.1 General Notation

The following list describes the notation that is used throughout the thesis.

- $\mathbb{K}^{m\times n}$ is the space of $m \times n$ matrices with coefficients in $\mathbb{K} \in \{\mathbb{R}, \mathbb{C}\}$, $\mathbb{R}, \mathbb{C}$ being real and complex numbers, respectively.
- If $x = [x_j]_{j=1}^n$ is a vector, then
  - x^T denotes the transpose of x;
  - x^H denotes the conjugate transpose of x;
- If $A = [a_{ij}]_{i,j=1}^n \in \mathbb{C}^{n\times n}$, we denote by
  - A^T the transpose of A;
  - A^H the conjugate transpose of A;
  - rank(A) the rank of A;
  - tr(A) the trace of A;

  - $\det(A)$ the determinant of A;
  - $\lambda_j(A)$, $j = 1, \ldots, n$, the eigenvalues of A;
  - $\sigma_j(A)$, $j = 1, \ldots, n$ the singular values of A;
  - $\Lambda(A)$ the spectrum of A.

- If A, $B \in \mathbb{C}^{n\times n}$, we write
  - $A \geq B$ if A and B are Hermitian and $A - B$ is Hermitian Positive Semi-Definite (HPSD);
  - $A > B$ if A and B are Hermitian and $A - B$ is Hermitian Positive Definite (HPD);

- $O_{n,m}$ is the $n \times m$ zero matrix. When the dimension is clear from the context, the subscript is omitted.

- I_m is the $m \times m$ identity matrix. When the dimension is clear from the context, the subscript is omitted.

- Y_m is the $m \times m$ anti-identity matrix.

- F_m is the $m \times m$ Fourier matrix, that is

$$F_m = \frac{1}{\sqrt{m}} \left[\mathrm{e}^{-\hat{\imath}\frac{2\pi ij}{m}} \right]_{i,j=0}^{m-1}.$$

- $\mathbf{o}_n$ denotes the vector $[0, 0, \ldots, 0]^T \in \mathbb{R}^n$.

- $\mathbf{e}_n$ denotes the vector $[1, 1, \ldots, 1]^T \in \mathbb{R}^n$.

- μ_k denotes the Lebesgue measure in $\mathbb{R}^k$. If not specified otherwise, "measure" always refers to the Lebesgue measure.

- $\hat{\imath}$ is the imaginary unit, that is $\hat{\imath}^2 = -1$.

- If $A \in \mathbb{C}^{n_1\times n_2}$ and $B \in \mathbb{C}^{m_1\times m_2}$, the Kronecker product of A and B is the $n_1m_1 \times n_2m_2$ matrix defined by

$$A \otimes B = [a_{ij}B]_{i=1,\ldots,n_1\,,\,j=1,\ldots,n_2} = \begin{bmatrix} a_{11}B & \ldots & a_{1n_2}B \\ a_{21}B & \ldots & a_{2n_2}B \\ \vdots & \ddots & \vdots \\ a_{n_11}B & \ldots & a_{n_1n_2}B \end{bmatrix}.$$

- Let D be a measurable subset of $\mathbb{R}^k$ and let $f_m, f : D \to \mathbb{C}$ be measurable functions for all $m \in \mathbb{N}$. We say that the sequence f_m converges to f in measure and we write

$$f_m \to f \text{ in measure}$$

  if, for every $\varepsilon > 0$,

$$\lim_{m\to\infty} \mu_k\{|f_m - f| > \varepsilon\} = 0.$$

- Given D a measurable subset of $\mathbb{R}^k$, we denote by
  - $L^p(D)$ the space of measurable functions $f : D \to \mathbb{C}$ such that
$$\int_D |f|^p \, \mathrm{d}\mu_k < \infty, \quad 1 \leq p < \infty;$$
  - $L^\infty(D)$ the space of measurable functions $f : D \to \mathbb{C}$ such that
$$\operatorname{ess\,sup}_D |f| < \infty.$$
- If D is a measurable subset of $\mathbb{R}^k$, given $f \in L^p(D)$, the quantity $\|f\|_p$ is the L^p-norm of f, that is
$$\|f\|_p = \begin{cases} \left(\int_D |f|^p \, \mathrm{d}\mu_k\right)^{\frac{1}{p}}, & \text{if} \quad 1 \leq p < \infty, \\ \operatorname{ess\,sup}_D |f|, & \text{if} \quad p = \infty. \end{cases}$$
- Given two sequences $\{a_n\}_n$ and $\{b_n\}_n$, with $a_n \geq 0$ and $b_n > 0$ for all n, the notation $a_n = O(b_n)$ means that there exists a constant C, independent of n, such that $a_n \leq C b_n$ for all n and the notation $a_n = o(b_n)$ means that $a_n / b_n \to 0$ as $n \to \infty$.
- A vector $\mathbf{i} = (i_1, i_2, \ldots, i_k) \in \mathbb{Z}^k$ is called a k-index or simply a multi-index.
- For all $\mathbf{n} = (n_1, n_2, \ldots, n_k) \in \mathbb{Z}^k$ we define the multi-index length by $\mathcal{N}(\mathbf{n}) = n_1 n_2 \ldots n_k$.
- $\mathbf{0}$, $\mathbf{1}$, $\mathbf{2}$,... respectively indicate $(0, 0, \ldots, 0)$, $(1, 1, \ldots, 1)$, $(2, 2, \ldots, 2)$, ...
- For all $\mathbf{n}$, $\mathbf{m} \in \mathbb{Z}^k$, $\mathbf{n} \leq \mathbf{m}$ means $n_i \leq m_i$, $\forall\, i = 1, \ldots, k$.
- If $\mathbf{n}$, $\mathbf{m} \in \mathbb{Z}^k$ are such that $\mathbf{n} \leq \mathbf{m}$, the multi-index range $\mathbf{n}, \ldots, \mathbf{m}$ is the set
$$\{\mathbf{j} \in \mathbb{Z}^k \, : \, \mathbf{n} \leq \mathbf{j} \leq \mathbf{m}\}.$$
- Given $\mathbf{n}, \mathbf{m} \in \mathbb{Z}^k$, with $\mathbf{n} \leq \mathbf{m}$, the notations $\sum_{\mathbf{j}=\mathbf{n}}^{\mathbf{m}}$, $\prod_{\mathbf{j}=\mathbf{n}}^{\mathbf{m}}$ and $\bigotimes_{\mathbf{j}=\mathbf{n}}^{\mathbf{m}}$ respectively indicate the summation, product, and Kronecker product over all multi-indices $\mathbf{j} = \mathbf{n}, \ldots, \mathbf{m}$.
- If $\mathbf{m} \in \mathbb{N}^k$ then
$$x = [x_{\mathbf{i}}]_{\mathbf{i}=\mathbf{1}}^{\mathbf{m}}$$
is a vector of size $\mathcal{N}(\mathbf{m})$ whose components $x_{\mathbf{i}}$, $\mathbf{i} = \mathbf{1}, \ldots, \mathbf{m}$ are sorted in accordance with the lexicographic ordering. Similarly
$$A = [a_{\mathbf{ij}}]_{\mathbf{i},\mathbf{j}=\mathbf{1}}^{\mathbf{m}}$$
is the $\mathcal{N}(\mathbf{m}) \times \mathcal{N}(\mathbf{m})$ matrix whose components are indexed by two multi-indices, both varying in $\mathbf{1}, \ldots, \mathbf{m}$ according the lexicographic ordering.

I.2 Matrix Norms

In the following subsections we include the definitions and the properties of two fundamental classes of matrix norms, namely p-norms and Schatten p-norms.

I.2.1 p-norms

Given $1 \leq p \leq \infty$ and $x \in \mathbb{C}^n$, we denote by $\|x\|_p$ the p-norm of x, i.e.,

$$\|x\|_p = \begin{cases} \left(\sum_{i=1}^n |x_i|^p\right)^{1/p}, & \text{if } 1 \leq p < \infty, \\ \max_{i=1,\dots,n} |x_i|, & \text{if } p = \infty. \end{cases} \tag{I.1}$$

The p-norm of a matrix $A \in \mathbb{C}^{n\times n}$ is the matrix norm induced by the vector norm $\|\cdot\|_p$,

$$\|A\|_p = \max_{x\in\mathbb{C}^n,\, \|x\|_p=1} \|Ax\|_p.$$

The 2-norm is also known as the spectral norm. Being an operator norm on $\mathbb{C}^{n\times n}$, the matrix p-norm satisfies the inequality $\rho(A) \leq \|A\|_p$ and has the sub-multiplicative property, that is

$$\|AB\|_p \leq \|A\|_p \|B\|_p, \qquad \forall A, B \in \mathbb{C}^{n\times n}. \tag{I.2}$$

In for some special values of p, a formula for the computation of $\|A\|_p$ is available [65]:

$$\|A\|_1 = \max_{j=1,\dots,n} \sum_{i=1}^n |a_{ij}|, \tag{I.3}$$

$$\|A\|_2 = \sqrt{\rho(A^H A)} = \sqrt{\lambda_{\max}(A^H A)}, \tag{I.4}$$

$$\|A\|_\infty = \max_{i=1,\dots,n} \sum_{j=1}^n |a_{ij}|; \tag{I.5}$$

From Formula (I.4) it is straightforward to see that the 2-norm is unitarily invariant, that is,

$$\|A\|_2 = \|UAV\|_2 \tag{I.6}$$

for all $A \in \mathbb{C}^{n\times n}$ and all unitary matrices $U, V \in \mathbb{C}^{n\times n}$.

Moreover, we report the inequality

$$\|A\|_2 \leq \sqrt{\|A\|_1 \|A\|_\infty}, \qquad A \in \mathbb{C}^{n\times n}, \tag{I.7}$$

which is useful in combination with equations (I.3)–(I.5) to estimate the spectral norm of a matrix using its elements. For a proof, see [65, Corollary 2.3.2].

By means of the 2-norm, we define the condition number $\kappa(A)$ of an invertible matrix A as the quantity

$$\kappa(A) = \|A\|_2 \left\|A^{-1}\right\|_2.$$

Notice that if A is normal the condition number can be written as

$$\kappa(A) = \|A\|_2 \left\|A^{-1}\right\|_2 = \frac{\sigma_{\max}(A)}{\sigma_{\min}(A)} = \frac{\max_j |\lambda_j(A)|}{\min_j |\lambda_j(A)|}.$$

Finally, if $A \in \mathbb{C}^{n\times n}$ is HPD, then we define the Euclidean norm weighted by A of a vector $x \in \mathbb{C}^n$ as

$$\|x\|_A = \left\|A^{1/2} x\right\|_2$$

and, consequently, we define the weighted Euclidean norm of a matrix $B \in \mathbb{C}^{n\times n}$ as

$$\|B\|_A = \left\|A^{1/2} B A^{-1/2}\right\|_2.$$

I.2.2 Schatten p-norms

Given $1 \leq p \leq \infty$ and a matrix $A \in \mathbb{C}^{n\times n}$, we denote by $|||A|||_p$ the Schatten p-norm of A, which is defined by

$$|||A|||_p = \left\| [\sigma_j(A)]_{j=1}^n \right\|_p,$$

that is, the Schatten p-norm of A is the p-norm of the vector having as elements the singular values of A. Since the singular values possess a unitary invariance property, this implies that all Schatten p-norms are unitarily invariant.

In what follows we list and describe three significant examples of Schatten p-norms.

- The Schatten ∞-norm coincides with the spectral norm. In fact, the quantity $|||A|||_\infty$ is equal to $\sigma_{\max}(A)$ by definition, which implies that $|||A|||_\infty = \|A\|_2$.
- The Schatten 2-norm $|||A|||_2$ is also known as the Frobenius norm. It coincides with the 2-norm of the vector containing all the elements of A, that is

$$|||A|||_2 = \left(\sum_{i,j=1}^{n} |x_{ij}|^2 \right)^{1/2}, \qquad A \in \mathbb{C}^{n\times n}. \tag{I.8}$$

- The Schatten 1-norm $|||A|||_1$ is also known as trace norm.

I.3 Matrix-valued Functions and Eigenvalue Functions

In the following subsection we deal with the concepts of matrix-valued function and its eigenvalue functions which are broadly used throughout the whole thesis. Given D a measurable subset of $\mathbb{R}^k$, a matrix-valued function brings values $\boldsymbol{\vartheta} \in D$ into the space of square matrices $\mathbb{C}^{s\times s}$.

In general, we state that a matrix-valued function $\mathbf{f}$ possesses a property such as measurability, continuity, and boundedness, if all its components $f_{ij} : D \to \mathbb{C},\ i,j = 1,\dots,s$, possess the same property. We denote by $L^p(D,s)$ the space of the functions such that all their components lay in $L^p(D)$.

Given a function $\mathbf{f} \in L^p(D,s)$, we define

$$\|\mathbf{f}\|_p = \begin{cases} \left(\int_D \|\mathbf{f}(\boldsymbol{\vartheta})\|_p^p \, d\boldsymbol{\vartheta}\right)^{1/p}, & \text{if} \quad 1 \leq p < \infty, \\ \operatorname{ess\,sup}_{\boldsymbol{\vartheta}\in D} \|\mathbf{f}(\boldsymbol{\vartheta})\|_\infty, & \text{if} \quad p = \infty. \end{cases}$$

For a Hermitian matrix-valued function $\mathbf{f}$ we write $\mathbf{f} \geq 0$ (resp. $\mathbf{f} > 0$) if, for almost all $\boldsymbol{\vartheta} \in D$, $\mathbf{f}(\boldsymbol{\vartheta})$ is a non-negative (resp. positive) definite matrix.

In the case where all the eigenvalues of the matrix $\mathbf{f}(\boldsymbol{\vartheta})$ are real for almost every $\boldsymbol{\vartheta} \in D$, we can sort the eigenvalues of the matrix $\mathbf{f}(\boldsymbol{\vartheta})$ in increasing order for almost every $\boldsymbol{\vartheta} \in D$. Hence, the eigenvalue function $\lambda_i(\mathbf{f})$ is well defined as the function taking the value of the i-th largest eigenvalue of $\mathbf{f}(\boldsymbol{\vartheta})$.

Further, adding the hypothesis that $\mathbf{f}$ is a continuous matrix-valued function defined on an interval, the existence and continuity of the eigenvalue functions of $\mathbf{f}$ is proven in [19, Section VI.1] and we summarize the result in the following lemma.

Lemma I.3.1. *Let $\mathbf{f}$ be a continuous map from an interval Q into the space of $s \times s$ matrices such that the eigenvalues of $\mathbf{f}(\vartheta)$ are real for all $\vartheta \in Q$. Then there exist continuous functions $\lambda_1(\mathbf{f}(\vartheta)), \lambda_2(\mathbf{f}(\vartheta)), \dots, \lambda_s(\mathbf{f}(\vartheta))$ such that, for each $\vartheta \in Q$, are the eigenvalues of $\mathbf{f}(\vartheta)$.*

I.4 Asymptotic Distribution of Matrix-Sequences

In the present section we provide definitions and results for the analysis of the spectral and singular value properties of a generic matrix-sequence $\{A_n\}_n$, where the matrices A_n have dimension that increases with n. We also introduce the concepts of clustering, asymptotic distributions, and approximating classes of sequences.

I.4.1 Eigenvalue and Singular Value Distributions of Matrix-Sequences

Before detailing the concepts on asymptotic distributions, we give the definition of cluster of the eigenvalues of a matrix-sequence, which is fundamental in the analysis of preconditioning strategies for Krylov solvers, as we see in Section I.9.3.

Definition I.4.1. *Let $S \subseteq \mathbb{C}$ be a non-empty subset of $\mathbb{C}$ and let $\{A_n\}_n$ be a matrix-sequence, with A_n of increasing size d_n. We say that $\{A_n\}_n$ is strongly clustered at S (in the sense of the eigenvalues), or equivalently that the eigenvalues of $\{A_n\}_n$ are strongly clustered at S, if, for every $\varepsilon > 0$, we have*

$$\#\{j \in \{1, \dots, d_n\} : \ \lambda_j(A_n) \notin D(S, \varepsilon)\} = O(1), \quad as \quad n \to \infty, \tag{I.9}$$

where $D(S, \varepsilon) = \bigcup_{z \in S} \{\omega \in \mathbb{C} : |\omega - z| < \varepsilon\}$.

Furthermore, we say that $\{A_n\}_n$ is (weakly) clustered at S (in the sense of the eigenvalues), or equivalently that the eigenvalues of $\{A_n\}_n$ are (weakly) clustered at S, if $O(1)$ is replaced with $o(d_n)$ in the previous relationships.

When the eigenvalues of a matrix-sequence are clustered at 0 with a weaker meaning, the sequence is said to be sparsely vanishing in the eigenvalue sense. We formally introduce the latter concept in the following definition, specifying the meaning of sparsely vanishing in both the singular value and in the eigenvalue sense.

Definition I.4.2. *Let $\{A_n\}_n$ be a matrix-sequence, with A_n of increasing size d_n. We say that $\{A_n\}_n$ is sparsely vanishing (s.v.) if*

$$\lim_{M \to \infty} \limsup_{n \to \infty} \frac{\#\{i \in \{1, ..., d_n\} : \sigma_i(A_n) < 1/M\}}{d_n} = 0.$$

Moreover, we say that $\{A_n\}_n$ is sparsely vanishing (s.v.) in the eigenvalue sense if

$$\lim_{M \to \infty} \limsup_{n \to \infty} \frac{\#\{i \in \{1, ..., d_n\} : |\lambda_i(A_n)|) < 1/M\}}{d_n} = 0.$$

Throughout the current subsection, we follow all standard notation and terminology introduced in [62]: let $\mathbb{K}$ be either $\mathbb{R}$ or $\mathbb{C}$, then we denote with $C_c(\mathbb{K})$ the space of complex-valued continuous functions defined on $\mathbb{K}$ with bounded support. The following properties hold.

If $\mathbf{g} : D \subset \mathbb{R}^k \to \mathbb{C}^{s \times s}$ is a measurable function defined on a set D with measure $\mu_k(D)$ such that $0 < \mu_k(D) < \infty$, then the expressions $\eta_{\mathbf{g}}^{(\sigma)}$ and $\eta_{\mathbf{g}}^{(\lambda)}$ denote the functionals described by the following relations

$$\eta_{\mathbf{g}}^{(\sigma)} : C_c(\mathbb{R}) \to \mathbb{C} \quad \text{and} \quad \eta_{\mathbf{g}}^{(\sigma)}(F) = \frac{1}{\mu_k(D)} \int_D \frac{\sum_{i=1}^{s} F\left(\sigma_i(\mathbf{g})\right)(\boldsymbol{\vartheta})}{s} \, d\boldsymbol{\vartheta},$$

$$\eta_{\mathbf{g}}^{(\lambda)} : C_c(\mathbb{C}) \to \mathbb{C} \quad \text{and} \quad \eta_{\mathbf{g}}^{(\lambda)}(F) = \frac{1}{\mu_k(D)} \int_D \frac{\sum_{i=1}^{s} F\left(\lambda_i(\mathbf{g})\right)(\boldsymbol{\vartheta})}{s} \, d\boldsymbol{\vartheta},$$

Definition I.4.3. [62, Definition 3.1](Singular value and eigenvalue distribution of a matrix-sequence) *Let $\{A_n\}_n$ be a matrix-sequence with $A_n \in \mathbb{C}^{d_n \times d_n}$.*

1. *We say that $\{A_n\}_n$ has an asymptotic singular value distribution described by a functional $\eta : C_c(\mathbb{R}) \to \mathbb{C}$, and we write $\{A_n\}_n \sim_\sigma \eta$, if*

$$\lim_{n\to\infty} \frac{1}{d_n} \sum_{j=1}^{d_n} F(\sigma_j(A_n)) = \eta(F), \quad \forall F \in C_c(\mathbb{R}).$$

   If $\eta = \eta_{\mathbf{f}}^{(\sigma)}$ for some measurable $\mathbf{f} : D \subset \mathbb{R}^k \to \mathbb{C}^{s\times s}$ defined on a set D with $0 < \mu_k(D) < \infty$, we say that $\{A_n\}_n$ has an asymptotic singular value distribution described by $\mathbf{f}$ and we write $\{A_n\}_n \sim_\sigma \mathbf{f}$. In this case, the function $\mathbf{f}$ is referred to as the singular value symbol of the matrix-sequence $\{A_n\}_n$.

2. *We say that $\{A_n\}_n$ has an asymptotic eigenvalue (or spectral) distribution described by a functional $\eta : C_c(\mathbb{C}) \to \mathbb{C}$, and we write $\{A_n\}_n \sim_\lambda \eta$, if*

$$\lim_{n\to\infty} \frac{1}{d_n} \sum_{j=1}^{d_n} F(\lambda_j(A_n)) = \eta(F), \quad \forall F \in C_c(\mathbb{C}).$$

   If $\eta = \eta_{\mathbf{f}}^{(\lambda)}$ for some measurable $\mathbf{f} : D \subset \mathbb{R}^k \to \mathbb{C}^{s\times s}$ defined on a set D with $0 < \mu_k(D) < \infty$, we say that $\{A_n\}_n$ has an asymptotic eigenvalue (or spectral) distribution described by $\mathbf{f}$ and we write $\{A_n\}_n \sim_\lambda \mathbf{f}$. In this case, the function $\mathbf{f}$ is referred to as the eigenvalue (or spectral) symbol of the matrix-sequence $\{A_n\}_n$.

Recalling that a function f is sparsely vanishing if and only if its set of zeros is of Lebesgue measure zero, we report the following result, which is proven in [62, Chapter 9, pp. 165–166]).

Theorem I.4.1. *The following statements are true.*

1. *Assume $\{A_n\}_n \sim_\sigma f$. Then $\{A_n\}_n$ is sparsely vanishing if and only if f is sparsely vanishing.*

2. *Assume $\{A_n\}_n \sim_\lambda f$. Then $\{A_n\}_n$ is sparsely vanishing in the eigenvalue sense if and only if f is sparsely vanishing.*

3. *Assume $\{A_n\}_n$ is given and assume that every matrix A_n is normal. Then $\{A_n\}_n$ is sparsely vanishing if and only if $\{A_n\}_n$ is sparsely vanishing in the eigenvalue sense.*

Finally, we recall that the essential range $\mathcal{ER}(f)$ of a function $f : D \subset \mathbb{R} \to \mathbb{C}$ is defined as the set of points $z \in \mathbb{C}$ such that, for every $\varepsilon > 0$, the measure of the set $\{f(\vartheta) \in \{\omega \in \mathbb{C} : |\omega - z| < \varepsilon\}$ is positive. The previous notion has a direct relation with the concept of spectral distribution. Indeed, if $\{A_n\}_n \sim_\lambda f$, then $\{A_n\}_n$ is weakly clustered at the essential range $\mathcal{ER}(f)$. See [62] for more details.

I.4.2 Approximating Classes of Sequences

Next, we introduce the definition and a key lemma on approximating classes of sequences [118], which is used for completing the proofs of the results presented in **Chapters II–III**.

Definition I.4.4. [62, Definition 5.1](approximating class of sequences)

Let $\{A_n\}_n$ be a matrix-sequence and let $\{\{B_{n,m}\}_n\}_m$ be a sequence of matrix-sequences. We say that $\{\{B_{n,m}\}_n\}_m$ is an approximating class of sequences (a.c.s.) for $\{A_n\}_n$ if the following condition is met: for every m there exist n_m, $c(m)$, $\omega(m)$ such that, for $n \geq n_m$,

$$A_n = B_{n,m} + R_{n,m} + N_{n,m},$$

$$rank\ R_{n,m} \leq c(m)n \quad and \quad \|N_{n,m}\|_2 \leq \omega(m),$$

where the quantities n_m, $c(m)$, and $\omega(m)$ depend only on m and

$$\lim_{m\to\infty} c(m) = \lim_{m\to\infty} \omega(m) = 0.$$

We use $\{B_{n,m}\}_n \xrightarrow{\text{a.c.s. wrt } m} \{A_n\}_n$ to denote that $\{\{B_{n,m}\}_n\}_m$ is an a.c.s. for $\{A_n\}_n$.

Lemma I.4.2. [62, Corollary 5.1] *Let $\{A_n\}_n, \{B_{n,m}\}_n$ be matrix-sequences and let $f, f_m : D \subset \mathbb{R}^k \to \mathbb{C}$ be measurable functions defined on a set D with $0 < \mu_k(D) < \infty$. Suppose that*

1. *$\{B_{n,m}\}_n \sim_\sigma f_m$ for every m,*
2. *$\{B_{n,m}\}_n \xrightarrow{a.c.s.\ wrt\ m} \{A_n\}_n$,*
3. *$f_m \to f$ in measure.*

Then

$$\{A_n\}_n \sim_\sigma f.$$

Furthermore, if the first assumption is replaced by $\{B_{n,m}\}_n \sim_\lambda f_m$ for every m, given that the other two assumptions are left unchanged, and all the involved matrices are Hermitian, then we conclude that $\{A_n\}_n \sim_\lambda f$.

I.5 Toeplitz Structures

The current section is devoted to Toeplitz matrices, a topic that is recurrent throughout the whole thesis and that has been the subject of several books [22, 24, 25, 67]. A matrix is said to have a Toeplitz structure if it has constant diagonals, either element by element or in a block sense. As we rigorously state in the following subsections, in some cases the components can be seen as the Fourier coefficients of a function that "generates" the Toeplitz matrix. The type of domain (either $[-\pi, \pi]$ or $[-\pi, \pi]^k$) and codomain (either the complex field or the space of $s \times s$ complex matrices) of the generating function gives rise to different kinds of Toeplitz matrices, see Table I.1 for a complete overview.

Type of generating function		Associated Toeplitz matrix	
univariate scalar	$f(\vartheta) : [-\pi, \pi] \to \mathbb{C}$	unilevel scalar	$T_n[f] \in \mathbb{C}^{n\times n}$
k-variate scalar	$f(\boldsymbol{\vartheta}) : [-\pi, \pi]^k \to \mathbb{C}$	multilevel scalar	$T_{\mathbf{n}}[f] \in \mathbb{C}^{\mathcal{N}(\mathbf{n})\times\mathcal{N}(\mathbf{n})}$
univariate matrix-valued	$\mathbf{f}(\vartheta) : [-\pi, \pi] \to \mathbb{C}^{s\times s}$	unilevel block	$T_n[\mathbf{f}] \in \mathbb{C}^{sn\times sn}$
k-variate matrix-valued	$\mathbf{f}(\boldsymbol{\vartheta}) : [-\pi, \pi]^k \to \mathbb{C}^{s\times s}$	multilevel block	$T_{\mathbf{n}}[\mathbf{f}] \in \mathbb{C}^{s\mathcal{N}(\mathbf{n})\times s\mathcal{N}(\mathbf{n})}$

Table I.1: Various types of generating function and the associated Toeplitz matrices.

I.5.1 Unilevel Scalar Toeplitz Matrices

A matrix of the form

$$A = [a_{i-j}]_{i,j=1}^{n} = \begin{bmatrix} a_0 & a_{-1} & a_{-2} & \cdots & \cdots & a_{-(n-1)} \\ a_1 & \ddots & \ddots & \ddots & & \vdots \\ a_2 & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & a_{-2} \\ \vdots & & \ddots & \ddots & \ddots & a_{-1} \\ a_{n-1} & \cdots & \cdots & a_2 & a_1 & a_0 \end{bmatrix}, \tag{I.10}$$

is called a Toeplitz matrix. Notice that the (i, j)-th entry of A depends only on the difference $i - j$, which means that the components are constant along each diagonal.

It is straightforward to see that the Toeplitz matrix $[a_{i-j}]_{i,j=1}^{n}$ can be written as the sum

$$[a_{i-j}]_{i,j=1}^{n} = \sum_{k=-(n-1)}^{n-1} a_k J_n^{(k)}, \tag{I.11}$$

where, for $k \in \mathbb{Z}$,

$$\left[J_n^{(k)}\right]_{ij} = \begin{cases} 1, & \text{if } i - j = k, \\ 0, & \text{otherwise} \end{cases}. \tag{I.12}$$

Throughout this subsection, we assume that $f \in L^1([-\pi, \pi])$ and is periodically extended to the real line. The Fourier coefficients of f are denoted by

$$\hat{f}_k = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(\vartheta) \mathrm{e}^{-\mathrm{i}\, k\vartheta}\, d\vartheta, \qquad k \in \mathbb{Z}. \tag{I.13}$$

The n-th Toeplitz matrix associated with f is defined as

$$T_n[f] = \left[\hat{f}_{i-j}\right]_{i,j=1}^{n} = \sum_{k=-(n-1)}^{n-1} \hat{f}_k\, J_n^{(k)}. \tag{I.14}$$

We call $\{T_n[f]\}_n$ the sequence of Toeplitz matrices associated with f (or generated by f), which in turn is referred to as the generating function of $\{T_n[f]\}_n$.

In the following list, we describe how some properties of the generating function reflect to the associated Toeplitz matrix, see [62, 99] for more details.

1. If f is real-valued, then $T_n[f]$ is Hermitian for all n.

2. If f is even, then $T_n[f]$ is symmetric for all n.

3. If f is real-valued and even, then $T_n[f]$ is real and symmetric for all n.

4. If f is a trigonometric polynomial, that is,

$$f(\vartheta) = \sum_{k=-r}^{r} \hat{f}_k \, \mathrm{e}^{\hat{\imath}\, k\vartheta},$$

then, for n sufficiently large, $T_n[f]$ is a banded matrix, with bandwidth bounded by $2r+1$.

We conclude this subsection by reporting two theorems that are useful to extract information from the generating function on the eigenvalues and singular values of the associated Toeplitz matrix, for the proof see [62].

Theorem I.5.1. *Assume that $f \in L^1([-\pi,\pi])$ is real a.e. and let*

$$m_f = \operatorname*{ess\,inf}_{\vartheta\in[-\pi,\pi]} f(\vartheta), \qquad M_f = \operatorname*{ess\,sup}_{\vartheta\in[-\pi,\pi]} f(\vartheta).$$

Then

$$\Lambda(T_n[f]) \subseteq [m_f, M_f], \quad n \in \mathbb{N}.$$

If we also assume that $m_f < M_f$, then

$$\Lambda(T_n[f]) \subset (m_f, M_f), \quad n \in \mathbb{N}.$$

Theorem I.5.2. *Let $f \in L^p([-\pi,\pi])$, $n \in \mathbb{N}$ and $1 \le p \le \infty$. Then*

$$|||T_n[f]|||_p \le \frac{n^{1/p}}{(2\pi)^{1/p}} \|f\|_{L^p}.$$

In particular, for $p = \infty$ we have

$$\|T_n[f]\|_2 = |||T_n[f]|||_\infty \le \|f\|_{L^\infty}.$$

I.5.2 Block and Multilevel Block Toeplitz Matrices

The current subsection is dedicated to the generalization of the concept of unilevel scalar Toeplitz matrix. We start with the definition of block-Toeplitz matrix, which has been conceptualized from the idea that the entries a_k of the matrix $A_n = [a_{i-j}]_{i,j=1}^n$ could be matrices themselves.

Given $A_{-(n-1)}, \ldots, A_{n-1} \in \mathbb{C}^{s\times s}$, the $sn \times sn$ block matrix A defined by

$$A = [A_{i-j}]_{i,j=1}^n = \begin{bmatrix} A_0 & A_{-1} & A_{-2} & \cdots & \cdots & A_{-(n-1)} \\ A_1 & \ddots & \ddots & \ddots & & \vdots \\ A_2 & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & A_{-2} \\ \vdots & & \ddots & \ddots & \ddots & A_{-1} \\ A_{n-1} & \cdots & \cdots & A_2 & A_1 & A_0 \end{bmatrix},$$

is said to be a block-Toeplitz matrix.

Analogously to the scalar case, a block-Toeplitz matrix of the form $T_n(\mathbf{f})$ is associated to a matrix-valued function $\mathbf{f} \in L^1([-\pi,\pi], s)$. We formalize the latter statement in the following definition.

Definition I.5.1. *Let the Fourier coefficients of a given function* $\mathbf{f} \in L^1([-\pi,\pi], s)$ *be*

$$\hat{f}_j := \frac{1}{2\pi} \int_{[-\pi,\pi]} \mathbf{f}(\vartheta) \mathrm{e}^{-\hat{\imath} j\vartheta} d\vartheta \in C^{s\times s}, \qquad j \in \mathbb{Z}.$$

Then, the block-Toeplitz matrix associated with $\mathbf{f}$ *is the matrix of order* sn *given by*

$$T_n[\mathbf{f}] = \sum_{|j|<n} J_n^{(j)} \otimes \hat{f}_j,$$

where the term $J_n^{(j)}$ *is defined in (I.12). The matrix-sequence* $\{T_n[\mathbf{f}]\}_n$ *is called the block-Toeplitz sequence generated by* f*, that in turn is referred to as the generating function of* $\{T_n[\mathbf{f}]\}_n$*.*

If the blocks are Toeplitz matrices themselves, the matrix is said to be a block-Toeplitz matrix with Toeplitz blocks, or BTTB matrix. However, in the case of a block-Toeplitz sequence $\{T_n[\mathbf{f}]\}_n$ generated by $\mathbf{f} \in L^1([-\pi,\pi], s)$, the blocks have a fixed dimension, that is, the block-size does not depend on n. BTTB matrices suggest instead an additional generalization of the concept of Toeplitz matrix-sequences, that is, the extension to sequences of BTTB matrices in which both the block-size and the number of blocks depend on n. Such matrices are said to be 2-level Toeplitz matrix-sequences, and, in general, are substantially different from block-Toeplitz matrix-sequences if we consider the type of generating function, which in this setting is bivariate and scalar-valued.

The following definition formalizes all the previous considerations and extends the definition of Toeplitz matrix to the most general setting that we consider, that is, the case of multilevel block-Toeplitz matrix-sequences associated with multivariate matrix-valued generating functions.

Definition I.5.2. *Let the Fourier coefficients of a given function* $\mathbf{f} \in L^1([-\pi,\pi]^k, s)$ *be defined as*

$$\hat{f}_{\mathbf{j}} := \frac{1}{(2\pi)^k} \int_{[-\pi,\pi]^k} \mathbf{f}(\boldsymbol{\vartheta}) \mathrm{e}^{-\hat{\imath}\langle \mathbf{j}, \boldsymbol{\vartheta}\rangle} \, \mathrm{d}\boldsymbol{\vartheta} \in \mathbb{C}^{s\times s}, \qquad \mathbf{j} = (j_1, \ldots, j_k) \in \mathbb{Z}^k,$$

where $\langle \mathbf{j}, \boldsymbol{\vartheta}\rangle = \sum_{t=1}^k j_t \vartheta_t$*.*

Given a k*-index* $\mathbf{n} = (n_1, n_2, \ldots, n_k)$*, the* $\mathbf{n}$*-th* k*-level* $s \times s$ *Toeplitz matrix associated with* $\mathbf{f}$ *is the matrix of order* $sn_1n_2\ldots n_k$ *given by*

$$T_{\mathbf{n}}[\mathbf{f}] = \sum_{\mathbf{j}=-(\mathbf{n}-\mathbf{1})}^{\mathbf{n}-\mathbf{1}} J_{n_1}^{(j_1)} \otimes \cdots \otimes J_{n_k}^{(j_k)} \otimes \hat{f}_{\mathbf{j}},$$

where $\mathbf{j} = (j_1, \ldots, j_k) \in \mathbb{N}^k$ *and the matrices of the form* $J_m^{(h)}$ *are defined in (I.12).*

I.5.3 Asymptotic Distribution of Toeplitz Sequences

The singular value and spectral distribution of Toeplitz matrix-sequences have been well studied in the past few decades. Ever since Szegő in [67] showed that the eigenvalues of the Toeplitz matrix $T_n[f]$ generated by real-valued $f \in L^\infty([-\pi,\pi])$ are asymptotically distributed as f, such result has undergone many generalizations and extensions. Under the same assumption on f, Avram and Parter [6, 103] proved that the singular values of $T_n[f]$ are distributed as $|f|$ and Tyrtyshnikov [134, 135, 140] later extended the latter result for $T_n[f]$ generated by complex-valued $f \in L^1([-\pi,\pi])$.

The generalized Szegő theorem that describes the singular value and spectral distribution of Toeplitz sequences generated by $f \in L^1([-\pi,\pi])$ is given as follows. We refer to [140] for the original results and [62, Theorem 6.5] for a proof that is based on the notion of approximating class of sequences given in Definition I.4.4.

Theorem I.5.3. *Suppose $f \in L^1([-\pi,\pi])$. Let $T_n[f]$ be the Toeplitz matrix generated by f. Then*

$$\{T_n[f]\}_n \sim_\sigma f.$$

Moreover, if f is real-valued, then

$$\{T_n[f]\}_n \sim_\lambda f.$$

Furthermore, Tilli [133] generalized the proof to the block-Toeplitz setting and, in particular, we report the following theorem, which is the extension of the eigenvalue result to the case of multivariate Hermitian matrix-valued generating functions.

Theorem I.5.4. *Let $\mathbf{f} \in L^1([-\pi,\pi]^k, s)$ be a Hermitian matrix-valued function with $k \geq 1, s \geq 2$. Then,*

$$\{T_{\mathbf{n}}[\mathbf{f}]\}_{\mathbf{n}\in\mathbb{N}^k} \sim_\lambda \mathbf{f}.$$

I.6 Circulant Matrices

Circulant matrices are special Toeplitz matrices which possess the additional property that each column vector is a circular shift of the preceding column vector. On one hand, all the theory presented in Section I.5 for Toeplitz matrices remains valid for circulant matrices. On the other hand, circulant matrices of a fixed size n form an algebra of matrices unitarily diagonalized by the Fourier matrix F_n. In what follows we focus on the latter aspect.

We begin with the definition of $\mathbf{n}$-th Fourier sum of a matrix-valued function $\mathbf{f} \in L^1([-\pi,\pi]^k, s)$, which is given by

$$\left(S_{\mathbf{n}}[\mathbf{f}]\right)(\boldsymbol{\vartheta}) = \sum_{j_1=1-n_1}^{n_1-1} \cdots \sum_{j_k=1-n_k}^{n_k-1} \hat{f}_{\mathbf{j}} \mathrm{e}^{\hat{\imath}\langle \mathbf{j},\boldsymbol{\vartheta}\rangle}, \qquad \langle \mathbf{j},\boldsymbol{\vartheta}\rangle = \sum_{t=1}^{k} j_t \vartheta_t. \tag{I.15}$$

A particular uniform sampling of the function in (I.15) is crucial in the construction of the circulant matrix associated with $\mathbf{f}$, as we see in the following definition.

Definition I.6.1. *Let $\mathbf{f} \in L^1([-\pi,\pi]^k, s)$ be a matrix-valued function and let $\mathbf{n} = (n_1, \ldots, n_k)$ be a k-index. Then, the $\mathbf{n}$-th circulant matrix generated by $\mathbf{f}$ is the matrix of order $s\mathcal{N}(\mathbf{n})$ defined by:*

$$\mathcal{C}_{\mathbf{n}}[\mathbf{f}] = (F_{\mathbf{n}} \otimes I_s) D_{\mathbf{n}}[\mathbf{f}] (F_{\mathbf{n}} \otimes I_s)^H, \tag{I.16}$$

where

$$D_{\mathbf{n}}[\mathbf{f}] = \underset{\mathbf{0} \le \mathbf{r} \le \mathbf{n-1}}{\operatorname{diag}} \left(S_{\mathbf{n}}[\mathbf{f}]\right) \left(\vartheta_{\mathbf{r}}^{(\mathbf{n})}\right) \tag{I.17}$$

is a block-diagonal matrix and

$$\vartheta_{\mathbf{r}}^{(\mathbf{n})} = 2\pi \frac{\mathbf{r}}{\mathbf{n}}, \quad F_{\mathbf{n}} = \frac{1}{\sqrt{\mathcal{N}(\mathbf{n})}} \left(\mathrm{e}^{-\hat{\imath}\left\langle \mathbf{j}, \vartheta_{\mathbf{r}}^{(\mathbf{n})} \right\rangle}\right)_{\mathbf{j},\mathbf{r}=\mathbf{0}}^{\mathbf{n-1}}. \tag{I.18}$$

Notice that in the case where f is a univariate scalar trigonometric polynomial, the n-th Fourier sum coincides with f if n is large enough, and hence we can write the circulant matrix generated by f as

$$\mathcal{C}_n[f] = F_n \underset{i \in \mathcal{I}_n}{\operatorname{diag}} \left(f(\vartheta_i^{(n)})\right) F_n^H, \tag{I.19}$$

where the grid points $\vartheta_i^{(n)}$ are $\frac{2\pi i}{n}$ and i belongs to the index range $\mathcal{I}_n = \{0, \ldots, n-\ 1\}$. From (I.19) it is clear that, for n large enough, the eigenvalues of $\mathcal{C}_n[f]$ are given by the evaluations of f at the grid points.

Example 1. *Consider the trigonometric polynomial $p(\vartheta) = 2 - 3\mathrm{e}^{-\hat{\imath}\vartheta}$. According to equation I.19, the circulant matrix generated by p of order 4 is*

$$\begin{aligned}
\mathcal{C}_4[p] =& F_4 \begin{bmatrix} 2-3\mathrm{e}^{-\hat{\imath}\vartheta_0^{(4)}} & 0 & 0 & 0 \\ 0 & 2-3\mathrm{e}^{-\hat{\imath}\vartheta_1^{(4)}} & 0 & 0 \\ 0 & 0 & 2-3\mathrm{e}^{-\hat{\imath}\vartheta_2^{(4)}} & 0 \\ 0 & 0 & 0 & 2-3\mathrm{e}^{-\hat{\imath}\vartheta_3^{(4)}} \end{bmatrix} F_4^H = \\
& \frac{1}{4} \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & -\hat{\imath} & -1 & \hat{\imath} \\ 1 & -1 & 1 & -1 \\ 1 & \hat{\imath} & -1 & -\hat{\imath} \end{bmatrix} \begin{bmatrix} -1 & 0 & 0 & 0 \\ 0 & 2+3\hat{\imath} & 0 & 0 \\ 0 & 0 & 5 & 0 \\ 0 & 0 & 0 & 2-3\hat{\imath} \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & -\hat{\imath} & -1 & \hat{\imath} \\ 1 & -1 & 1 & -1 \\ 1 & \hat{\imath} & -1 & -\hat{\imath} \end{bmatrix}^H = \\
& \begin{bmatrix} 2 & -3 & 0 & 0 \\ 0 & 2 & -3 & 0 \\ 0 & 0 & 2 & -3 \\ -3 & 0 & 0 & 2 \end{bmatrix}.
\end{aligned}$$

Hence, the eigenvalues $-1, 2 \pm 3\hat{\imath}, 5$ of $\mathcal{C}_4[p]$ are obtained by evaluating the generating function p on the grid points $\vartheta_j^{(4)}$, $j = 0, \ldots,$.

Analogously, if $\mathbf{f}$ is a univariate $s \times s$ matrix-valued trigonometric polynomial the block-circulant matrix generated by $\mathbf{f}$ can be written as

$$\mathcal{C}_n[\mathbf{f}] = (F_n \otimes I_s) \underset{i \in \mathcal{I}_n}{\operatorname{diag}} \left(\mathbf{f}(\vartheta_i^{(n)})\right) (F_n^H \otimes I_s), \tag{I.20}$$

where $\underset{i\in\mathcal{I}_n}{\text{diag}}\left(\mathbf{f}(\vartheta_i^{(n)})\right)$ is the block-diagonal matrix with the block-diagonal elements being the evaluation of $\mathbf{f}$ on the grid points $\vartheta_i^{(n)}$ for $i \in \mathcal{I}_n$. Given the block structure of the decomposition in (I.20), it is clear that the eigenvalues of $\mathcal{C}_n[\mathbf{f}]$ are given by $\lambda_t\left(\mathbf{f}\left(\vartheta_i^{(n)}\right)\right)$, varying $t = 1, \ldots, s$ and $i \in \mathcal{I}_n$.

We conclude the current subsection by recalling that, when circulant matrix-sequences with elements of increasing size d_n are involved, operations such as the matrix-vector product, the solution of a linear system and the computation of eigenvalues have computational cost $O(d_n \log d_n)$. Indeed, the essence of calculations with circulants is the exploitation of the Fast Fourier Transform (FFT) algorithm for multiplying a vector by the Fourier matrix [32, 138].

I.7 Generalized Locally Toeplitz Sequences

In the sequel, we briefly present the class of Generalized Locally Toeplitz (GLT) sequences [12, 121, 122] in their multilevel block form. GLT sequences constitute a $*$-algebra of matrix-sequences to which multilevel block-Toeplitz matrix-sequences with Lebesgue integrable generating functions belong. The formal definition of the GLT class requires rather technical tools, hence here we only list some properties, which are sufficient for studying the asymptotic distributions of the matrix-sequences that we deal with in **Chapter III**. See [12, 62, 63] for complete discussions on the topic.

GLT1 Each GLT sequence $\{A_{\mathbf{n}}\}_{\mathbf{n}}$ has a singular value symbol $\tilde{\mathbf{f}} : [0,1]^k \times [-\pi,\pi]^k \to \mathbb{C}^{s\times s}$. If all the matrices of the sequence are Hermitian, then the distribution also holds in the eigenvalue sense. We call $\tilde{\mathbf{f}}(\mathbf{x},\boldsymbol{\vartheta})$ the (GLT) symbol of $\{A_{\mathbf{n}}\}_{\mathbf{n}}$ and we write $\{A_{\mathbf{n}}\}_{\mathbf{n}} \sim_{\text{GLT}} \tilde{\mathbf{f}}$.

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

GLT3 Every Toeplitz sequence generated by a function $\mathbf{f} \in L^1([-\pi,\pi]^k, s)$ is a GLT sequence and its symbol is $\tilde{\mathbf{f}}(\mathbf{x},\boldsymbol{\vartheta}) = \mathbf{f}(\boldsymbol{\vartheta})$, with the specifications reported in Item **GLT1**.

GLT4 Every sequence which is distributed as the constant zero in the singular value sense is a GLT sequence with symbol $\tilde{\mathbf{f}} \equiv \mathbf{0}$.

GLT5 $\{A_{\mathbf{n}}\}_{\mathbf{n}} \sim_{\text{GLT}} \tilde{\mathbf{f}}$ if and only if there exist GLT sequences $\{B_{\mathbf{n},m}\}_{\mathbf{n}} \sim_{\text{GLT}} \tilde{\mathbf{f}}_m$ such that $\tilde{\mathbf{f}}_m$ converges to $\tilde{\mathbf{f}}$ in measure and $\{\{B_{\mathbf{n},m}\}_{\mathbf{n}}\}_m$ is an a.c.s. for $\{A_{\mathbf{n}}\}_{\mathbf{n}}$.

The advantage of dealing with Hermitian matrix-sequences is clear from **GLT1**. Indeed, in this setting, we can use these GLT properties to study also the asymptotic spectral features of the involved matrix-sequences. However, useful relaxations of such hypothesis are introduced and discussed in [62].

I.8 Functions of Matrices

Let h be a real analytic function centred at $z_0 = 0$ with radius of convergence r. If $|z| < r$, we can represent $h(z)$ through its Taylor series expansion in $z_0 = 0$, that is $h(z) = \sum_{k=0}^{\infty} b_k z^k$. We exploit this representation to define the corresponding matrix function $h(A)$, with A being an n-by-n matrix. Notice that, given a real analytic function h through its explicit Taylor series expansion in 0, we denote both the function defined on a subset of $\mathbb{C}$ and the function defined on a subset of $\mathbb{C}^{n\times n}$ by h.

Assume that $\Lambda(A) \subset \{z \in \mathbb{C} : |z| < r\}$, then Theorem 4.7 in [74] assures that the series $\sum_{k=0}^{\infty} b_k A^k$ converges. Hence, $h(A)$ is well-defined by

$$h(A) = \sum_{k=0}^{\infty} b_k A^k.$$

Now, we want to investigate the latter definition in the Toeplitz case. Let us consider the Toeplitz matrix $T_n[f] \in \mathbb{R}^{n\times n}$ generated by a function $f \in L^\infty([-\pi,\pi])$ with real Fourier coefficients. Recalling Theorem I.5.2 and the relation $\rho(T_n[f]) \le \|T_n[f]\|_2$, we see that $\rho(T_n[f]) < \|f\|_\infty$. Hence, if we take a real analytic function $h(z)$ with radius of convergence r such that $\|f\|_\infty < r$, then $h(T_n[f])$ is well-defined.

In **Chapter II** we consider a symmetrization strategy for the Toeplitz matrix $T_n[f]$ with real components that consists in pre-multiplying it by the anti-identity matrix and obtain the real symmetric matrix $Y_n T_n[f]$. Looking more closely at the latter procedure, we see that the symmetry of $Y_n T_n[f]$ is a consequence of the persymmetry of $T_n[f]$, which exactly means that $Y_n T_n[f] = T_n[f]^T Y_n$. In order to extend the applicability of the symmetrization procedure to $h(T_n[f])$, as we do in **Chapter III**, one needs to prove that $h(T_n[f])$ is persymmetric, and in fact this is done in [78]. We report the result for completeness.

Lemma I.8.1. [78, Lemma 6] *Assume that $h(z)$ is analytic on $|z| < r$. If $A_n \in \mathbb{R}^{n\times n}$ with $\rho(A_n) < r$ is (real) persymmetric, i.e. $Y_n A_n = A_n^T Y_n$, then $h(A_n)$ is also (real) persymmetric.*

Since the coefficients b_k, with integer k, are all real, we deal with real symmetric matrices $Y_n h(T_n[f])$.

I.9 Iterative Methods

The current section is dedicated to iterative methods, which represent a convenient tool for the solution of large linear systems in the case where the coefficient matrix is sparse or possesses an exploitable structure.

Let us fix an invertible matrix $A \in \mathbb{C}^{n\times n}$ and a vector $b \in \mathbb{C}^n$. The linear system

$$Ax = b$$

has exactly one solution $x = A^{-1}b$.

A convergent iterative method theoretically consists in the construction of a sequence of iterates $\{x^{(k)}\}_k$ such that the quantity $\|x^{(k)} - x\|_2$ tends to 0 as k tends to ∞. However, for an actual implementation of the iterative method, one needs to choose an index $\tilde{k}$ such that the

procedure stops at iteration $\tilde{k}$ and returns $x^{(\tilde{k})}$ as the solution of the linear system, approximated up to a desired precision. In other words, the iterates $x^{(k)}$ constitute successive approximations of the solution x, and one needs to choose when the approximation is good enough for stopping the iteration. For this purpose, let us define the k-th error by

$$\mathrm{e}^{(k)} = x^{(k)} - x$$

and the k-th residual by

$$r^{(k)} = b - Ax^{(k)}.$$

Ideally, a stopping criterion would be based on the quantity $\mathrm{e}^{(k)}$, because it measures how close to the true solution the k-th iterate is. In practice, this is not possible since x is unknown and in many cases the stopping criterion involves the residual.

If a problem is ill-conditioned, we need to pay particular attention to the efficiency of the chosen iterative method. To this end, Axelsson and Neytcheva [8] have proposed two criteria to judge the performance of a method in the case of linear systems stemming from differential problems: the optimal rate of convergence – independent of the level number – and the optimal order of computational complexity – proportional to the degrees of freedom of the problem. In a subsequent work, Serra [111] has generalized such criteria to the case of general iterative methods for nested linear systems $A_n x_n = b_n$, where $\{A_n\}$ is an asymptotically ill-conditioned class, as follows:

'Opt1': An iterative method is said optimal in the sense of the convergence rate if the convergence speed is independent of the dimension of the matrix A_n.

'Opt2': An iterative method is said optimal with respect to the arithmetic cost if the cost of a single iteration, as a function of n, has the same asymptotic order than the cost of a generic product between the matrix A_n and a given vector y.

We say that a method is optimal if it possesses both property **'Opt1'** and **'Opt2'**. In order to develop an efficient iterative method for the solution of our linear system, from property **'Opt2'** we see that one iteration of the method should have a reasonable computational cost, possibly proportional to the cost of the matrix-vector product with matrix A_n. Property **'Opt1'** suggests that the aforementioned index $\tilde{k}$ such that $x^{(\tilde{k})}$ approximates the true solution up to the desired precision should not depend on the matrix-size.

In the following subsections we recall the key features of some well-known classes of iterative methods. Firstly, in Subsection I.9.1 we define stationary iterative solvers, focusing in particular on the Jacobi and Gauss–Seidel methods. In Subsection I.9.2 we present Krylov subspace methods and see how the study of the spectral properties of the involved matrices can be crucial in the a priori estimates of the error and the residual behaviours. Moreover, we provide the basics of preconditioning and report some known strategies in the Toeplitz case. Finally, Subsection I.9.4 is dedicated to algebraic multigrid methods.

I.9.1 Stationary Methods

Given an invertible matrix $A \in \mathbb{C}^{n\times n}$ and a vector $b \in \mathbb{C}^n$, a stationary iterative method for the solution of the linear system $Ax = b$ consists in choosing an initial guess $x^{(0)} \in \mathbb{C}^n$ and

computing the successive iterates with the formula

$$x^{(k+1)} = Sx^{(k)} + q, \tag{I.21}$$

where $S \in \mathbb{C}^{n\times n}$ is referred to as the iteration matrix and $q \in \mathbb{C}^n$ is a fixed vector. The peculiarity of stationary iterative methods is that the matrix S is fixed, that is, it does not depend on the step k.

The iteration matrix plays a crucial role for the convergence analysis of the associated method. Indeed, the following result provides a sufficient and necessary condition that links the spectral radius $\rho(S)$ to the convergence of the stationary iterative procedure.

Proposition I.9.1. *[108, Theorem 4.1] Let S be a square matrix such that $\rho(S) < 1$. Then $I - S$ is non-singular and the iteration (I.21) converges for any b and $x^{(0)}$. Conversely, if the iteration (I.21) converges for any b and $x^{(0)}$, then $\rho(S) < 1$.*

A general strategy to develop a stationary iterative method is to consider the decomposition

$$A = M - (M - A),$$

where M is an invertible matrix in $\mathbb{C}^{n\times n}$. It is straightforward to see that x is the solution of the linear system $Ax = b$ if and only if the equality

$$x = (I - M^{-1}A)x + M^{-1}b,$$

is verified and, exploiting this observation, given $x^{(0)} \in \mathbb{C}^n$ we can define the method

$$x^{(k+1)} = (I - M^{-1}A)x^{(k)} + M^{-1}b, \tag{I.22}$$

for which the iteration matrix S is equal to $(I - M^{-1}A)$.

In the following, we present the iteration structure of the Richardson, Jacobi and Gauss-Seidel methods. For this purpose, let us define the matrices $B = [b_{ij}]_{i,j=1}^n$ and $D = [d_{ij}]_{i,j=1}^n$ from $A = [a_{ij}]_{i,j=1}^n$ by the formulae

$$b_{ij} = \begin{cases} -a_{ij}, & \text{if } i > j, \\ 0, & \text{if } i \leq j \end{cases} \quad ; \qquad d_{ij} = \begin{cases} a_{ij}, & \text{if } i = j, \\ 0, & \text{if } i \neq j \end{cases} \quad .$$

Given $\omega \in \mathbb{C}\backslash\{0\}$, the relaxed Richardson, Jacobi and Gauss-Seidel methods are defined by iterations of the form (I.22), where

- if we choose $M = \frac{1}{\omega}I$, we obtain the relaxed Richardson method;
- if we choose $M = \frac{1}{\omega}D$, we obtain the relaxed Jacobi method;
- if we choose $M = \frac{1}{\omega}D - B$, we obtain the relaxed Gauss-Seidel method.

For a detailed convergence analysis of these methods we remand to [108].

I.9.2 Krylov Methods

Krylov subspace methods are a successful class of iterative solvers for a system of linear equations of the form

$$Ax = b$$

that consists in selecting the k-th iterate $x^{(k)}$ from the affine space

$$x^{(0)} + \text{span}\left\{r^{(0)}, Ar^{(0)}, A^2r^{(0)}, \dots, A^{k-1}r^{(0)}\right\}$$

such that the residual $r^{(k)}$ is orthogonal to a given subspace $\mathcal{L}_k$. The feature that characterizes one Krylov solver from the others is the choice of the subspace $\mathcal{L}_k$.

One of the most celebrated Krylov subspace methods is the Conjugate Gradient (CG) method, developed by Hestenes and Stiefel [71]. However, this method requires for convergence that the matrix of coefficients is a HPD matrix, which is quite a strong restriction. In 1975 Paige and Saunders [102] developed the Minimal Residual (MINRES) method for Hermitian – but in general indefinite – matrices, which is the case of the matrix-sequences that we study in **Chapters II–III**. A method that can successfully be applied to an even larger class of linear systems is the Generalized Minimal Residual (GMRES) method, developed by Saad and Schultz [109] in 1986, which is suitable also for non-Hermitian matrices.

On the other hand, the CG and MINRES methods possess a significant advantage with respect to the GMRES method: the convergence rates can be estimated using only the eigenvalues of the system matrix. For instance, for the CG method the Axelsson–Lindskog estimates hold [7], while for the MINRES method analogous results can be deduced from the following inequality [66], which provides a sharp bound for the residual at iteration k:

$$\|r^{(k)}\|_2/\|r^{(0)}\|_2 \leq \min_{\substack{p_k \in \mathbb{R}_k[\lambda] \\ p_k(0)=1}} \max_{i=1,\dots,n} |p_k(\lambda_i)|, \tag{I.23}$$

where $\mathbb{R}_k[\lambda]$ is the space of polynomials with coefficients in $\mathbb{R}$ of degree less than or equal to k. If the eigenvalues of the Hermitian system matrix are known, the convergence rate of MINRES can be studied a priori choosing an appropriate polynomial p_k in the expression $\max_{i=1,\dots,n} |p_k(\lambda_i)|$.

From Equation (I.23), it is immediate to see that the knowledge of the spectral features of the system matrix is crucial in the development of some Krylov solvers. Moreover, the estimate provides a first intuitive reason why some eigenvalue distributions are more desirable than others for the convergence of the MINRES method. Namely, if the eigenvalues are clustered around a single non-zero point α, the convergence rate is satisfactory: for instance, we can consider the polynomial $p_k(\lambda) = (1 - \lambda/\alpha)^k$, for which the quantity $|p_k(\lambda)|$ is small at all the points near α.

I.9.3 Preconditioning

In the case of sequences of structured linear systems of the form

$$\{A_n x_n = b_n\}_n, \qquad A_n \in \mathbb{C}^{d_n \times d_n}, \qquad b_n \in \mathbb{C}^{d_n} \tag{I.24}$$

stemming from the discretization of linear PDEs, it often happens that the condition number of the system matrix A_n diverges to infinity as n increases. In many cases, this property is in

contrast with the desired cluster of the spectrum of A_n around a single point that we cited in the previous subsection. In order to accelerate the convergence, we can use the technique of preconditioning.

The theoretical idea behind preconditioning consists in the substitution of the systems (I.24) with the preconditioned ones

$$\{P_n^{-1}A_n x_n = P_n^{-1}b_n\}_n, \tag{I.25}$$

where $P_n \in \mathbb{C}^{d_n \times d_n}$ is a HPD matrix. Since the goal of preconditioning is to accelerate the convergence of the iterative method, it is evident that the operation of solving a system with matrix P_n should not be as computationally expensive as the solution of the initial system.

Summarizing the considerations that we made in the previous paragraphs, we can say that, ideally, the preconditioner P_n should satisfy the following two requirements:

a) for all $c_n \in \mathbb{C}^{d_n}$ the solution of the system $P_n y_n = c_n$ has computational cost proportional to that of the matrix-vector product with matrix A_n;

b) either $\kappa(P_n^{-1}A_n)$ is bounded from above by a constant independent of n or $\{P_n^{-1}A_n - I_n\}_n$ is strongly clustered at 0.

Trivially, the option $P_n = A_n$ satisfies condition b), but in general it definitely fails to satisfy the first requirement and of course it is not a sensible choice. The development of an efficient preconditioner should well-balance the two conditions with the construction of a matrix P_n "close" to A_n, but not as computationally costly to invert.

In the Toeplitz setting, many satisfactory solutions have been studied (see [28, 99, 114] and references therein). One possibility is to look for a preconditioner with a circulant structure. This choice automatically satisfies requirement a), since the computational cost of the solution of a linear system with a (multilevel block) circulant coefficient matrix is proportional the cost of the matrix-vector product with a (multilevel block) Toeplitz matrix. Indeed, both operations can be performed by using only few FFTs, as we recalled in Section I.6.

In what follows, we report two different strategies for circulant preconditioning in the scalar Toeplitz setting that are efficient under specific assumptions on the generating function:

- the Strang preconditioner for $T_n[f]$ is the circulant matrix $S_n \in \mathbb{C}^{n\times n}$ having the vector $[s_0, s_{-1}, \ldots, s_{-n+1}]$ as the first row and the vector $[s_0, s_1, \ldots, s_{n-1}]^T$ as the first column, which are defined by the formula

$$s_i = \begin{cases} \hat{f}_i, & 0 \le i \le \lfloor \frac{n}{2} \rfloor; \\ \hat{f}_{i-n}, & \lfloor \frac{n}{2} \rfloor < i < n; \\ s_{n+i}, & -n < i < 0. \end{cases}$$

  See [49, 116] for optimality results in the case where f belongs to the Dini–Lipschitz class.

- The Frobenius optimal preconditioner for $T_n[f]$ is the circulant matrix $\tilde{C}_n \in \mathbb{C}^{n\times n}$ that minimizes the Frobenius norm of $T_n[f] - C_n$, where C_n ranges over the set of circulant matrices, that is,

$$\tilde{C}_n = \underset{C_n \text{ circulant}}{\arg\min} \; |||T_n[f] - C_n|||_2.$$

If f is a positive continuous function, then the Frobenius optimal preconditioner is a suitable choice, see [30, 114, 116].

I.9.4 Multigrid Methods

Multigrid methods (MGM) are iterative procedures that aim at solving efficiently a linear system of large size by creating a proper sequence of linear systems of decreasing dimensions obtained by consecutive projections. In the current subsection, we present the Two-Grid Method (TGM) and the V-cycle method.

Let $A_n \in \mathbb{C}^{n\times n}$, and x_n, $b_n \in \mathbb{C}^n$. Let $P_{n,m} \in \mathbb{C}^{n\times m}$, $m < n$, be a full-rank matrix and let us consider two stationary iterative methods: the method $\mathcal{V}_{n,\text{pre}}$, with iteration matrix $V_{n,\text{pre}}$, and $\mathcal{V}_{n,\text{post}}$, with iteration matrix $V_{n,\text{post}}$.

Given an initial guess $x_n^{(0)} \in \mathbb{C}^n$, an iteration of a TGM is given by the following steps:

$$x_n^{(k+1)} = \mathcal{TGM}(A_n, x_n^{(k)}, b_n)$$

0. $x_n^{\text{pre}} = \mathcal{V}_{n,\text{pre}}^{\nu_{\text{pre}}}(A_n, b_n, x_n^{(k)})$	Pre-smoothing iterations
1. $r_n = b_n - A_n x_n^{\text{pre}}$ 2. $r_m = P_{n,m}^H r_n$ 3. $A_m = P_{n,m}^H A_n P_{n,m}$ 4. Solve $A_m y_m = r_m$ 5. $\hat{x}_n = x_n^{\text{pre}} + P_{n,m} y_m$	Coarse Grid Correction (CGC)
6. $x_n^{(k+1)} = \mathcal{V}_{n,\text{post}}^{\nu_{\text{post}}}(A_n, b_n, \hat{x}_n)$	Post-smoothing iterations

Steps 1. $\to$ 5. define the Coarse Grid Correction (CGC) that depends on the grid transfer operator $P_{n,m}$, while step 0. and step 6. consist, respectively, in applying ν_{pre} times a pre-smoother and ν_{post} times a post-smoother of the given iterative methods. Step 3. defines the coarser matrix A_m according to the Galerkin approach, which ensures that the CGC is an algebraic projector, that is, the matrix

$$\text{CGC}(A_n, P_{n,m}) = \left[I_n - P_{n,m}\left(P_{n,m}^H A_n P_{n,m}\right)^{-1} P_{n,m}^H A_n\right]$$

is such that $\text{CGC}(A_n, P_{n,m})^2 = \text{CGC}(A_n, P_{n,m})$. An algebraic projector has eigenvalues 0 and 1, which means that a stationary method with iteration matrix $\text{CGC}(A_n, P_{n,m})$ does not converge (see Theorem I.9.1) and this consideration highlights the crucial role played by the smoothers. Combining smoothing steps and CGC, the TGM is a stationary method defined by the following matrix

$$\text{TGM}(A_n, V_{n,\text{pre}}^{\nu_{\text{pre}}}, V_{n,\text{post}}^{\nu_{\text{post}}}, P_{n,m}) = V_{n,\text{post}}^{\nu_{\text{post}}}\left[I_n - P_{n,m}\left(P_{n,m}^H A_n P_{n,m}\right)^{-1} P_{n,m}^H A_n\right] V_{n,\text{pre}}^{\nu_{\text{pre}}}.$$

For the convergence analysis of structured matrices, the results are based on the Ruge-Stüben theory [107] for TGM. In particular, we report a fundamental theorem on the TGM convergence, whose proof and details are contained in [107, Theorem 5.2] and [4, Remark 2.2].

Theorem I.9.2. *Assume that the pre-smoothing step 0. is not present. Let A_n be a positive definite matrix of size n and let $V_{n,\text{post}}$ be defined as in the* TGM *algorithm. Assume*

(a) $\exists \alpha_{\text{post}} > 0 \,:\, \|V_{n,\text{post}} x_n\|_{A_n}^2 \leq \|x_n\|_{A_n}^2 - \alpha_{\text{post}} \|x_n\|_{A_n^2}^2, \qquad \forall x_n \in \mathbb{C}^n,$

(b) $\exists \gamma > 0 \,:\, \min_{y \in \mathbb{C}^m} \|x_n - P_{n,m} y\|_2^2 \leq \gamma \|x_n\|_{A_n}^2, \qquad \forall x_n \in \mathbb{C}^n.$

Then $\gamma \geq \alpha_{\text{post}}$ *and*

$$\|\text{TGM}(A_n, I, V_{n,\text{post}}^{\nu_{\text{post}}}, P_{n,m})\|_{A_n} \leq \sqrt{1 - \alpha_{\text{post}}/\gamma}.$$

Conditions (a) and (b) are usually called "smoothing property" and "approximation property", respectively.

Since α_{post} and γ are independent of n, if the assumptions of Theorem I.9.2 are satisfied, then the resulting TGM also has an optimal rate of convergence. In other words, the number of iterations in order to reach a given accuracy ε can be bounded from above by a constant independent of n (possibly depending on the parameter ε).

The computational flaw of the TGM algorithm is the exact solution of the error equation required by step 4., an operation that can be extremely expensive if the system matrix is of large size. The V-cycle method remedies this fault by consecutively restricting the problem until it is so small that the error equation can be easily solved.

Indeed, the standard V-cycle method is obtained replacing the direct solution at step 4. with a recursive call of the TGM applied to the coarser linear system $A_{m_\ell} y_{m_\ell} = r_{m_\ell}$, where ℓ represents the level. The recursion is usually stopped at level $\overline{\ell}$ when $m_{\overline{\ell}}$ becomes small enough for solving cheaply step 4. with a direct solver.

Chapter II

Asymptotic Spectral Distributions of Symmetrized Toeplitz Sequences

In the present chapter we analyse the spectral features of the symmetrization of Toeplitz matrices of the form $T_n[f]$, generated by a function $f \in L^1([-\pi,\pi])$ defined on $[-\pi,\pi]$ and periodically extended to the whole real line. In particular, we consider the case where the Fourier coefficients of f are real, hence, from the definition in I.5, the corresponding $T_n[f]$ is real. The object of our investigation is the real matrix $Y_nT_n[f]$ obtained pre-multiplying $T_n[f]$ by the anti-identity matrix $Y_n \in \mathbb{R}^{n\times n}$ defined as

$$Y_n = \begin{bmatrix} & & 1 \\ & \iddots & \\ 1 & & \end{bmatrix}.$$

Note that $Y_nT_n[f]$ is a Hankel matrix, that is, it is a matrix with constant elements along the skew-diagonals, and hence it is symmetric.

The matrix Y_n is a unitary matrix and, by the definition and properties of the singular value decomposition [80], this implies that $T_n[f]$ and $Y_nT_n[f]$ possess the same singular values. Conversely, the presented one-sided symmetrization strategy produces significant changes in the eigenvalues. Think for instance of the Toeplitz matrix

$$T_n\left[\mathrm{e}^{-\hat{\imath}\vartheta}\right] = \begin{bmatrix} 0 & 1 & & \\ & \ddots & \ddots & \\ & & 0 & 1 \\ & & & 0 \end{bmatrix},$$

which is real non-symmetric and has all the eigenvalues equal to 0. The symmetrized version

$$Y_nT_n\left[\mathrm{e}^{-\hat{\imath}\vartheta}\right] = \begin{bmatrix} & & & 0 \\ & & 0 & 1 \\ & \iddots & \iddots & \\ 0 & 1 & & \end{bmatrix},$$

is instead a diagonalizable matrix with rank $n-1$ and hence its null eigenvalue has multiplicity only 1. It is straightforward to see that all the other eigenvalues are equal to 1 and -1 with

roughly the same multiplicity. Notice that the generating function $\mathrm{e}^{-\hat{\imath}\vartheta}$ is in no way related to the spectrum of the non-symmetric Toeplitz matrix $T_n\left[\mathrm{e}^{-\hat{\imath}\vartheta}\right]$, while the eigenvalues of $Y_nT_n\left[\mathrm{e}^{-\hat{\imath}\vartheta}\right]$ are approximatively half described by the modulus of $\mathrm{e}^{-\hat{\imath}\vartheta}$, which is identically equal to 1, and approximatively half described by the opposite of the modulus of $\mathrm{e}^{-\hat{\imath}\vartheta}$.

In this chapter, we formalize the latter considerations providing the spectral distribution of a matrix-sequence of the form $\{Y_nT_n[f]\}_n$ in the case where $f \in L^1([-\pi,\pi])$ has real Fourier coefficients [53]. The basic structure of the spectral symbol of $\{Y_nT_n[f]\}_n$ is intuitively clear both from the example above and from the results shown in [77], where the authors prove that roughly half of the eigenvalues of $Y_nT_n[f]$ are negative and roughly half of the eigenvalues of $Y_nT_n[f]$ are positive, when the dimension n of the matrix is sufficiently large and f is sparsely vanishing. In the main results of the chapter, Theorem II.1.2 and Corollary II.1.2.1, we prove that $\{Y_nT_n[f]\}_n$ is distributed as $\phi_{|f|}$ in the eigenvalue sense, where we define

$$\phi_{|f|}(\vartheta) = \begin{cases} |f(\vartheta)|, & \vartheta \in [0,2\pi], \\ -|f(-\vartheta)|, & \vartheta \in [-2\pi,0) \end{cases},$$

and this informally means that roughly half of the eigenvalues of $Y_nT_n[f]$ are positive and they are approximated by a uniform sampling of $|f|$ and roughly half of the eigenvalues are negative and they are approximated by a uniform sampling of $-|f|$.

As we saw in Subsection I.9.2, symmetry is a particularly desirable property for a matrix when we want to solve an associated linear system with an iterative method. Indeed, if the matrix is symmetric a method such as the MINRES can be employed and it is possible to study a priori the convergence rates of the algorithm if the eigenvalues are known. The symmetrization procedure that we analyse in this chapter was introduced by Pestana and Wathen [104] for the very purpose of developing a competitive method for the solution of real non-symmetric Toeplitz systems. Namely, they introduced an absolute value circulant preconditioner $|C_n|$ and showed, under certain assumptions, that the preconditioned matrix $|C_n|^{-1}Y_nT_n[f]$ can be decomposed into the sum of an involutory matrix, a low rank matrix, and a small norm matrix. Due to the observed clustered spectra around ± 1 of $|C_n|^{-1}Y_nT_n[f]$, rapid convergence of Krylov subspace methods such as MINRES can be expected. Exploiting the spectral distribution results on $\{Y_nT_n[f]\}_n$, in Theorem II.2.1 we prove, under analogous assumptions, that the preconditioned matrix-sequence $\{|C_n|^{-1}Y_nT_n[f]\}_n$ has spectral symbol ϕ_1, and this result permits us to analyse in detail the efficiency of classes of circulant preconditioners $|C_n|$ obtained from the relevant literature, such as the Strang preconditioner and the Frobenius optimal preconditioner.

The findings presented in the following sections are published in [53]. In the following we highlight the main results section by section. In Section II.1 we first give a distribution result regarding the eigenvalues of special 2-by-2 block matrix-sequences, whose generality goes beyond the specific case under consideration. Moreover, we report the main results on the asymptotic distributions of $\{Y_nT_n[f]\}_n$, both in the scalar and in the block-Toeplitz case, see Theorems II.1.2–II.1.3. In Section II.2, we provide the eigenvalue distribution of the preconditioned matrix-sequences $\{|C_n|^{-1}Y_nT_n[f]\}_n$ under specific assumptions on the circulant preconditioner $|C_n|$. Finally, in Sections II.3–II.4 we provide and critically discuss a selection of numerical experiments concerning different Toeplitz matrices $T_n[f]$ and the corresponding circulant preconditioners.

It is worth noting that analogous results have been obtained independently in [93], making use of the powerful *-algebra structure of the GLT sequences that we introduced in Section

I.7. Conversely, our analysis is based on the notion of approximating class of sequences that we defined in Subsection I.4.2 and that constitutes one of the prerequisites for the GLT theory, as can be seen in [62].

II.1 Spectral Results on $\{Y_nT_n[f]\}_n$

We open the present section fixing the notation for a class of functions with a particular structure, which are used throughout the current and the next chapters.

Given $D \subset \mathbb{R}^k$ with $0 < \mu_k(D) < \infty$, we define $\tilde{D}$ as $D \bigcup D_p$, where $p \in \mathbb{R}^k$ and $D_p = p + D$, with the constraint that D and D_p have non-intersecting interior part, that is $D^\circ \bigcap D_p^\circ = \emptyset$. In this way $\mu_k(\tilde{D}) = 2\mu_k(D)$. Given any function $g : D \to \mathbb{C}^{s\times s}$, we define $\psi_{g,p} \equiv \psi_g$ over $\tilde{D}$ in the following manner

$$\psi_{g,p} \equiv \psi_g(x) = \begin{cases} g(x), & x \in D, \\ -g(x-p), & x \in D_p,\ x \notin D. \end{cases} \tag{II.1}$$

The following theorem is of wide interest when dealing with special 2×2 block matrix-sequences. Even though the result is quite intuitive if we consider the relation between the eigenvalues of the block matrix

$$\begin{bmatrix} O & A \\ A^H & O \end{bmatrix}$$

and the singular values of the matrix $A \in \mathbb{C}^{m\times m}$, for clarity we provide a highly detailed proof and, moreover, we treat a general case where the blocks are rectangular matrices.

Theorem II.1.1. *Suppose $k_n = o(n)$ with $k_n \in \mathbb{Z}$ and $A(n) \in \mathbb{C}^{(\lceil n/2\rceil+k_n)\times(\lfloor n/2\rfloor-k_n)}$. Let $B_n, E_n \in \mathbb{C}^{n\times n}$ be Hermitian matrices such that*

$$B_n = \begin{bmatrix} O_{\lceil n/2\rceil+k_n,\lceil n/2\rceil+k_n} & A(n) \\ A(n)^H & O_{\lfloor n/2\rfloor-k_n,\lfloor n/2\rfloor-k_n} \end{bmatrix} + E_n.$$

If $\{A(n)\}_n \sim_\sigma g$, where $g : D \to \mathbb{C}$ is a non-negative function defined over a measurable set D with positive, finite Lebesgue measure, and $\{E_n\}_n \sim_\sigma 0$, then

$$\{B_n\}_n \sim_\lambda \psi_g$$

over the domain $\tilde{D}$, with ψ_g as in (II.1).

Proof. For the sake of notational simplicity, we set $A = A(n)$ and we define the auxiliary matrix G_n as follows

$$G_n = \begin{bmatrix} O_{\lceil n/2\rceil+k_n,\lceil n/2\rceil+k_n} & A \\ A^H & O_{\lfloor n/2\rfloor-k_n,\lfloor n/2\rfloor-k_n} \end{bmatrix}.$$

Fixing n and supposing $k_n \geq 0$, we define $m = \lfloor n/2\rfloor - k_n$ and $M = \lceil n/2\rceil + k_n$. Then, we consider the (full) singular value decomposition of $A = U_M \Sigma V_m^H$, where U_M, V_m are unitary matrices of size M and m, respectively, and Σ is the rectangular diagonal matrix containing the singular values $\sigma_1, \ldots, \sigma_m$. We have

$$G_n = \begin{bmatrix} U_M & O_{M,m} \\ O_{m,M} & V_m \end{bmatrix} \begin{bmatrix} O_{M,M} & \Sigma \\ \Sigma^T & O_{m,m} \end{bmatrix} \begin{bmatrix} U_M^H & O_{M,m} \\ O_{m,M} & V_m^H \end{bmatrix} \tag{II.2}$$

which is similar to

$$S_n = \begin{bmatrix} O_{M,M} & \Sigma \\ \Sigma^T & O_{m,m} \end{bmatrix}.$$

Notice that the matrix Σ can be written as

$$\Sigma = \begin{bmatrix} \tilde{\Sigma}_m \\ O_{k,m} \end{bmatrix}, \qquad \tilde{\Sigma}_m = \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_m \end{bmatrix}, \qquad k = M - m, \tag{II.3}$$

where $\Sigma = \tilde{\Sigma}_m$ if $k = 0$. Under the hypothesis that $k_n \geq 0$, if the fixed n is even, the index k is equal to $2k_n$. Otherwise, it is equal to $2k_n + 1$.

Using (II.3), the matrix S_n can be written as

$$S_n = \begin{bmatrix} O_{M,M} & \Sigma \\ \Sigma^T & O_{m,m} \end{bmatrix} = \begin{bmatrix} O_{m,m} & O_{m,k} & \tilde{\Sigma}_m \\ O_{k,m} & O_{k,k} & O_{k,m} \\ \tilde{\Sigma}_m & O_{m,k} & O_{m,m} \end{bmatrix},$$

where, if $k = 0$, the central row and column are not present and which, up to similarity by an obvious permutation, can be written as the direct sum of $O_{k,k}$ and

$$\begin{bmatrix} O_{m,m} & \tilde{\Sigma}_m \\ \tilde{\Sigma}_m & O_{m,m} \end{bmatrix}.$$

The latter matrix is 2×2 block circulant and hence can be diagonalized by the 2×2 block Fourier matrix so that

$$\begin{bmatrix} O_{m,m} & \tilde{\Sigma}_m \\ \tilde{\Sigma}_m & O_{m,m} \end{bmatrix} = \frac{\sqrt{2}}{2} \begin{bmatrix} I_m & I_m \\ I_m & -I_m \end{bmatrix} \begin{bmatrix} \tilde{\Sigma}_m & O_{m,m} \\ O_{m,m} & -\tilde{\Sigma}_m \end{bmatrix} \frac{\sqrt{2}}{2} \begin{bmatrix} I_m & I_m \\ I_m & -I_m \end{bmatrix}.$$

Therefore, putting together the above information, we can write the factorization

$$S_n = \begin{bmatrix} O_{m,m} & O_{m,k} & \tilde{\Sigma}_m \\ O_{k,m} & O_{k,k} & O_{k,m} \\ \tilde{\Sigma}_m & O_{m,k} & O_{m,m} \end{bmatrix} = Q_n \begin{bmatrix} \tilde{\Sigma}_m & O_{m,k} & O_{m,m} \\ O_{k,m} & O_{k,k} & O_{k,m} \\ O_{m,m} & O_{m,k} & -\tilde{\Sigma}_m \end{bmatrix} Q_n,$$

where Q_n is the orthogonal matrix

$$Q_n = \frac{\sqrt{2}}{2} \begin{bmatrix} I_m & O_{m,k} & I_m \\ O_{k,m} & \sqrt{2} I_k & O_{k,m} \\ I_m & O_{m,k} & -I_m \end{bmatrix}$$

given by the direct sum of the identity of size k and of the previous 2×2 block Fourier matrix. Thus, we know that G_n is similar to the block diagonal matrix

$$\begin{bmatrix} \tilde{\Sigma}_m & O_{m,k} & O_{m,m} \\ O_{k,m} & O_{k,k} & O_{k,m} \\ O_{m,m} & O_{m,k} & -\tilde{\Sigma}_m \end{bmatrix}. \tag{II.4}$$

Hence (II.4) implies that we can write the eigenvalues of the matrix G_n for the case $k_n \geq 0$. A similar factorization can be obtained for $k_n < 0$, by defining $m = \lceil n/2 \rceil + k_n$ and $M = \lfloor n/2 \rfloor - k_n$.

In particular, the eigenvalues of G_n are given by the set of the singular values of A_n, the set of the negation of the singular values of A_n and, in addition to these, at most $k = o(n)$ zero eigenvalues. From the latter, it is transparent that

$$\{G_n\}_n \sim_\lambda \psi_g.$$

Finally, since all the involved matrices are Hermitian and the perturbation matrix-sequence is zero distributed, i.e, $\{E_n\}_n \sim_{\lambda,\sigma} 0$, the desired result follows directly from the second part of Lemma I.4.2, taking into account that $\{\{G_n\}_n\}_m$ is a constant class of sequences (that is not depending on the variable m) and it is nevertheless an a.c.s for $\{B_n\}_n$. □

Employing Theorem II.1.1, we now prove the following central result on the spectral distribution of symmetrized Toeplitz sequences.

Theorem II.1.2. *Suppose $f \in L^1([-\pi,\pi])$ with real Fourier coefficients and let $Y_n \in \mathbb{R}^{n\times n}$ be the anti-identity matrix. Let $T_n[f] \in \mathbb{R}^{n\times n}$ be the Toeplitz matrix generated by f. Then*

$$\{Y_nT_n[f]\}_n \sim_\lambda \psi_{|f|}$$

with $\psi_{|f|}$ defined as in (II.1) over the domain $\tilde{D}$ with $D = [0, 2\pi]$ and $p = -2\pi$.

Proof. We let $H_\nu[f,-]$ be the ν-by-ν Hankel matrix generated by f containing the Fourier coefficients from $\hat{f}_{-1}$ in position $(1,1)$ to $\hat{f}_{-2\nu+1}$ in position (ν,ν). Analogously, we let $H_\nu[f,+]$ be the ν-by-ν Hankel matrix generated by f containing the Fourier coefficients from $\hat{f}_1$ in position $(1,1)$ to $\hat{f}_{2\nu-1}$ in position (ν,ν).

We start by considering the case of even n and writing $Y_nT_n[f]$ as a 2-by-2 block matrix of size $n = 2\nu$, i.e.

$$Y_nT_n[f] = \begin{bmatrix} Y_\nu H_\nu[f,+]Y_\nu & Y_\nu T_\nu[f] \\ Y_\nu T_\nu[f] & H_\nu[f,-] \end{bmatrix}.$$

Note that for Lebesgue integrable f, $H_\nu[f,+]$ is exactly the Hankel matrix generated by f according to the definition given in [51]: in that paper it was proved that $\{H_\nu[f,+]\}_n \sim_\sigma 0$. Since in our setting $H_\nu[f,+]$ is symmetric for every ν, it follows that $\{H_\nu[f,+]\}_n \sim_\lambda 0$. Hence, with Y_ν being both symmetric and orthogonal, we deduce that the matrix is symmetric with the same singular values as $H_\nu[f,+]$. Therefore

$$\{Y_\nu H_\nu[f,+]Y_\nu\}_n \sim_{\lambda,\sigma} 0.$$

Similarly, we have

$$\{H_\nu[f,-]\}_n \sim_{\lambda,\sigma} 0$$

since $H_\nu[f,-] = H_\nu[\bar{f},+]$ and $\bar{f}$ (being the conjugate of f) is Lebesgue integrable if and only if f is Lebesgue integrable.

Therefore, the matrix-sequence $\{Y_nT_n[f]\}_n$ can be written as the sum of the matrix-sequence whose eigenvalues are clustered at zero

$$\{E_n\}_n = \left\{ \begin{bmatrix} Y_\nu H_\nu[f,+]Y_\nu & O_{\nu,\nu} \\ O_{\nu,\nu} & H_\nu[f,-] \end{bmatrix} \right\}_n$$

and the matrix-sequence

$$\left\{ \begin{bmatrix} O_{\nu,\nu} & Y_\nu T_\nu[f] \\ Y_\nu T_\nu[f] & O_{\nu,\nu} \end{bmatrix} \right\}_n$$

whose eigenvalues are $\pm\sigma_j(Y_\nu T_\nu[f]) = \pm\sigma_j(T_\nu[f])$, $j = 1, \ldots, \nu$.

Hence, the claimed thesis follows from Theorem II.1.1 with $g = |f|$, $A = A^H = A^T = Y_\nu T_\nu[f]$, and $k_n = 0$.

In the case where n is odd, the analysis is of the same type as before with a few slight technical changes.

By setting $\nu = \lfloor n/2 \rfloor$, $\mu = \lceil n/2 \rceil$, $v = \left[\hat{f}_\nu, \ldots, \hat{f}_1\right]^T$, and $w = \left[\hat{f}_{-1}, \ldots, \hat{f}_{-\nu}\right]^T$ we have

$$Y_n T_n[f] = \begin{bmatrix} Y_\nu H_\nu[f \cdot \mathrm{e}^{-\hat{\imath}\vartheta}, +] Y_\nu & v & Y_\nu T_\nu[f] \\ v^T & \hat{f}_0 & w^T \\ Y_\nu T_\nu[f] & w & H_\nu[f \cdot \mathrm{e}^{\hat{\imath}\vartheta}, -] \end{bmatrix}, \tag{II.5}$$

provided that we exclude the trivial case $n = 1$. Let us consider the matrices

$$E'_n = \begin{bmatrix} Y_\mu H_\mu[f \cdot \mathrm{e}^{\hat{\imath}\vartheta}, +] Y_\mu & O_{\mu,\nu} \\ O_{\nu,\mu} & H_\nu[f \cdot \mathrm{e}^{\hat{\imath}\vartheta}, -] \end{bmatrix},$$

$$Y_\mu H_\mu[f \cdot \mathrm{e}^{\hat{\imath}\vartheta}, +] Y_\mu = \begin{bmatrix} Y_\nu H_\nu[f \cdot \mathrm{e}^{-\hat{\imath}\vartheta}, +] Y_\nu & v \\ v^T & \hat{f}_0 \end{bmatrix},$$

$$E''_n = \begin{bmatrix} O_{\nu,\nu} & \mathbf{o}_\nu & O_{\nu,\nu} \\ \mathbf{o}_\nu{}^T & 0 & w^T \\ O_{\nu,\nu} & w & O_{\nu,\nu} \end{bmatrix},$$

and define $E_n = E'_n + E''_n$. From (II.5), it is evident that the matrix-sequence $\{Y_n T_n[f]\}_n$ can be written as the sum of the matrix-sequence $\{E_n\}_n$, whose eigenvalues are clustered at zero, and the matrix-sequence

$$\left\{ \begin{bmatrix} O_{\nu,\nu} & \mathbf{o}_\nu & Y_\nu T_\nu[f] \\ \mathbf{o}_\nu{}^T & 0 & \mathbf{o}_\nu{}^T \\ Y_\nu T_\nu[f] & \mathbf{o}_\nu & O_{\nu,\nu} \end{bmatrix} \right\}_n$$

whose eigenvalues are 0 with multiplicity 1 and $\pm\sigma_j(Y_\nu T_\nu[f])$, $j = 1, \ldots, \nu$. Note that the unitary nature of Y_ν implies again that $\sigma_j(Y_\nu T_\nu[f]) = \sigma_j(T_\nu[f])$, $j = 1, \ldots, \nu$.

Consequently, the claimed thesis follows from Theorem II.1.1 with $g = |f|$,

$$A = A(n) = \begin{bmatrix} Y_\nu T_\nu[f] \\ \mathbf{o}_\nu^T \end{bmatrix}, \qquad A^H = A(n)^H = A(n)^T = \begin{bmatrix} Y_\nu T_\nu[f] & \mathbf{o}_\nu \end{bmatrix},$$

and $k_n = 0$. □

The following corollary provides a different spectral symbol for the matrix-sequence $\{Y_n T_n[f]\}_n$, obtained by "rearranging" the function $\psi_{|f|}$. Indeed, the concept of rearrangement has a precise technical meaning and a discussion on the topic can be found in [62, Section 3.2].

Corollary II.1.2.1. *Suppose $f \in L^1([-\pi,\pi])$ with real Fourier coefficients and $Y_n \in \mathbb{R}^{n\times n}$ is the anti-identity matrix. Let $T_n[f] \in \mathbb{R}^{n\times n}$ be the Toeplitz matrix generated by f. Then,*

$$\{Y_n T_n[f]\}_n \sim_\lambda \phi_{|f|}$$

over the domain $[-2\pi, 2\pi]$ with ϕ_g defined in the following way

$$\phi_g(\vartheta) = \begin{cases} g(\vartheta), & \vartheta \in [0, 2\pi], \\ -g(-\vartheta), & \vartheta \in [-2\pi, 0). \end{cases}$$

Proof. We observe that $\phi_{|f|}$ is a rearrangement of $\psi_{|f|}$, that is, for all F continuous with bounded support we have

$$\int_{-2\pi}^{2\pi} F\left(\phi_{|f|}(\vartheta)\right)\, d\vartheta = \int_{-2\pi}^{2\pi} F\left(\psi_{|f|}(\vartheta)\right)\, d\vartheta.$$

Hence, by the very definition of spectral distribution, we have $\{Y_n T_n[f]\}_n \sim_\lambda \phi_{|f|}$ if and only if $\{Y_n T_n[f]\}_n \sim_\lambda \psi_{|f|}$. Therefore, the desired result is an immediate consequence of Theorem II.1.2. □

Considering a real-valued generating function f, we remark that the spectral distribution of $\{Y_n T_n[f]\}_n$ is in stark contrast to that of $\{T_n[f]\}_n$ provided by the generalized Szegő theorem (Theorem I.5.3), even though their singular value distributions are equivalent.

Finally, the techniques given in this section can be adapted verbatim to the case of Toeplitz structures generated by $s \times s$ matrix-valued functions, namely, the following theorem holds.

Theorem II.1.3. *Suppose that $\mathbf{f} \in L^1([-\pi,\pi], s)$ is an $s \times s$ matrix-valued function defined on $[-\pi,\pi]$. Let $T_n[\mathbf{f}] \in \mathbb{C}^{sn\times sn}$ be the block-Toeplitz matrix generated by $\mathbf{f}$. Then*

$$\{(Y_n \otimes I_s) T_n[\mathbf{f}]\}_n \sim_\lambda \psi_{|\mathbf{f}|}, \qquad |\mathbf{f}| = (\mathbf{f}\,\mathbf{f}^H)^{1/2},$$

over the domain $\tilde{D}$ with $D = [0, 2\pi]$ and $p = -2\pi$, where $\psi_{|\mathbf{f}|}$ is defined in (II.1).

II.2 Spectral Results on Preconditioned Matrix-Sequences

In the following theorems we use the results of the previous subsection in order to deal with the eigenvalue distribution of certain preconditioned matrix-sequences. In particular, we investigate the assumptions on the matrix $C_n = F_n \Lambda_n F_n^H$, where Λ_n is a diagonal matrix, such that its absolute value $|C_n|$ defined by

$$\begin{aligned} |C_n| &= (C_n^H C_n)^{1/2} \\ &= (C_n C_n^H)^{1/2} \\ &= F_n |\Lambda_n| F_n^H, \end{aligned} \tag{II.6}$$

provides a weak cluster to ± 1 of the eigenvalues of the preconditioned matrix-sequence $\{|C_n|^{-1} Y_n T_n[f]\}_n$.

Theorem II.2.1. *Suppose $f \in L^1([-\pi,\pi])$ with real Fourier coefficients and let $Y_n \in \mathbb{R}^{n\times n}$ be the anti-identity matrix. Let $T_n[f] \in \mathbb{R}^{n\times n}$ be the Toeplitz matrix generated by f. Then*

$$\{|C_n|^{-1}Y_nT_n[f]\}_n \sim_\lambda \psi_1 = \phi_1$$

over the domain $\tilde{D}$ with $D = [0, 2\pi]$ and $p = -2\pi$ under the assumption that $\{C_n\}_n$ is a circulant matrix-sequence of invertible matrices such that

$$\{C_n^{-1}T_n[f]\}_n \sim_\sigma 1.$$

Proof. For C_n being non-singular, the matrix $|C_n|$ is symmetric positive definite and, hence, the matrices

$$|C_n|^{-1}Y_nT_n[f] \quad \text{and} \quad |C_n|^{-1/2}Y_nT_n[f]|C_n|^{-1/2}$$

are well defined and similar. They share the same eigenvalues clustered around $\{-1, 1\}$ by [104], under the assumption that $\{C_n^{-1}T_n[f]\}_n$ is clustered around 1 in the singular value sense. Also, by the Sylvester inertia law, the matrices

$$|C_n|^{-1/2}Y_nT_n[f]|C_n|^{-1/2} \quad \text{and} \quad Y_nT_n[f]$$

have exactly the same inertia, namely the same number of positive, negative, and zero eigenvalues. Also, by [77, Theorem 4.1], we know that the matrix $Y_nT_n[f]$ has $n/2 + o(n)$ positive eigenvalues, $n/2+o(n)$ negative eigenvalues, and $o(n)$ zero eigenvalues for large enough n. Therefore, by combining the above statements, we deduce that the matrix $|C_n|^{-1}Y_nT_n[f]$ possesses $n/2 + o(n)$ eigenvalues clustered around 1 and $n/2 + o(n)$ eigenvalues clustered around -1.

A simple check shows that the latter statement is equivalent to writing

$$\{|C_n|^{-1}Y_nT_n[f]\}_n \sim_\lambda \psi_1 = \phi_1$$

over the domain $\tilde{D}$ with $D = [0, 2\pi]$ and $p = -2\pi$. □

We now complement the previous theorem with a short discussion regarding the hypothesis $\{C_n^{-1}T_n[f]\}_n \sim_\sigma 1$. Note that we can extend the result to the case where C_n is not necessarily invertible. For this purpose, we denote by $C_n^\dagger$ the pseudo-inverse of a circulant matrix C_n, which is obtained by taking the singular value decomposition of C_n and replacing every non-zero singular value by its reciprocal. If we consider $C_n^\dagger$ instead of C_n^{-1}, the assumption that f is sparsely vanishing implies the presence of at most $o(n)$ zero eigenvalues in both the matrix C_n and the preconditioned matrix $C_n^\dagger T_n[f]$. Recalling the definitions in Subsection I.9.3 and the analysis in [49, 117], we have the following picture.

A) When C_n is the Strang preconditioner for $T_n[f]$, the key assumption $\{C_n^\dagger T_n[f]\}_n \sim_\sigma 1$ holds if f is sparsely vanishing and belongs to the Dini-Lipschitz class (see for example [49, Item 2, Proposition 2.1]) which is a proper subset of the continuous 2π-periodic functions.

B) When C_n is the Frobenius optimal preconditioner for $T_n[f]$, the key assumption $\{C_n^\dagger T_n[f]\}_n \sim_\sigma 1$ holds if f is sparsely vanishing and simply Lebesgue integrable (such a general result was proved quite elegantly by combining the Korovkin theory [117] and the GLT analysis in [62]).

We summarize the interplay among Theorem II.2.1 and Items **A** and **B** in the following general result.

Theorem II.2.2. *Suppose $f \in L^1([-\pi,\pi])$ with real Fourier coefficients and let $Y_n \in \mathbb{R}^{n\times n}$ be the anti-identity matrix. Let $T_n[f] \in \mathbb{R}^{n\times n}$ be the Toeplitz matrix generated by f and assume that f is sparsely vanishing. Then*

$$\{|C_{n,*}|^{-1}Y_nT_n[f]\}_n \sim_\lambda \psi_1 = \phi_1$$

over the domain $\tilde{D}$ with $D = [0, 2\pi]$ and $p = -2\pi$, under the assumption that either

α) *f belongs to the Dini-Lipschitz class, C_n is the Strang preconditioner, and $C_{n,*}$ is the stabilized Strang preconditioner where all the zero eigenvalues are replaced by 1 (or by any other suitable constant different from zero) or*

β) *C_n is the Frobenius optimal preconditioner and $C_{n,*}$ is the stabilized Frobenius optimal preconditioner where all the zero eigenvalues are replaced by 1 (or by any other suitable constant different from zero).*

Proof. By combining Theorem II.2.1 and the aforementioned Item **A**, we deduce that $\{C_n^\dagger T_n[f]\}_n \sim_\sigma 1$ and $\{|C_n|^\dagger Y_nT_n[f]\}_n \sim_\lambda \psi_1 = \phi_1$. Since f is sparsely vanishing, the number of zero eigenvalues of $\{C_n\}_n$ is at most $o(n)$, and both $\{C_n - C_{n,*}\}_n$ and $\{|C_n|^\dagger - |C_{n,*}|^{-1}\}_n$ are clustered around zero. Hence, the assertion under the assumption α) follows. Using the exactly same arguments with Item **B**, the assertion under assumption β) can be shown.

□

The above theorem covers the range of applicability of the preconditioned MINRES technique described in [104]. Regarding the analysis wherein, it is worth observing that the circulant matrix $\tilde{C}_n = F_n\tilde{\Lambda}_nF_n^H$, where $\tilde{\Lambda}_n$ is the diagonal matrix in the eigendecomposition of C_n with all entries divided by their module, is not involutory as claimed in [104, Eq. (3.4), P. 276]. In fact, it is simply unitary: indeed its eigenvalues have unit modulus, but in general they are not real. Hence, it is orthogonal when C_n is real.

Finally, we point out that the quality of clustering of the preconditioners in Theorems II.2.1 and II.2.2 depends on that of the standard circulant based preconditioning whose analysis is available in the relevant literature (see [99] and the references therein).

II.3 Numerical Tests on the Spectral Distribution of $\{Y_nT_n[f]\}_n$

In the current section we numerically show that the results obtained in Section II.1 are true in the cases of both trigonometric polynomials and more generic functions in $L^1([-\pi,\pi])$.

In order to numerically support Theorem II.1.2, we show that for large enough n the eigenvalues of $Y_nT_n[f]$ are approximately equal to the samples of $\psi_{|f|}$ over a uniform grid in $[-2\pi, 2\pi]$, with the possible exception of a small number of outliers. We also remark that the function $\phi_{|f|}$ in Corollary II.1.2.1 has the same property, being a rearrangement of $\psi_{|f|}$.

Surprisingly, we observe that the forecasts provided by our theorems concerning the symbols are highly accurate and go beyond the scope of our developed theory, so the corresponding investigation will be a subject for future research.

We highlight the fact that the matrix $Y_nT_n[f]$ is symmetric for any n, so the quantities $\lambda_j(Y_nT_n[f])$ are real for $j = 1, \ldots, n$. In particular we order the eigenvalues of $Y_nT_n[f]$ according to the evaluation of $\psi_{|f|}$ (respectively $\phi_{|f|}$) on the following uniform grid in $[-2\pi, 2\pi]$:

$$\vartheta_{j,n} = -2\pi + j\frac{4\pi}{n}, \qquad j = 1, \ldots, n. \tag{II.7}$$

Thus, in our experiments, we first compute the quantities $\psi_{|f|}(\vartheta_{j,n})$ and $\phi_{|f|}(\vartheta_{j,n})$ for a fixed n and then compare them with the properly sorted eigenvalues $\lambda_j(Y_nT_n[f])$, $j = 1, \ldots, n$. The quantities $\lambda_j(Y_nT_n[f])$ are computed with MATLAB's `eig` function.

In Example 2, we give numerical evidence of the fact that $\lambda_j(Y_nT_n[f])$ and $\psi_{|f|}(\vartheta_{j,n})$ are approximately equal for a real-valued, even trigonometric polynomial. In Example 3, considering a trigonometric polynomial, we compare the quantities $\lambda_j(Y_nT_n[f])$ with both $\psi_{|f|}(\vartheta_{j,n})$ and $\phi_{|f|}(\vartheta_{j,n})$, and observe that they are approximately equal with the exception of three outliers. In Example 4 we give numerical evidence of Theorem II.1.2 for a continuous function in $L^1([-\pi, \pi])$ and in Example 5 we do the same for a discontinuous piecewise constant function in $L^1([-\pi, \pi])$.

Example 2. *We consider the real-valued, even trigonometric polynomial $f : [-\pi, \pi] \to \mathbb{R}$ defined by*

$$f(\vartheta) = 2 - 12\cos(\vartheta).$$

The n-by-n Toeplitz matrix generated by f is

$$T_n[f] = \begin{bmatrix} 2 & -6 & & \\ -6 & \ddots & \ddots & \\ & \ddots & \ddots & -6 \\ & & -6 & 2 \end{bmatrix}.$$

Notice that $T_n[f]$ is banded and symmetric. The multiplication by Y_n produces the following matrix:

$$Y_nT_n[f] = \begin{bmatrix} & & -6 & 2 \\ & \iddots & \iddots & -6 \\ -6 & \iddots & \iddots & \\ 2 & -6 & & \end{bmatrix}.$$

Figure II.1 shows that the properly sorted eigenvalues of $Y_nT_n[f]$ are approximately equal to the samples of $\psi_{|f|}$ over $\vartheta_{j,n}$ for all $j = 1, \ldots, n$. The plot is made for $n = 300$. This result is expected from the statement of Theorem II.1.2 and there are no outliers in this case.

Example 3. *In this example we deal with a trigonometric polynomial $f : [-\pi, \pi] \to \mathbb{C}$, defined as*

$$f(\vartheta) = 4 + 2\mathrm{e}^{-\hat{\imath}\vartheta} + 2\mathrm{e}^{-2\hat{\imath}\vartheta} + 9\mathrm{e}^{-3\hat{\imath}\vartheta} + \mathrm{e}^{\hat{\imath}\vartheta}.$$

Hence, the function f generates a real, banded Toeplitz matrix $T_n[f]$. Differently from Example 2, the matrix $T_n[f]$ in this case is not symmetric. Nevertheless, the premultiplication by Y_n produces the symmetric matrix $Y_nT_n[f]$ with real eigenvalues $\lambda_j(Y_nT_n[f])$, $j = 1, \ldots, n$.

For this example, we compare the eigenvalues of $Y_nT_n[f]$ with the samples of $\psi_{|f|}$ in Figure II.2 and those with $\phi_{|f|}$ in Figure II.3. In both figures, we observe that the spectrum of $Y_nT_n[f]$

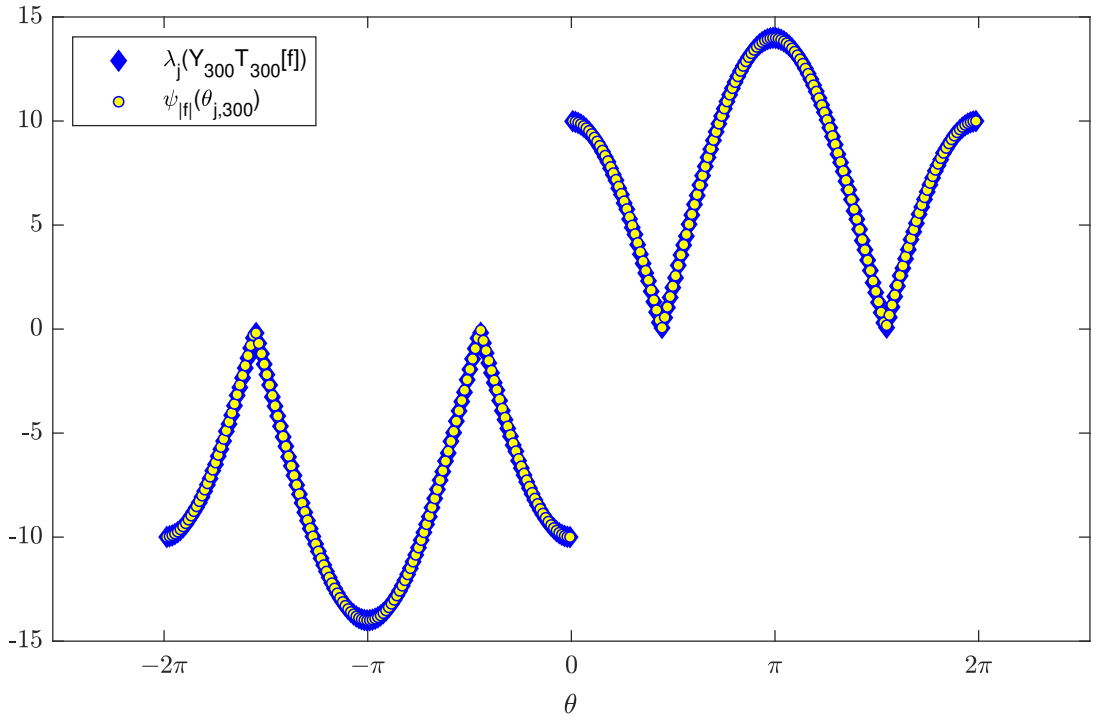

Figure II.1: Example 2, a comparison between the eigenvalues $\lambda_j(Y_n T_n[f])$ and the samples $\psi_{|f|}(\vartheta_{j,n})$, for $f(\vartheta) = 2 - 12\cos(\vartheta)$ and $n = 300$.

is well approximated by both the evaluations of $\psi_{|f|}$ and $\phi_{|f|}$, except for the presence of three outliers.

The presence of such eigenvalues, which are not captured by the sampling of $\psi_{|f|}$ and $\phi_{|f|}$, is in line with the behaviour predicted by Theorem II.1.2 and Corollary II.1.2.1. In fact, this agrees well with the concept of spectral distribution formalized in Definition I.4.3.

Example 4. *Let us consider the function $f : [-\pi, \pi] \to \mathbb{R}$ given by*

$$f(\vartheta) = \vartheta^2,$$

periodically extended to the real line.

The function f is not a trigonometric polynomial, and consequently the matrices $T_n[f]$ are dense for all n. In fact, the Fourier coefficients of f are explicitly given by the formulae

$$\begin{cases} \hat{f}_0 = \frac{\pi^2}{3}, \\ \hat{f}_k = (-1)^k \frac{2}{k^2}, \quad k = \pm 1, \pm 2, \dots . \end{cases}$$

This expression can be derived by a direct computation of the quantities

$$\hat{f}_k = \frac{1}{\pi} \int_0^\pi \vartheta^2 \cos(-k\vartheta)\, d\vartheta.$$

In this example, we set $n = 200$ and evaluate $\psi_{|f|}$ on the points of the grid $\vartheta_{j,n}$. Recalling that f is defined on $[-\pi, \pi]$ and periodically extended to the real line, we can write the following explicit formulae for f in $[0, 2\pi]$:

$$f(\vartheta) = \begin{cases} \vartheta^2, & \vartheta \in [0, \pi], \\ (\vartheta - 2\pi)^2, & \vartheta \in (\pi, 2\pi] \end{cases} .$$

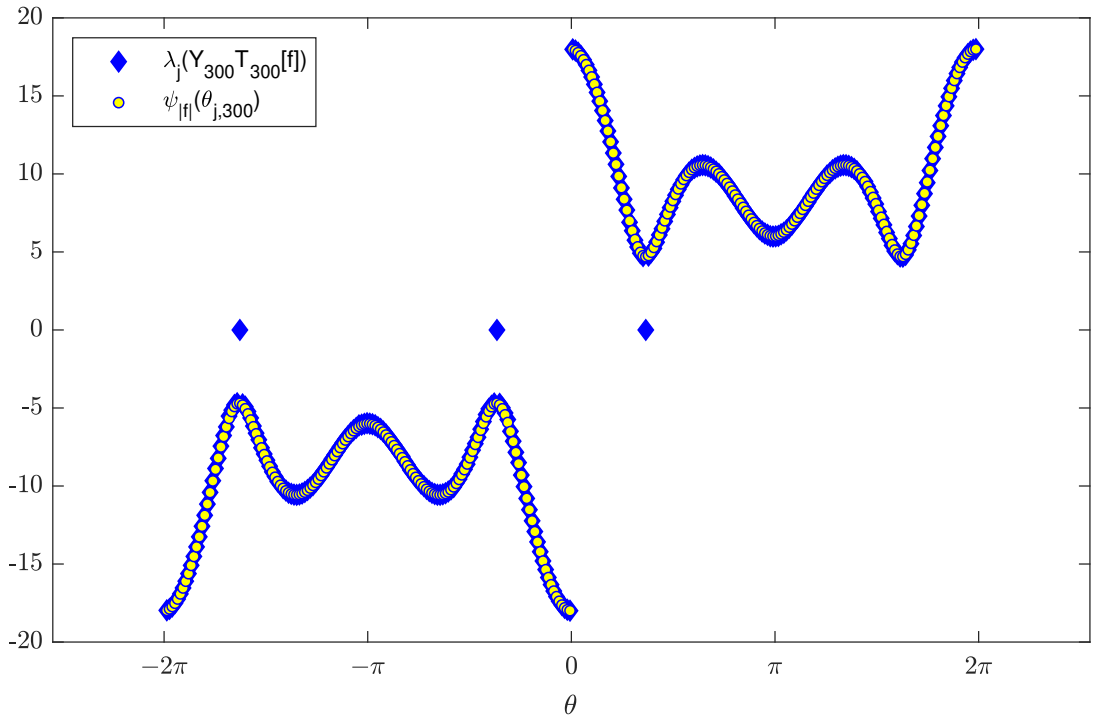

Figure II.2: Example 3, a comparison between the eigenvalues $\lambda_j(Y_nT_n[f])$, $j = 1, \ldots, n$, and the samples $\psi_{|f|}(\vartheta_{j,n})$, for $f(\vartheta) = 4 + 2\mathrm{e}^{-\mathrm{i}\vartheta} + 2\mathrm{e}^{-2\mathrm{i}\vartheta} + 9\mathrm{e}^{-3\mathrm{i}\vartheta} + \mathrm{e}^{\mathrm{i}\vartheta}$ for $n = 300$.

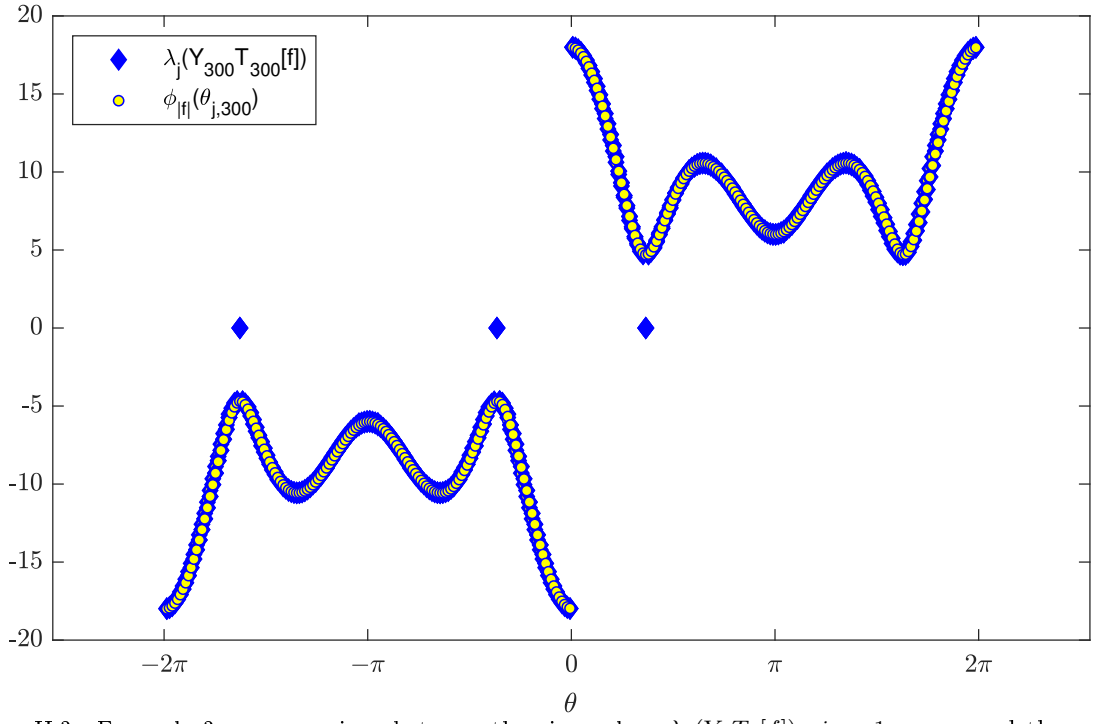

Figure II.3: Example 3, a comparison between the eigenvalues $\lambda_j(Y_nT_n[f])$, $j = 1, \ldots, n$, and the samples $\phi_{|f|}(\vartheta_{j,n})$, for $f(\vartheta) = 4 + 2\mathrm{e}^{-\mathrm{i}\vartheta} + 2\mathrm{e}^{-2\mathrm{i}\vartheta} + 9\mathrm{e}^{-3\mathrm{i}\vartheta} + \mathrm{e}^{\mathrm{i}\vartheta}$ for $n = 300$.

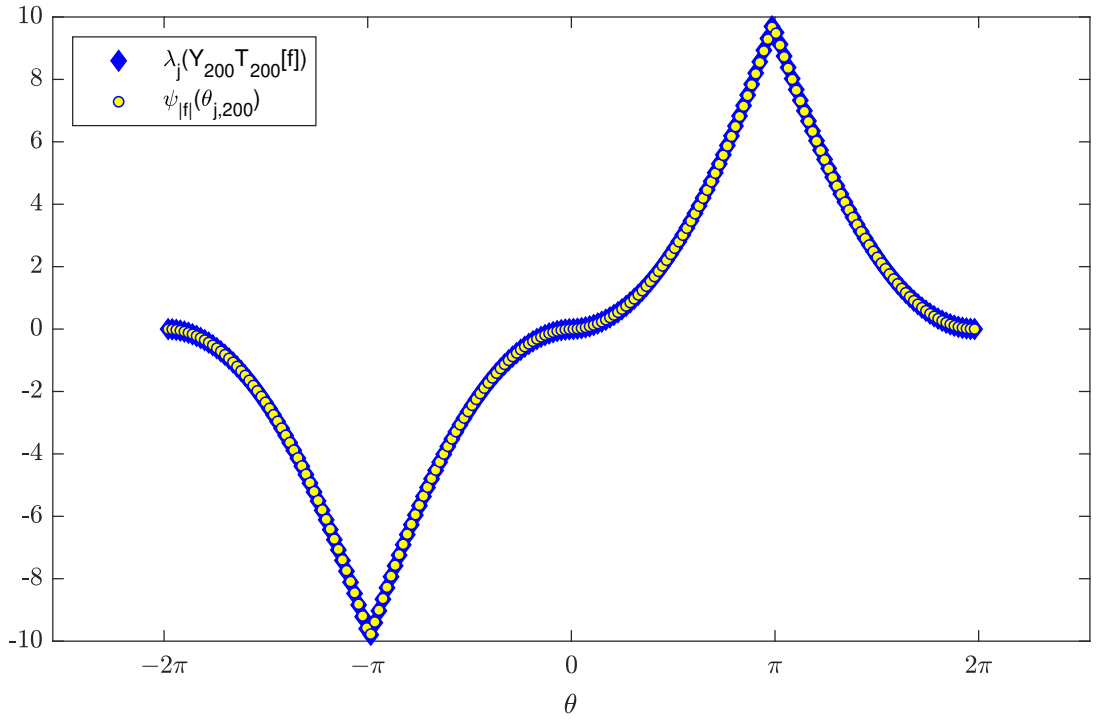

Figure II.4: Example 4, a comparison between the eigenvalues $\lambda_j(Y_n T_n[f])$ and the samples $\psi_{|f|}(\vartheta_{j,n})$, for $f(\vartheta) = \vartheta^2$ and $n = 200$.

As a consequence of the definition of f, we have that the associated function $\psi_{|f|}$ is piecewisely defined in the following 4 subintervals

$$\psi_{|f|}(\vartheta_{j,n}) = \begin{cases} -(\vartheta_{j,n} + 2\pi)^2, & \forall j = 1, \dots, \frac{n}{4}, \\ -(\vartheta_{j,n})^2, & \forall j = \frac{n}{4} + 1, \dots, \frac{n}{2}, \\ (\vartheta_{j,n})^2, & \forall j = \frac{n}{2} + 1, \dots, \frac{3n}{4}, \\ (\vartheta_{j,n} - 2\pi)^2, & \forall j = \frac{3n}{4} + 1, \dots, n \end{cases}.$$

In Figure II.4, we numerically show that the quantities $\psi_{|f|}(\vartheta_{j,n})$ approximate the eigenvalues $\lambda_j(Y_n T_n[f])$ for all $j = 1, \dots, n$, computed with MATLAB's `eig` *function. This result is expected from Theorem II.1.2, which holds for generic functions in $L^1([-\pi, \pi])$ with real Fourier coefficients.*

Example 5. *In the current example, we give numerical evidence of the distribution result of Theorem II.1.2 under the hypothesis that f is a discontinuous function $f : [-\pi, \pi] \to \mathbb{R}$, piecewisely defined by the formulae*

$$f(\vartheta) = \begin{cases} 5, & \vartheta \in [-\pi, -\pi/2), \\ 2, & \vartheta \in [-\pi/2, \pi/2), \\ 5, & \vartheta \in [\pi/2, \pi], \end{cases}$$

and periodically extended to the real line.

We fix $n = 80$ and compute $\psi_{|f|}$ on the whole grid $\vartheta_{j,n}$ with a procedure similar to that in Example 4. In Figure II.5, we show that the sampling $\psi_{|f|}(\vartheta_{j,n})$ is an approximation of the eigenvalues of the matrix $Y_n T_n[f]$ up to a constant number of outliers.

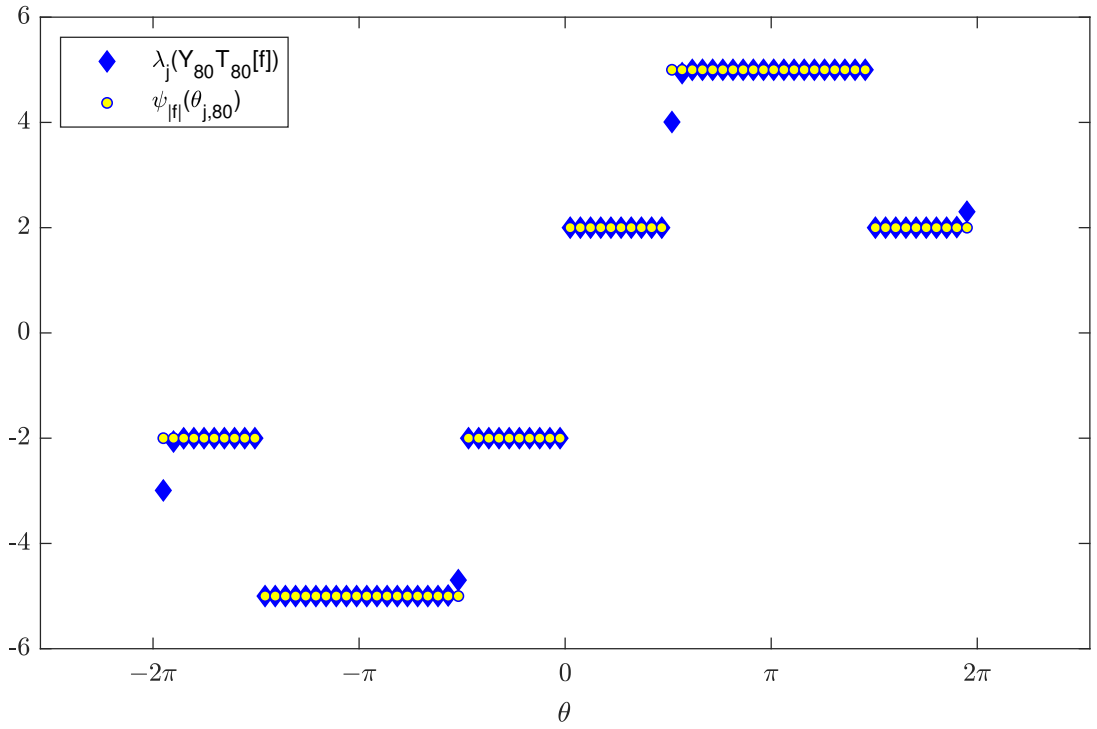

Figure II.5: Example 5, comparison between the eigenvalues $\lambda_j(Y_nT_n[f])$, $j = 1, \dots, n$, and the samples $\psi_{|f|}(\vartheta_{j,n})$, for the piecewise constant f for $n = 80$.

Example 6. *In the last example of this subsection we focus on the spectral distribution of the symmetrized Toeplitz sequence associated with a matrix-valued function* $\mathbf{f} : [-\pi, \pi] \to \mathbb{R}^{2\times 2}$ *given by*

$$\mathbf{f}(\vartheta) = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} 10 + 2\cos\vartheta & 0 \\ 0 & 2 - \cos\vartheta \end{bmatrix} \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}.$$

To numerically verify the distribution result of Theorem II.1.3 in this matrix-valued setting we need to compare the eigenvalues of the matrix $(Y_n \otimes I_s)T_n[\mathbf{f}]$ *and the evaluation of the eigenvalue functions of* $\psi_{|\mathbf{f}|}$ *on the uniform grid* $\vartheta_{j,n}$*, that is, the quantities* $\lambda_1\left(\psi_{|\mathbf{f}|}\right)(\vartheta_{j,n})$ *and* $\lambda_2\left(\psi_{|\mathbf{f}|}\right)(\vartheta_{j,n})$*. We choose* $n = 100$*, in this setting we can evaluate* $\psi_{|\mathbf{f}|}$ *on the uniform grid* $\vartheta_{j,n}$ *and then compute the quantities* $\lambda_1\left(\psi_{|\mathbf{f}|}(\vartheta_{j,n})\right)$ *and* $\lambda_2\left(\psi_{|\mathbf{f}|}(\vartheta_{j,n})\right)$ *for* $j = 1, \dots, n$*. Figure II.6 shows that the considered sampling of the eigenvalue functions approximates the eigenvalues of the matrix* $(Y_n \otimes I_s)T_n[\mathbf{f}]$ *well. Moreover, we observe the four branches of eigenvalues* $[-12, -8] \cup [-3, -1] \cup [1, 3] \cup [8, 12]$ *as described by Theorem II.1.3.*

II.4 Numerical Tests on Preconditioned Matrix-Sequences

In the current section we illustrate the predicted behaviour of the eigenvalues of the preconditioned matrix-sequences in Theorem II.2.1 for different choices of generating functions and circulant preconditioners.

In particular, in Example 7 we focus on f being a trigonometric polynomial. In Example 8 we fix f to be a quadratic function and in Example 9 we consider a discontinuous piecewise constant generating function.

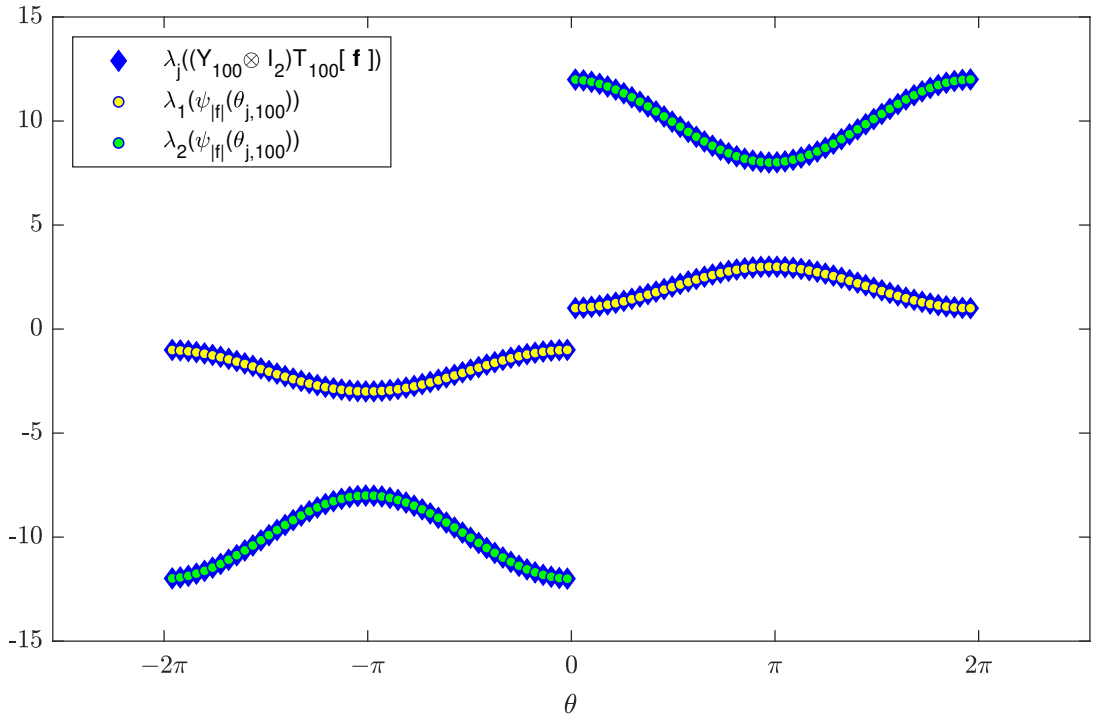

Figure II.6: Example 6, comparison between the eigenvalues $\lambda_j((Y_n \otimes I_2)T_n[\mathbf{f}])$ and the eigenvalue functions of $\psi_{|\mathbf{f}|}$ evaluated on the grid $\vartheta_{j,n}$, for the matrix-valued function $\mathbf{f}$ and $n = 100$.

In the following examples, we first verify that the condition $\{C_n^\dagger T_n[f]\}_n \sim_\sigma 1$ holds for the specific choices of generating function f and circulant preconditioner C_n. We prove this either using the discussion after Theorem II.2.1 (for Examples 7 and 8) or numerically (for Example 9).

Once such a hypothesis is verified, we graphically show that the eigenvalues of $\{|C_n|^{-1}Y_nT_n[f]\}_n$ are distributed as the function ψ_1 over $[-2\pi, 2\pi]$.

Example 7. *We consider the trigonometric polynomial*

$$f(\vartheta) = 2 - 2\mathrm{e}^{-\hat{\imath}\vartheta} - 3\mathrm{e}^{\hat{\imath}\vartheta}.$$

Since f is a nonzero polynomial, it is obviously sparsely vanishing and belongs to the Dini-Lipschitz class. Thus, we can use either Item **A** *or* **B** *after Theorem II.2.1 to realize that $\{C_n^\dagger T_n[f]\}_n \sim_\sigma 1$. We follow Item* **A** *(Item* **B** *is analogous), choosing C_n as the Strang preconditioner for $T_n[f]$.*

In Figure II.7, we plot the eigenvalues of $|C_n|^{-1}Y_nT_n[f]$ for different values of n. For both $n = 500$ and $n = 1000$, we observe that the values $\lambda_j(|C_n|^{-1}Y_nT_n[f])$ are distributed as the function ψ_1, as predicted by Theorem II.2.1. In fact, except for a constant number of outliers, half of the eigenvalues are equal to -1 and the other half are equal to 1.

Example 8. *We consider the generating function*

$$f(\vartheta) = \vartheta^2.$$

The discussion following Theorem II.2.1 assures us that, in this case, we can use both the Strang preconditioner and the Frobenius optimal preconditioner. For the current example, we show the results obtained from the two types of preconditioners for different n.

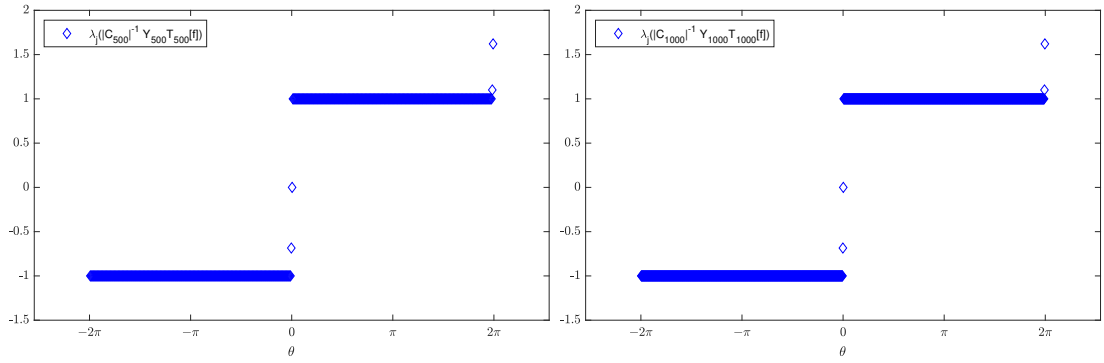

Figure II.7: Example 7, the eigenvalues of $|C_n|^{-1}Y_nT_n[f]$, where $f(\vartheta) = 2 - 2e^{-\hat{\imath}\vartheta} - 3e^{\hat{\imath}\vartheta}$, C_n is the Strang preconditioner, and $n = 500$ or 1000.

In Figure II.8, we plot the eigenvalues $\lambda_j(|C_n|^{-1}Y_nT_n[f])$, where C_n is the Strang preconditioner for $n = 157, 200, 589,$ or 1000. For all tested n, the largest eigenvalue $\lambda_n(|C_n|^{-1}Y_nT_n[f])$ is an outlier and becomes large quickly as n increases. Consequently, this large outlier is not plotted for a better visualization of the values $\lambda_j(|C_n|^{-1}Y_nT_n[f])$ for $j = 1, \ldots, n-1$.

Notice that the spectrum of $|C_n|^{-1}Y_nT_n[f]$ is divided into two sets with almost the same cardinality: the first contains the eigenvalues equal to -1 and the second contains those equal to 1. Finally, the outliers that do not belong to the previous group are infinitesimal in the dimension n of the matrix.

In Figure II.9, an analogous clustering of eigenvalues is shown using the Frobenius preconditioner for $n = 157, 200, 589,$ or 1000. In this second experiment, the Frobenius preconditioner gives us a worse result in terms of outliers. In fact, the number of outliers is significantly larger than that in the Strang preconditioner case. However, it is still infinitesimal with respect to n as expected from Theorem II.2.1.

Example 9. *In this last example, we consider the discontinuous function*

$$f(\vartheta) = \begin{cases} 5, & \vartheta \in [-\pi, -\pi/2), \\ 2, & \vartheta \in [-\pi/2, \pi/2), \\ 5, & \vartheta \in [\pi/2, \pi]. \end{cases}$$

In this case, instead of using Item **B**, *we show in Figure II.10 graphically that the property*

$$\{C_n^{\dagger}T_n[f]\}_n \sim_{\sigma} 1,$$

is true for the Strang preconditioner.

In Figure II.11, we plot the eigenvalues $\lambda_j(|C_n|^{-1}Y_nT_n[f])$, $j = 1, \ldots, n-1$, for $n = 500$ or 1000. In both cases, the eigenvalue $\lambda_n(|C_n|^{-1}Y_nT_n[f])$ is an outlier of large magnitude and therefore we do not plot it as before.

The clustering of the spectrum around ± 1 numerically confirms the distribution result on the preconditioned matrix-sequence $\{|C_n|^{-1}Y_nT_n[f]\}_n$ in a more general hypothesis of Theorem II.2.1.

In this Chapter we focused the eigenvalue distribution of sequences of the form $\{Y_nT_n[f]\}_n$. In the next Chapter we extend the analysis to the case of symmetrization of matrix-sequences of

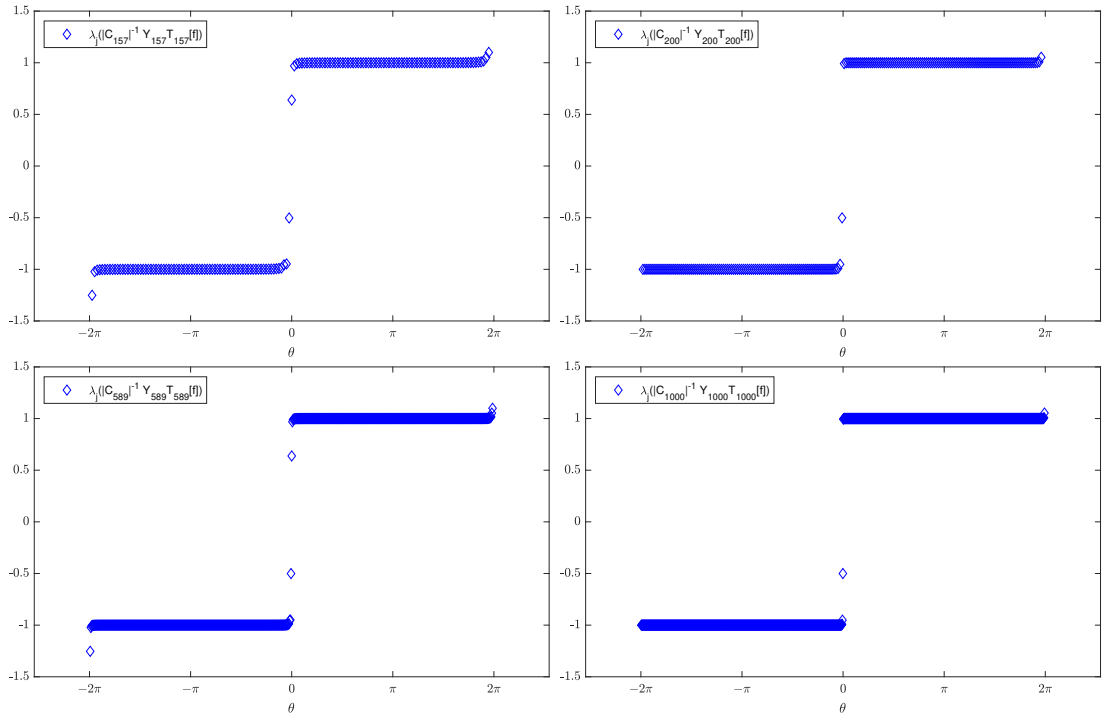

Figure II.8: Example 8, the eigenvalues of $|C_n|^{-1}Y_nT_n[f]$, where $f(\vartheta) = \vartheta^2$, C_n is the Strang preconditioner, and $n = 157, 200, 589$ or 1000. The largest eigenvalue $\lambda_n(|C_n|^{-1}Y_nT_n[f])$ is an outlier – approximately 10^4 for all values of n – and it is not plotted.

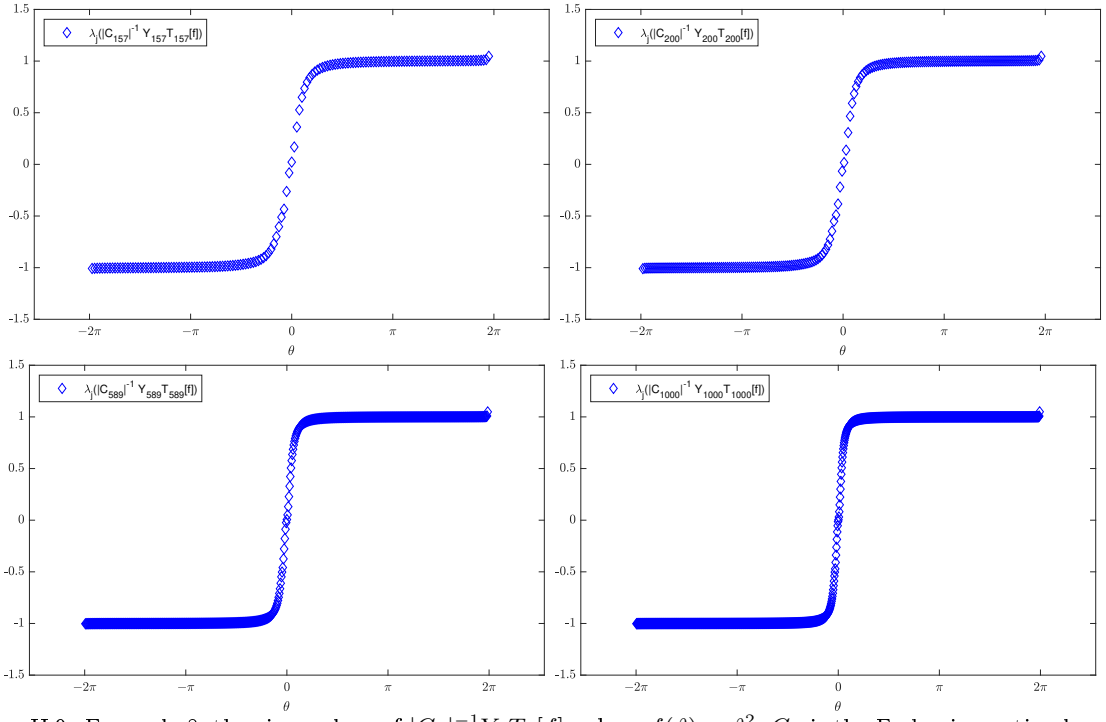

Figure II.9: Example 8, the eigenvalues of $|C_n|^{-1}Y_nT_n[f]$, where $f(\vartheta) = \vartheta^2$, C_n is the Frobenius optimal preconditioner, and $n = 157, 200, 589$, or 1000. The largest eigenvalue $\lambda_n(|C_n|^{-1}Y_nT_n[f])$ is an outlier – approximately 10^2 for all values of n – and it is not plotted.

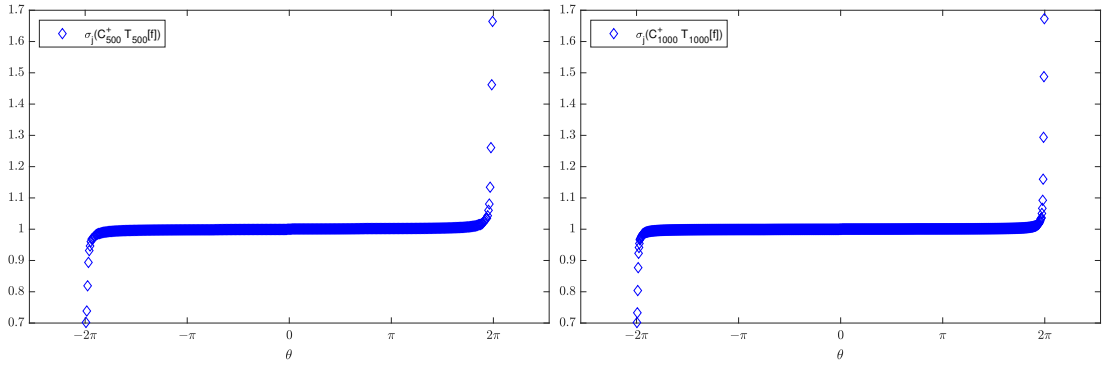

Figure II.10: Example 9, the singular values of $C_n^\dagger T_n[f]$, where f is piecewise constant, C_n is the Strang preconditioner, and $n = 500$ or 1000.

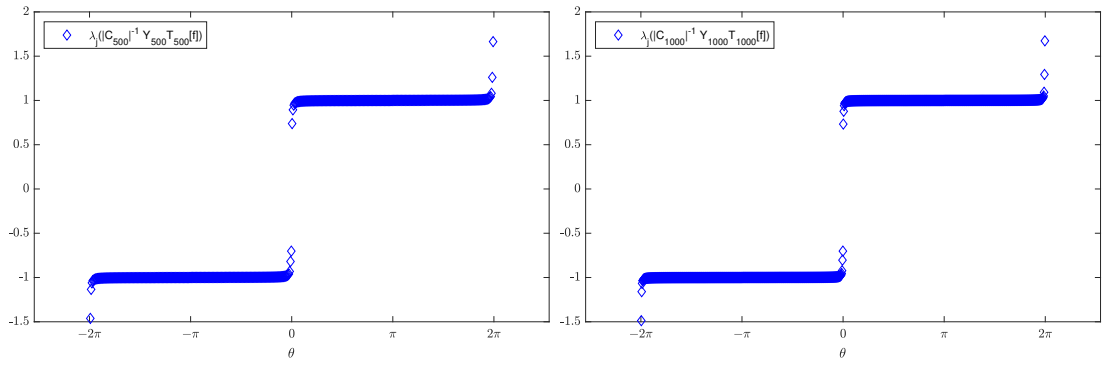

Figure II.11: Example 9, the eigenvalues of $|C_n|^{-1}Y_nT_n[f]$, where f is piecewise constant, C_n is the Strang preconditioner, and $n = 500$ or 1000.

the form $\{h(T_n[f])\}_n$, where h is an analytic function. In particular we investigate the singular value distribution of sequences of the form $\{h(T_n[f])\}_n$ and we provide a result on the spectral distribution of the symmetrizated sequence $\{Y_n h(T_n[f])\}_n$.

Chapter III

Asymptotic Spectral Distributions of Symmetrized Toeplitz Structure Functions

Following the numerical evidences in [79] and the algorithmic proposals in [78], the purpose of this chapter is to extend the results concerning the eigenvalue distribution of $\{Y_n T_n[f]\}_n$ that we obtained in **Chapter II** to the symmetrization of matrix-sequences of the form $\{h(T_n[f])\}_n$, where h is an analytic function.

Our work is motivated also by the fact that functions of Toeplitz matrices have crucial relevance in several applications. For instance, exponential functions of Toeplitz matrix-sequences arise from the discretization of integro-differential equations with a shift-invariant kernel [46]. Furthermore, trigonometric functions are involved in the case of the approximation by local methods of differential equations [75].

In particular, we consider a function f in $L^\infty([-\pi,\pi])$ with real Fourier coefficients and an analytic function h with convergence radius r such that $\|f\|_\infty < r$. Under these hypotheses, we prove that the matrix-sequence $\{h(T_n[f])\}_n$ is distributed in the singular value sense as $h \circ f$. We exploit this property to investigate the spectral distribution of the symmetrized sequence $\{Y_n h(T_n[f])\}_n$ and we prove that its spectral symbol is given by

$$\phi_{|h\circ f|}(\vartheta) = \begin{cases} |h \circ f(\vartheta)|, & \vartheta \in [0, 2\pi], \\ -|h \circ f(-\vartheta)|, & \vartheta \in [-2\pi, 0), \end{cases}$$

which has the same structure of the eigenvalue symbol of $\{Y_n T_n[f]\}_n$ that we derived in the previous chapter. The proof of the distribution result concerning the sequence $\{Y_n h(T_n[f])\}_n$ is based both on Theorem II.1.2 and on the features of the GLT theory that we introduced in Section I.7.

As we detailed in the introductory chapter, spectral distribution results represent key ingredients in the design and in the convergence analysis of multigrid methods and preconditioned Krylov solvers. Following this direction, in Section III.3 we numerically study the spectral properties of ad-hoc preconditioners for the previously analyzed symmetrized sequences. Thanks to the symmetry of the considered matrices, these preconditioners may also be used to fasten the convergence of Krylov solvers such as MINRES.

The results presented in the following sections are published in [52] and the chapter is outlined as follows. Firstly, in Section III.1, we give our main theorem on the asymptotic distributions of $\{h(T_n[f])\}_n$ and $\{Y_n h(T_n[f])\}_n$. Then, in Section III.2 we numerically support the derived distribution results for several choices of generating functions f and analytic functions h, including a significant example stemming from computational finance. Finally, in Section III.3 we analyse the examples of the previous section to define and compare different preconditioning strategies for the matrices $\{h(T_n[f])\}_n$.

III.1 Asymptotic Distributions of $\{h(T_n[f])\}_n$ and $\{Y_n h(T_n[f])\}_n$

In this section, we provide the main asymptotic distribution results on the sequences $\{h(T_n[f])\}_n$ and $\{Y_n h(T_n[f])\}_n$ in the case where $f \in L^\infty([-\pi,\pi])$ has real Fourier coefficients and h is a real analytic function in 0 with radius of convergence r such that $\|f\|_\infty < r$. Following the discussion in Section I.8, we notice that under these hypotheses the matrix $Y_n h(T_n[f])$ is real symmetric.

Furthermore, we stress that the function $h \circ f(\vartheta) = h(f(\vartheta))$ defined on $[-\pi,\pi]$ plays a very important role in the expression of the underlying symbols.

Lemma III.1.1. *Suppose $f \in L^\infty([-\pi,\pi])$ with real Fourier coefficients and let $Y_n \in \mathbb{R}^{n\times n}$ be the anti-identity matrix. Let $T_n[f] \in \mathbb{R}^{n\times n}$ be the Toeplitz matrix generated by f. Let $p(z)$ be a polynomial. Then*

$$\{p(T_n[f])\}_n \sim_\sigma p \circ f.$$

Proof. The thesis is an immediate consequence of Items **GLT1**, **GLT2**, **GLT3**, and of the fact that p is a polynomial, since $\{p(T_n[f])\}_n \sim_{\text{GLT}} \tilde{f} = p \circ f$. □

Theorem III.1.2. *Suppose $f \in L^\infty([-\pi,\pi])$ with real Fourier coefficients and let $Y_n \in \mathbb{R}^{n\times n}$ be the anti-identity matrix. Let $T_n[f] \in \mathbb{R}^{n\times n}$ be the Toeplitz matrix generated by f. Let $h(z)$ be a real analytic function in 0 with radius of convergence r such that $\|f\|_\infty < r$. Then we have the following asymptotic distributions:*

$$\{h(T_n[f])\}_n \sim_\sigma h \circ f \tag{III.1}$$

and

$$\{Y_n h(T_n[f])\}_n \sim_\lambda \psi_{|h\circ f|}. \tag{III.2}$$

where $\psi_{|h\circ f|}$ is defined as in (II.1) over the domain $\tilde{D}$ with $D = [0, 2\pi]$ and $p = -2\pi$.

Proof. Notice that the assumption $\|f\|_\infty < r$ implies $\|T_n[f]\|_2 < r$ and hence $\rho(T_n[f]) < r$ (see I.5.2). Consequently, as we detailed in Section I.8, Theorem 4.7 in [74] guarantees that $h(T_n[f])$ is well-defined.

If $|z| < r$, we can represent $h(z)$ through its Taylor series expansion in 0, that is $h(z) = \sum_{k=0}^{\infty} b_k z^k$. For $m \in \mathbb{N}$, we define the polynomial

$$p_m(z) = \sum_{k=0}^{m} b_k z^k.$$

We have the following properties:

1. $\{p_m(T_n[f])\}_n \sim_\sigma p_m \circ f$ for all $m \in \mathbb{N}$;
2. $\{\{p_m(T_n[f])\}_n\}_m$ is an a.c.s. for $\{h(T_n[f])\}_n$;
3. $p_m \circ f \to h \circ f$ in measure.

The first property is a consequence of Lemma III.1.1. The second property can be proven from the decomposition

$$h(T_n[f]) = p_m(T_n[f]) + (h(T_n[f]) - p_m(T_n[f])),$$

by observing that $\|h(T_n[f]) - p_m(T_n[f])\| < \varepsilon_m$ with

$$\lim_{m\to\infty} \varepsilon_m = 0,$$

taking into account Definition I.4.4.

For proving the third property, notice that the assumption $\|f\|_\infty < r$ guarantees that h is analytic in $f(\vartheta)$ almost everywhere on $\vartheta \in [-\pi, \pi]$. It follows that $p_m \circ f$ converges almost everywhere to $h \circ f$ and the convergence in measure is a consequence of the boundedness of the domain.

Hence, the objects $\{\{p_m(T_n[f])\}_n\}_m$, $\{h(T_n[f])\}_n$, p_m and h satisfy the assumptions of Lemma I.4.2, from which we can infer the first part of the thesis:

$$\{h(T_n[f])\}_n \sim_\sigma h \circ f.$$

Moreover, Property **GLT5** implies that the matrix-sequence $\{h(T_n[f])\}_n$ is GLT with symbol $h \circ f$.

For proving (III.2), let us define the quantity

$$\Delta_n(h, f) = h(T_n[f]) - T_n[h \circ f].$$

Since $h \circ f \in L^1([-\pi, \pi])$, by Theorem I.5.3 the Toeplitz matrix-sequence $\{T_n[h \circ f]\}_n$ is distributed in the singular value sense as $h \circ f$ and it is a GLT matrix-sequence. By (III.1), also $\{h(T_n[f])\}_n$ is distributed in the singular value sense as $h \circ f$ and it is a GLT matrix-sequence. Hence, Properties **GLT1-GLT2** imply that the GLT sequence $\{\Delta_n(h, f)\}_n$ is distributed as 0 in the singular value sense.

Since Y_n is a unitary matrix, also the matrix-sequence $\{Y_n\Delta_n(h, f)\}_n$ is zero-distributed in the singular value sense. From [62, Chapter 9] we know that $Y_n\Delta_n(h, f) \sim_\sigma 0$ if and only if $Y_n\Delta_n(h, f) = R_n + N_n$ with

$$\lim_{n\to\infty} \frac{\operatorname{rank}(R_n)}{n} = \lim_{n\to\infty} \|N_n\| = 0. \tag{III.3}$$

Note that, by Lemma I.8.1, the matrix $Y_n\Delta_n(h, f)$ is Hermitian for all n; from Properties **GLT1** and **GLT4** we see that the spectral distribution of the corresponding matrix-sequence is given by

$$\{Y_n\Delta_n(h, f)\}_n \sim_\lambda 0.$$

Thanks to the definition of $\Delta_n(h, f)$, we can write

$$\{Y_n h(T_n[f])\}_n = \{Y_n T_n[h \circ f]\}_n + \{Y_n\Delta_n(h, f)\}_n. \tag{III.4}$$

Then, the constant (not depending on m) class of sequences $\{\{B_n\}_n\}_m = \{Y_nT_n[h \circ f]\}_n$ is an a.c.s for $\{Y_nh(T_n[f])\}_n$. In fact, we can write $Y_nh(T_n[f])$ as in formula (III.4) and, from (III.3), we have that the matrix-sequence $\{Y_n\Delta_n(h,f)\}_n$ verifies the low- rank plus small-norm requirement of the Definition I.4.4.

As already stated, the function $h \circ f$ belongs to $L^1([-\pi,\pi])$, then, from Theorem II.1.2 it follows that

$$\{Y_nT_n[h \circ f]\}_n \sim_\lambda \psi_{|h\circ f|}. \tag{III.5}$$

Hence, the desired result

$$\{Y_nh(T_n[f])\}_n \sim_\lambda \psi_{|h\circ f|}$$

follows directly from the second part of Lemma I.4.2. □

III.2 Numerical Experiments on the Asymptotic Distributions of $\{Y_nh(T_n[f])\}_n$

In the present section we provide different examples in order to show that the statements of Theorem III.1.2 are numerically evident already in the case of really moderate matrix sizes. Indeed, we consider a function f in $L^\infty([-\pi,\pi])$ with real Fourier coefficients and an analytic function h with convergence radius r such that $\|f\|_\infty < r$ and we show that the singular values of the matrix $h(T_n[f])$ are well approximated by a uniform sampling of $|h \circ f|$ over its domain and that the eigenvalues of $Y_nh(T_n[f])$ are well approximated by a uniform sampling of $\psi_{|h\circ f|}$.

In particular, we consider the case where f is a trigonometric polynomial and h is an analytic function with convergent Taylor series in a neighbourhood of the origin (Examples 10–11) or a polynomial (Example 12). Furthermore, in Example 13 we study the spectral properties of the symmetrization of the exponential of a Toeplitz matrix generated by a high-degree trigonometric polynomial stemming from computational finance.

Example 10. *We take into consideration the analytic function $h(z) = \sin(z)$, whose Taylor series at 0 converges in the whole complex plane, and we consider the trigonometric polynomial $f(\vartheta) = \mathrm{e}^{\hat{\imath}\vartheta}$. Figure III.1 shows that for $n = 100$ the eigenvalues of $Y_nh(T_n[f])$ are well approximated by a uniform sampling of $\psi_{|h\circ f|}$ over $[-2\pi, 2\pi]$, except for the presence of one outlier.*

This behaviour numerically confirms the spectral distribution predicted by Theorem III.1.2. In fact, Definition I.4.3 contemplates the presence of eigenvalues not captured by the sampling of $\psi_{|h\circ f|}$.

Example 11. *We now consider the analytic function $h(z) = \log(1+z)$, whose Taylor series at 0 converges with the radius of convergence equals 1. Moreover, we take the trigonometric polynomial $f(\vartheta) = 0.5\mathrm{e}^{\hat{\imath}\vartheta}$, with $\|f\|_\infty < 1$ as Theorem III.1.2 demands. In Figure III.2 we can observe that, except again for one outlier, the eigenvalues of $Y_nh(T_n[f])$, for $n = 100$, are well approximated by a uniform sampling of $\psi_{|h\circ f|}$ over $[-2\pi, 2\pi]$.*

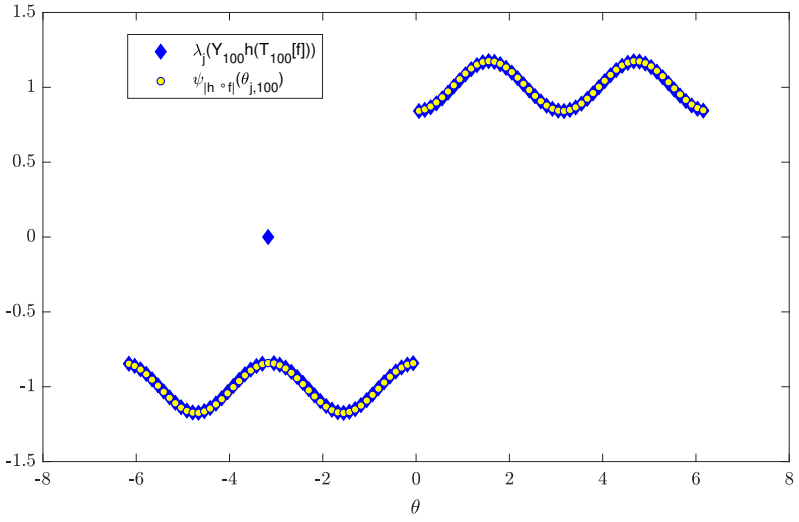

Figure III.1: Comparison between the eigenvalues of the symmetrized matrix $Y_{100}h(T_{100}[f])$ and the uniform sampling of $\psi_{|h\circ f|}$, over $[-2\pi, 2\pi]$, for $h(z) = \sin(z)$ and $f(\vartheta) = \mathrm{e}^{\hat{\imath}\vartheta}$.

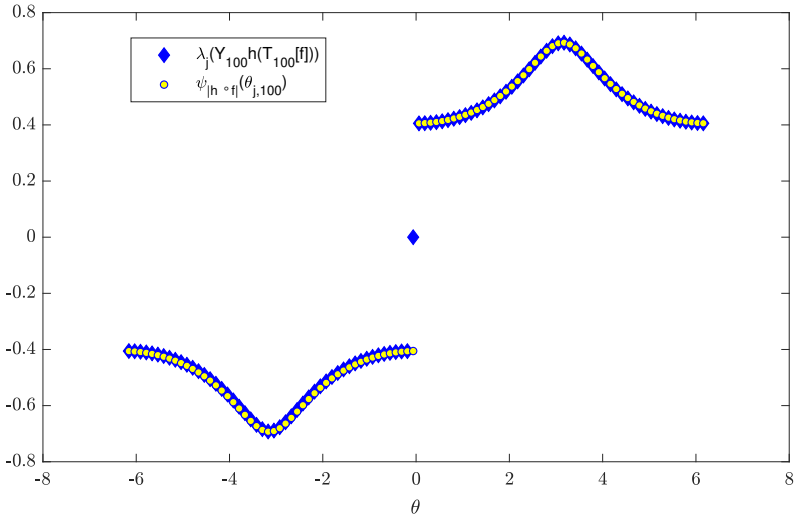

Figure III.2: Comparison between the eigenvalues of the symmetrized matrix $Y_{100}h(T_{100}[f])$ and the uniform sampling of $\psi_{|h\circ f|}$, over $[-2\pi, 2\pi]$, for $h(z) = \log(1+z)$ and $f(\vartheta) = 0.5\mathrm{e}^{\hat{\imath}\vartheta}$.

Example 12. *The example is taken from [78]. Following the same procedure of Examples 1-2, we plot in Figure III.3 the spectrum of $Y_n h(T_n[f])$, for $n = 200$, for the function $h(z) = 1+z+z^2$, whose Taylor series in 0 converges in the whole complex plane, and the trigonometric polynomial $f(\vartheta) = -\mathrm{e}^{\hat{\imath}\vartheta} + 1 + \mathrm{e}^{-\hat{\imath}\vartheta} + \mathrm{e}^{-\hat{\imath}2\vartheta} + \mathrm{e}^{-\hat{\imath}3\vartheta}$. In the present example we can observe that there are no outliers and the eigenvalues of $Y_n h(T_n[f])$ are approximated by the uniform sampling of $\psi_{|h\circ f|}$ over $[-2\pi, 2\pi]$. Moreover, in order to numerically confirm relation (III.1) of Theorem III.1.2, we verify that the singular values of the matrix $h(T_n[f])$ can be approximated by a uniform sampling of $|h \circ f|$ over $[0, 2\pi]$. Indeed, Figure III.4 shows that the expected approximation holds true already for a moderate size such as $n = 200$.*

Example 13. *The last example is a practical case taken from [90, 91]. Here we consider the*

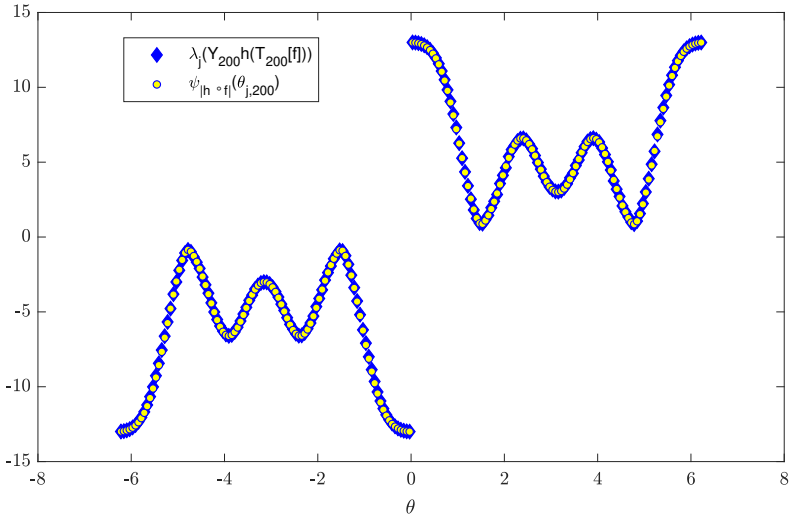

Figure III.3: Comparison between the eigenvalues of the symmetrized matrix $Y_{200}h(T_{200}[f])$ and the uniform sampling of $\psi_{|h\circ f|}$, over $[-2\pi, 2\pi]$, for $h(z) = 1 + z + z^2$ and $f(\vartheta) = -\mathrm{e}^{\hat{\imath}\vartheta} + 1 + \mathrm{e}^{-\hat{\imath}\vartheta} + \mathrm{e}^{-\hat{\imath}2\vartheta} + \mathrm{e}^{-\hat{\imath}3\vartheta}$.

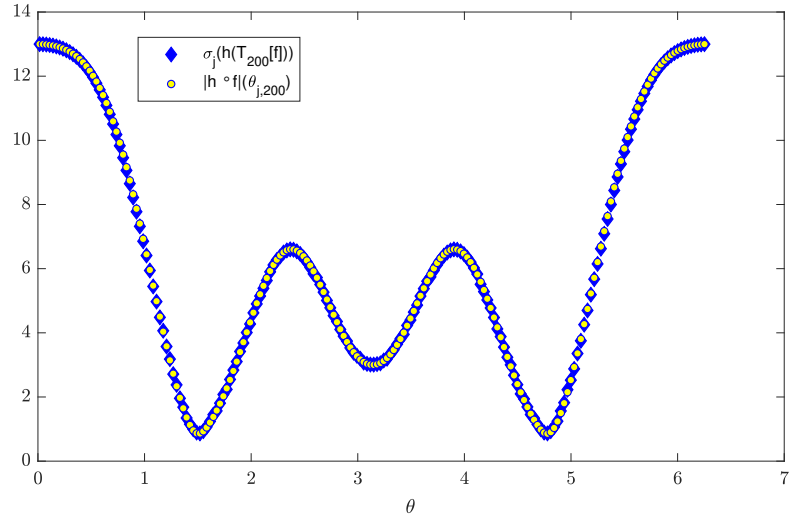

Figure III.4: Comparison between the singular values of the matrix $h(T_{200}[f])$ and the uniform sampling of $|h \circ f|$, over $[0, 2\pi]$, for $h(z) = 1 + z + z^2$ and $f(\vartheta) = -\mathrm{e}^{\hat{\imath}\vartheta} + 1 + \mathrm{e}^{-\hat{\imath}\vartheta} + \mathrm{e}^{-\hat{\imath}2\vartheta} + \mathrm{e}^{-\hat{\imath}3\vartheta}$.

case of the exponential of a real nonsymmetric Toeplitz matrix stemming from computational finance, in particular from the option pricing framework in jump-diffusion models, where a partial integro-differential equation (PIDE) needs to be solved. Indeed, the discretization of a PIDE can be transformed into a matrix exponential problem. In our notation, it is equivalent to consider the analytic function $h(z) = \mathrm{e}^z$*, whose Taylor series centred at* 0 *converges in the whole complex plane, and a trigonometric polynomial* $f(\vartheta) = \sum_{j=-n+1}^{n-1} \hat{f}_j \mathrm{e}^{\hat{\imath}j\vartheta}$ *defined by the following Fourier coefficients:*

$$\hat{f}_0 = -\nu^2 - \Delta x^2(r + \lambda - \lambda w(0)\Delta x); \tag{III.6}$$

$$\hat{f}_1 = \frac{\nu^2}{2} - \Delta x \frac{(2r - 2\lambda k - \nu^2)}{4} + \lambda w(-\Delta x)\Delta x^3; \tag{III.7}$$

$$\hat{f}_{-1} = \frac{\nu^2}{2} + \Delta x \frac{(2r - 2\lambda k - \nu^2)}{4} + \lambda w(\Delta x)\Delta x^3; \tag{III.8}$$

$$\hat{f}_j = \lambda \Delta x^3 w(-j\Delta x), \quad j \in \{-n+1, \ldots, -2, \} \cup \{2, \ldots, n-1\}. \tag{III.9}$$

where $w(s) = \frac{e^{-\frac{(s-\mu)^2}{2\sigma^2}}}{\sqrt{2\pi}\sigma}$, *is a normal distribution function with mean* μ *and standard deviation* σ, *the parameter* $k = e^{\mu + \frac{\sigma^2}{2}} - 1$ *is the expectation of the impulse function,* Δx *is the spatial step-size,* ν *is the stock return volatility,* r *is the risk-free interest rate, and* λ *is the arrival intensity of a Poisson process.*

Following the same procedure of Examples 1-3, we plot in Figure III.5 the spectrum of $Y_n h(T_n[f])$, *for* $n = 100$. *In the present example we can observe that there are no outliers and the eigenvalues of* $Y_n h(T_n[f])$ *are well approximated by the uniform sampling of* $\psi_{|h \circ f|}$ *over* $[-2\pi, 2\pi]$.

In addition, in order to numerically validate relation (III.1), in Figure III.6, for $n = 100$, *we compare the singular values of* $h(T_n[f])$ *and a uniform sampling of* $|h \circ f|$ *over* $[0, 2\pi]$.

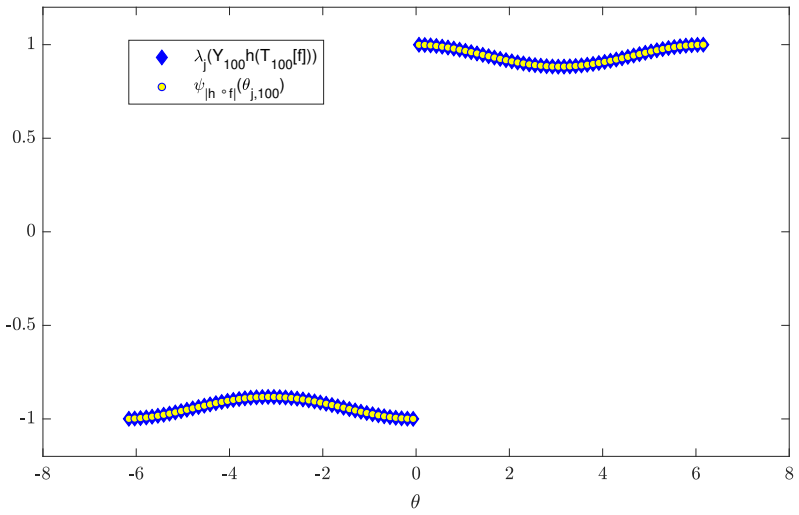

Figure III.5: Comparison between the eigenvalues of the symmetrized matrix $Y_{100}h(T_{100}[f])$ and the uniform sampling of $\psi_{|h \circ f|}$, over $[-2\pi, 2\pi]$, for $h(z) = e^z$ and $f(\vartheta) = \sum_{j=-99}^{99} \hat{f}_j e^{\hat{\imath} j \vartheta}$, with $\lambda = 0.1$, $\mu = -0.9$, $\nu = 0.25$, $\sigma = 0.45$, $r = 0.05$, and $\Delta x = \frac{4}{101}$.

III.3 Numerical Study of a Circulant Preconditioner

In the current section we exploit the derived spectral information on the matrix-sequences of the form $\{Y_n h(T_n[f])\}_n$ in order to speed up the convergence of the MINRES method for the related linear systems. For the latter purpose, we suggest a preconditioner P_n for the symmetrized

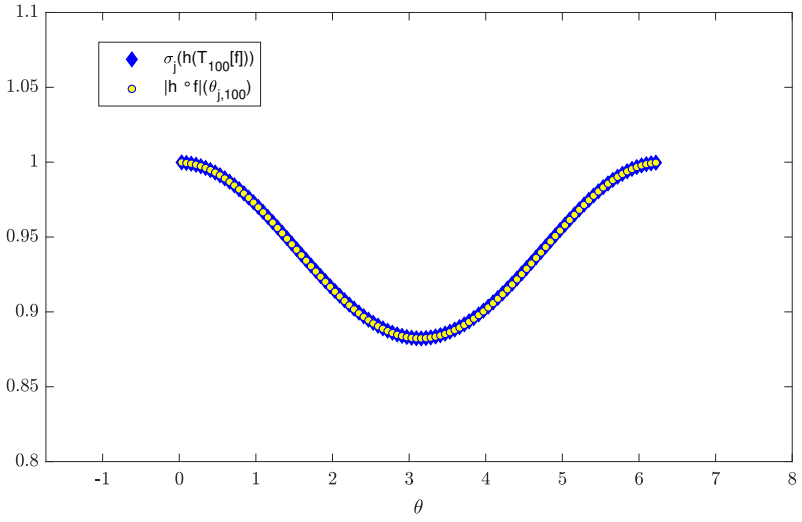

Figure III.6: Comparison between the singular values of the matrix $h(T_{100}[f])$ and the uniform sampling of $|h \circ f|$, over $[0, 2\pi]$, for $h(z) = \mathrm{e}^z$ and $f(\vartheta) = \sum_{j=-99}^{99} \hat{f}_j \mathrm{e}^{\hat{\imath} j \vartheta}$, with $\lambda = 0.1$, $\mu = -0.9$, $\nu = 0.25$, $\sigma = 0.45$, $r = 0.05$, and $\Delta x = \frac{4}{101}$.

matrix $Y_n h(T_n[f])$ and we numerically investigate the behaviour of the asymptotic spectrum of the preconditioned matrix-sequence $\{P_n^{-1} Y_n h(T_n[f])\}_n$.

For the development of a first preconditioning strategy we follow the approach introduced in [78] and we report the numerical evidence of the preconditioner efficiency in terms of eigenvalue clusters in Examples 14-16. Moreover, we also consider a further class of preconditioners whose efficiency is motivated by the theoretical results in Section II.2 and by the relation between the spectral distributions of $\{Y_n h(T_n[f])\}_n$ and $\{Y_n(T_n[h \circ f])\}_n$. The application of both strategies to the cases considered in Examples 11-13 shows that the two approaches are both valid and have a comparable performance.

In all the examples the construction of the preconditioner involves the concepts of absolute value of a circulant matrix that we defined in Section II.2 and of Frobenius optimal preconditioner that we introduced in Subsection I.9.3. For simplicity, in the present section the optimal Frobenius preconditioner for a Toeplitz matrix $T_n[f]$ is denoted by $c(T_n[f])$.

Example 14. *In this example we test the efficiency as preconditioner of the absolute value circulant matrix $|c(T_n[h \circ f])|$, for the symmetrized matrix $Y_n h(T_n[f])$, where $c(T_n[h \circ f])$ is the Frobenius optimal circulant preconditioner associated with the matrix $T_n[h \circ f]$. We consider the functions $h(z) = \log(1 + z)$ and $f(\vartheta) = 0.5\mathrm{e}^{\hat{\imath}\vartheta}$. This choice is motivated by the fact that the sequences $\{Y_n h(T_n[f])\}_n$ and $\{Y_n(T_n[h \circ f])\}_n$ share the same asymptotic spectral distribution described by $\psi_{|h \circ f|}$. Indeed in the following setting we have $h \circ f \in L^1([-\pi, \pi])$, then the results in Section II.2 suggest that $P_n = |c(T_n[h \circ f])|$ is a good preconditioner for the matrix-sequence $\{Y_n(T_n[h \circ f])\}_n$ and consequently for $\{Y_n h(T_n[f])\}_n$ as well. Moreover, the efficiency of the preconditioning strategy is highlighted if we compare the latter cluster result with the plot of the eigenvalues, sorted in the increasing order, of the non preconditioned matrix $Y_n h(T_n[f])$, shown in the top panel of Figure III.7. We highlight that the choice of the preconditioner is not unique. Indeed, we can precondition the sequence $\{Y_n h(T_n[f])\}_n$ following the approach*

introduced in [78], that is, we consider $P_n = |h(c(T_n[f]))|$, *where* $c(T_n[f])$ *is the Frobenius optimal circulant preconditioner associated with the matrix* $T_n[f]$. *We can see the efficiency of both strategies looking at Figure III.7 where we numerically confirm that the eigenvalues of the preconditioned matrix* $P_n^{-1}Y_n h(T_n[f])$, *for* $n = 512$ *are clustered around -1 and 1, up to* $o(n)$ *outliers. In particular, in the bottom left we use the preconditioner* $P_n = |c(T_n[h \circ f])|$, *and in the bottom right the preconditioner is* $P_n = |h(c(T_n[f]))|$.

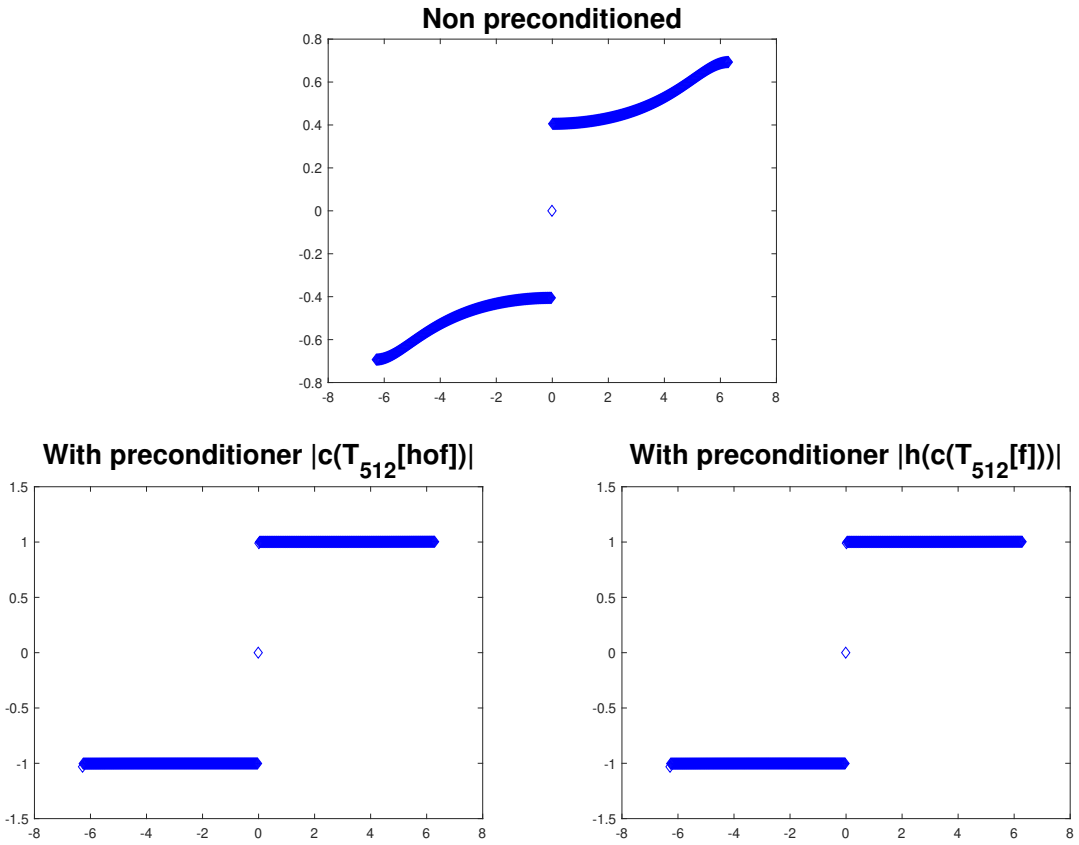

Figure III.7: The spectrum of the symmetrized matrix $Y_{512}h(T_{512}[f])$, for $h(z) = \log(1+z)$ and $f(\vartheta) = 0.5e^{\hat{\imath}\vartheta}$. Top: without preconditioner, bottom left: preconditioner $P_n = |c(T_n[h \circ f])|$, bottom right: preconditioner $P_n = |h(c(T_n[f]))|$.

Example 15. *In the present example we consider the functions as in Example 12, that is* $h(z) = 1+z+z^2$ *and* $f(\vartheta) = -e^{\hat{\imath}\vartheta}+1+e^{-\hat{\imath}\vartheta}+e^{-\hat{\imath}2\vartheta}+e^{-\hat{\imath}3\vartheta}$. *In Figure III.8, we show the behaviour of the eigenvalues of the matrix* $Y_{512}h(T_{512}[f])$ *with and without the use of a preconditioning strategy. In particular, on the top we plot the eigenvalues of the matrix* $Y_{512}h(T_{512}[f])$, *sorted in increasing order. In the bottom left and bottom right panels of Figure III.8 we test the efficiency of both preconditioning strategies described in the previous example. In both cases, we can observe that the eigenvalues of the preconditioned matrix are clustered at -1 and 1, up to* $o(n)$ *outliers.*

Example 16. *The last preconditioning test is performed on the case stemming from computational finance that we studied in Example 13. That is, we consider the case where* $h(z) = e^z$ *and* $f(\vartheta) = \sum_{j=-99}^{99} \hat{f}_j e^{\hat{\imath}j\vartheta}$, *with* a_j *defined as in (III.7)-(III.9). First, we apply the preconditioning*

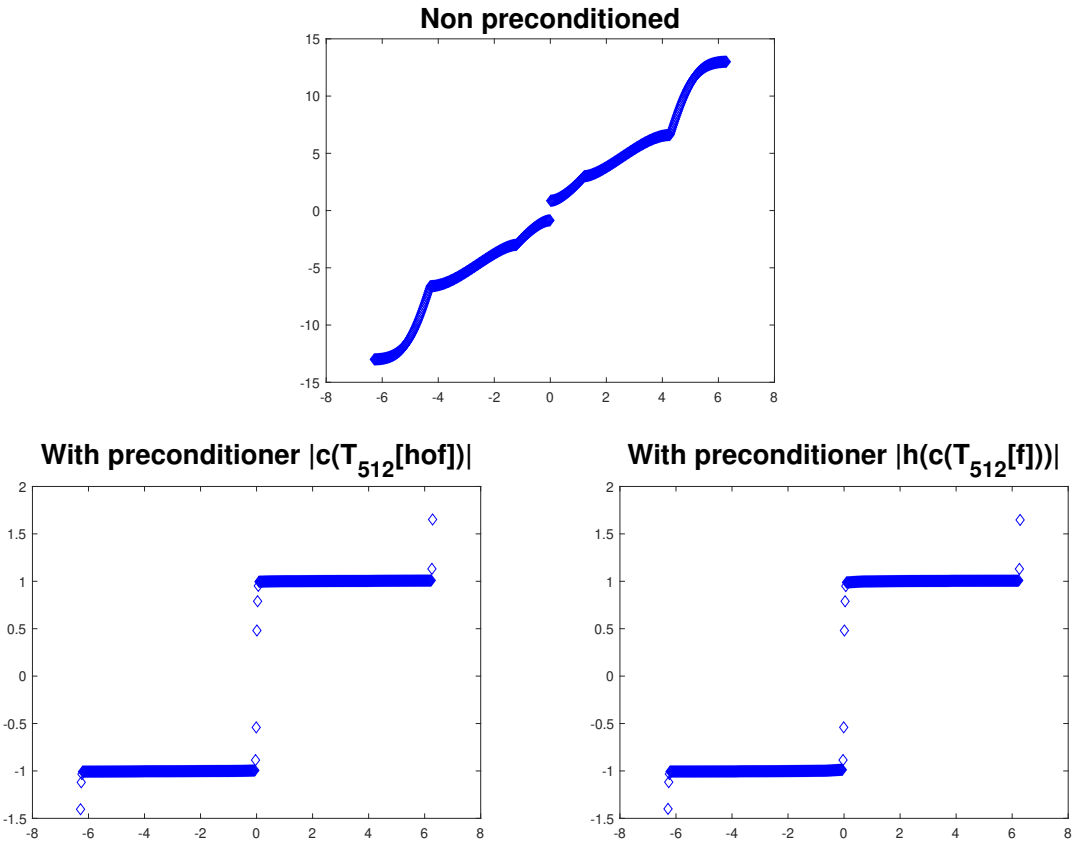

Figure III.8: The spectrum of the symmetrized matrix $Y_{512}h(T_{512}[f])$, for $h(z) = 1 + z + z^2$ and $f(\vartheta) = -\mathrm{e}^{\hat{\imath}\vartheta} + 1 + \mathrm{e}^{-\hat{\imath}\vartheta} + \mathrm{e}^{-\hat{\imath}2\vartheta} + \mathrm{e}^{-\hat{\imath}3\vartheta}$. Top: without preconditioner, bottom left: preconditioner $P_n = |c(T_n[h \circ f])|$, bottom right: preconditioner $P_n = |h(c(T_n[f]))|$.

strategy approach introduced in [78], that is, $P_n = |h(c(T_n[f]))|$. We can see the efficiency of the proposed strategy in the right panel of Figure III.9, where we observe that the eigenvalues of the preconditioned matrix $P_n^{-1}Y_n h(T_n[f])$, for $n = 100$ are clustered around -1 and 1, up to 2 outliers. Analogously, we can study the eigenvalues of the preconditioned matrix $P_{100}^{-1}Y_{100}T_{100}[f]$, where $P_{100} = |c(T_{100}[h \circ f])|$. Indeed, we have $h \circ f \in L^1([-\pi, \pi])$, then, applying the results in Section II.2, we have that P_{100} is a valid preconditioner for the matrix $Y_{100}h(T_{100}[f])$. The left panel of Figure III.9 confirms that the eigenvalues of the preconditioned matrix $P_{100}^{-1}Y_{100}h(T_{100}[f])$ are clustered around -1 and 1 up to 2 outliers.

For each example, we showed the validity of two different preconditioning strategies. However, we have seen that, for large enough matrix-sizes, the spectral results are remarkably similar. Other valid choices of preconditioning that give a slightly different effect on the spectrum of the preconditioned matrix can be considered. Moreover, we highlight that the strategy based on Theorem II.2.1 provides an entire class of preconditioners suitable for symmetrized Toeplitz structure functions. Indeed, a preconditioner in this class is the absolute value of any circulant matrix C_n such that the following singular value distribution is verified

$$\{C_n^{-1}T_n[h \circ f]\}_n \sim_\sigma 1. \tag{III.10}$$

Concerning the choice of the preconditioning strategy based on this requirement, we used the

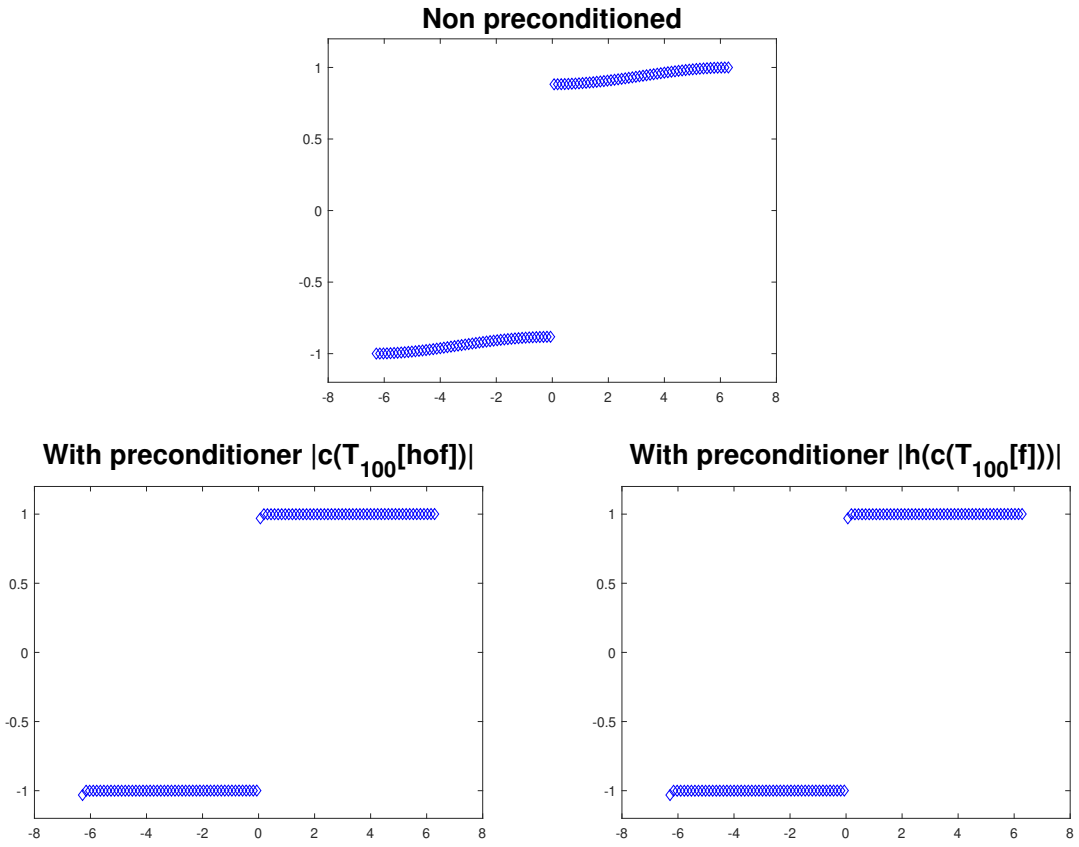

Figure III.9: The spectrum of the symmetrized matrix $Y_{100}h(T_{100}[f])$, for $h(z) = \mathrm{e}^z$ and $f(\vartheta) = \sum_{j=-99}^{99} \hat{f}_j \mathrm{e}^{\hat{\imath} j \vartheta}$, with $\lambda = 0.1$, $\mu = -0.9$, $\nu = 0.25$, $\sigma = 0.45$, $r = 0.05$, and $\Delta x = \frac{4}{101}$. Top: without preconditioner, bottom left: preconditioner $P_n = |c(T_n[h \circ f])|$, bottom right: preconditioner $P_n = |h(c(T_n[f]))|$.

Frobenius optimal circulant preconditioner, since, from the properties of the considered f and h, relation (III.10) is satisfied.

Finally, we highlight that the choice of the best preconditioning strategy among the two approaches that we analysed in the examples depends on the computational aspects in constructing the matrix P_n, which depend in turn on the information known about the specific example. For instance, the computational cost of the construction of the preconditioner $P_n = |c(T_n[h \circ f])|$ decreases if the Fourier coefficients of $h \circ f$ are known.

Chapter IV

Multigrid Methods for Block-Toeplitz Linear Systems

In the present chapter, we are mainly interested in solving large positive definite linear systems that possess a block-Toeplitz structure up to a low rank correction. Such systems are of great interest in many applications, such as numerical approximations of constant coefficient PDEs and coupled systems of integro-differential equations [36, 94].

After the seminal papers on multigrid methods for Toeplitz structures [27, 56, 57], the results have been extended to multidimensional problems and a V-cycle convergence analysis has been provided, see [4] and references therein. A multigrid method for block-Toeplitz matrices has been proposed in [83] and studied in the case of diagonal block generating functions. This was then adapted and further analysed for specific applications, like those considered in [38, 43], but the results are strictly related to the (multilevel) block-Toeplitz matrices in question. In practice, when the block symbol is not diagonal, there is still a substantial lack of an effective projection proposal and of a rigorous convergence analysis.

In this chapter we aim to fill this gap generalizing the existing convergence results in the scalar settings for linear systems with coefficient matrix in the circulant algebra associated with a matrix-valued symbol. According to the classical Ruge and Stüben convergence analysis in [107], we split the two-grid convergence analysis into the validation of a smoothing property and an approximation property. The smoothing property is proved for damped Jacobi with the relaxation parameter appropriately chosen in an interval depending on the symbol. For the proof of the approximation property, we provide a general theorem concerning the boundedness of a specific matrix-valued function $R(\vartheta)$ that depends both on the problem and on the grid transfer operator. However, the latter result is not straightforward to exploit for practical applications. A closer look at the derived condition on $R(\vartheta)$ highlights that the matrix-valued trigonometric polynomial that generates the block-circulant matrix used in the construction of the grid transfer operator needs to fulfil stricter conditions than the ones present in the scalar case. More specifically, we analyse a first case where the trigonometric polynomial in question is unitarily diagonalizable at all points and satisfies a specific commutativity condition. Moreover, we prove the approximation property for a grid transfer operator with a block symbol that might be non-diagonalizable.

We highlight that the theoretical analysis is performed for block-circulant matrices, in order

to exploit their intrinsic algebra structure, but the results can be forwarded to Toeplitz matrices. Indeed, this extension is possible thanks to the proof that the symbol analysis for Toeplitz matrices is an algebraic generalization of the local Fourier analysis of multigrid methods, see [37]. Moreover, if the block-Toeplitz matrices are banded, that is, they are generated by a matrix-valued trigonometric polynomial, they represent a low rank correction of block-circulant matrices with the same generating function. The latter consideration implies that the deterioration of the convergence rate is computationally acceptable, as we see through numerical examples in the next chapter.

The contents of the present chapter are in the process of being published in [20, 39] and the chapter is organized as follows. In Section IV.1 we give an overview on algebraic multigrid methods for circulant and Toeplitz matrices. In Section IV.2 we sketch the basic ideas for defining projecting operators for block-circulant matrices and we report two suitable sets of conditions on the matrix-valued trigonometric polynomial associated to the grid transfer operator. A convergence and optimality proof of the two-grid technique for both cases is reported in Section IV.3. Finally, in Section IV.4 we provide the generalization of the convergence results to multilevel block-circulant matrices, where the multilevel grid transfer operator possesses a tensor structure.

IV.1 Multigrid Methods for Toeplitz Matrices

As we already stated in Subsection I.9.4, the convergence analysis of the two-grid method splits into the validation of two separate conditions: the smoothing property and the approximation property. Regarding the latter, with reference to scalar structured matrices [4, 56], the optimality of two-grid methods is given in terms of choosing the proper conditions that the symbol p of a family of projection operators has to fulfil. Indeed, consider the Toeplitz matrix $T_n[f]$ with $n = (2^t - 1)$ generated by a non-negative trigonometric polynomial f. Suppose that f vanishes at exactly one point, which implies that the Toeplitz matrix $T_n[f]$ becomes ill–conditioned as n increases. Let ϑ_0 be the unique zero of f. Then, the optimality of the two-grid method applied to $T_n[f]$ is guaranteed if we choose a family of projection operators associated with a symbol p such that

$$\begin{aligned} &\limsup_{\vartheta\to\vartheta_0} \frac{|p(\eta)|^2}{f(\vartheta)} < \infty, \quad \eta \in \mathcal{M}(\vartheta), \\ &\sum_{\eta\in\Omega(\vartheta)} p^2(\eta) > 0, \end{aligned} \tag{IV.1}$$

where the sets $\Omega(\vartheta)$ and $\mathcal{M}(\vartheta)$ are the following corner and mirror points

$$\Omega(\vartheta) = \{\vartheta, (\vartheta + \pi) \mod 2\pi\}, \qquad \mathcal{M}(\vartheta) = \Omega(\vartheta) \setminus \{\vartheta\},$$

respectively.

Informally, the latter conditions mean that the optimality of the two-grid method is obtained by choosing the family of grid transfer operators associated to a symbol p such that $|p|^2(\vartheta) + |p|^2(\vartheta + \pi)$ does not have zeros and $|p|^2(\vartheta + \pi)/f(\vartheta)$ is bounded. For achieving the optimality of the V-cycle method, the second condition needs to be strengthened, see [4] for details.

As far as the smoothing property is concerned, a lot of results are present in the relevant literature for different stationary methods. See, for instance, [56] for an analysis of the best choice of the relaxation parameter for the relaxed Richardson method.

IV.2 Projecting Operators for Block-Circulant Matrices

The current section is dedicated to the construction of grid transfer operators suitable for block-circulant matrices. Indeed, as we outlined in the previous section, the choice of prolongation and restriction operators fulfilling the approximation property that we introduced in Subsection I.9.4 is crucial for multigrid convergence and optimality. In particular, the projector $P_{n,m}$ should be chosen in order that it projects the problem onto a coarser space by "cutting" the coefficient matrix and the resulting projected matrix should maintain the same block structure and properties of original matrix.

Let $K_{n,m}$ be the $n \times m$ down-sampling matrix, that is,

- when n is even: $m = \frac{n}{2}$ and $K_{n,m} = K_{n,m}^{Odd}$,
- when n is odd: $m = \frac{n-1}{2}$ and $K_{n,m} = K_{n,m}^{Even}$,

with $K_{n,m}^{Odd}$ and $K_{n,m}^{Even}$ defined as

$$K_{n,m}^{Odd} = \begin{bmatrix} 1 & & \\ 0 & & \\ \vdots & 1 & \\ & 0 & \\ & \vdots & \\ & & \ddots \\ & & 1 \\ & & 0 \end{bmatrix}_{n \times m}, \qquad K_{n,m}^{Even} = \begin{bmatrix} 0 & & \\ 1 & & \\ 0 & 0 & \\ \vdots & 1 & \\ & 0 & \\ & \vdots & \\ & & \ddots \\ & & 1 \\ & & 0 \end{bmatrix}_{n \times m}.$$

In particular, $K_{n,m}^{Odd}$ is the $n \times m$ matrix obtained by removing the even rows from the identity matrix of size n, that is it keeps the odd rows. On the other hand, $K_{n,m}^{Even}$ keeps the even rows. When n is even, $K_{n,m}^{Odd}$ performs the following manipulation of the Fourier frequencies:

$$(K_{n,m}^{Odd})^T F_n = \frac{1}{\sqrt{2}}[F_m \,|\, F_m]. \tag{IV.2}$$

This property of the Fourier matrix is the key to define a grid transfer operator $P_{n,m}^s$ that preserves the block-circulant structure at the coarser levels, where the superscript s indicates that the block-structured matrices have square blocks of size s.

Therefore, we define the structure of the grid transfer operators $P_{n,m}^s$ for the block-circulant matrix $\mathcal{C}_n[\mathbf{f}]$ generated by a matrix-valued trigonometric polynomial $\mathbf{f} : [-\pi, \pi] \to \mathbb{C}^{s \times s}$ as follows. Let n be even and of the form 2^t, $t \in \mathbb{N}$, such that the size of the coarser problem is $m = \frac{n}{2} = 2^{t-1}$. The projector $P_{n,m}^s$ is then constructed as the product between a matrix $\mathcal{C}_n[\mathbf{p}]$

in the circulant algebra, with $\mathbf{p}$ being a proper trigonometric polynomial that are discussed in the following subsections, and a cutting matrix $K_{n,m}^{Odd} \otimes I_s$. That is,

$$P_{n,m}^s = \mathcal{C}_n[\mathbf{p}](K_{n,m}^{Odd} \otimes I_s). \tag{IV.3}$$

The result of multiplying a $s \times s$ block matrix of dimension $sn \times sn$ by $K_{n,m}^{Odd} \otimes I_s$ is a $s \times s$ block matrix where just the even "block-columns" are maintained.

We are left to determine the conditions to be satisfied by $\mathcal{C}_n[\mathbf{p}]$ (or better by its generating function $\mathbf{p}$), in order to obtain a projector which is effective in terms of convergence. The used tool is an algebraic generalization of the Local Fourier Analysis of multigrid methods [37].

The same strategy can be applied when we deal with block-Toeplitz matrices generated by a matrix-valued trigonometric polynomial, instead of block-circulant matrices. Indeed, the only thing that should be adapted is the structure of the projector which slightly changes for block-Toeplitz matrices, in order to preserve the structure at coarser levels.

Hence, for a matrix-valued trigonometric polynomial $\mathbf{p}$, the projector matrix is

$$P_{n,m}^s = T_n[\mathbf{p}] \left(K_{n,m}^{Even} \otimes I_s\right). \tag{IV.4}$$

Note that in the Toeplitz case n should be chosen odd and of the form $2^t - 1$, $t \in \mathbb{N}$, such that the size of the coarser problem is $m = \frac{n-1}{2} = 2^{t-1} - 1$.

Finally, we mention that it is possible to consider the case where the size n of the coefficient matrix is divisible by a factor $g \gneq 2$ such that at the lower level the system is reduced to one of size n/g. Indeed, in this situation we can exploit a g-circulant based projectors [44, 100]. In particular, we can analogously repeat the TGM convergence result adopting a cutting matrix $(K_{n,n/g}) \otimes I_s$, where $K_{n,n/g} \in \mathbb{R}^{n \times n/g}$, of the form

$$K_{n,n/g} = [\delta_{i-gj}]_{i,j}, \quad i = 0, \ldots, n-1;\ j = 0, \ldots, n/g - 1, \qquad \delta_\ell = \begin{cases} 1 & \text{if } \ell \equiv 0 \,(\mathrm{mod}\, n), \\ 0 & \text{otherwise} \end{cases}.$$

IV.2.1 TGM Conditions: the Diagonalizable Case

Let $\mathcal{C}_n[\mathbf{f}]$ be the block-circulant matrix generated by a matrix-valued trigonometric polynomial $\mathbf{f} \geq 0$ and let us consider the grid transfer operator $P_{n,m}^s = \mathcal{C}_n[\mathbf{p}](K_{n,m}^{Odd} \otimes I_s)$, with $\mathbf{p}$ being a matrix-valued trigonometric polynomial. In the following, we provide and discuss a set of conditions on $\mathbf{p}$.

Define Θ_0 as the set of points ϑ such that $\lambda_j(\mathbf{f}(\vartheta)) = 0$ for some j. Assume that, for $\vartheta \in \Theta_0$, $\lambda_j(\mathbf{f}(\vartheta + \pi)) \neq 0$ for all $j = 1, \ldots, s$, which also implies that the set Θ_0 is a finite set. Choose $\mathbf{p}(\cdot)$ diagonalizable by a unitary matrix such that the following relations

$$\exists \delta \text{ s.t. } \left\| \mathbf{f}(\vartheta)^{-\frac{1}{2}} \mathbf{p}(\vartheta + \pi)^H \right\|_1 < \delta \quad \forall \vartheta \in [0, 2\pi) \backslash \Theta_0, \tag{IV.5}$$

$$\mathbf{p}(\vartheta)^H \mathbf{p}(\vartheta) + \mathbf{p}(\vartheta + \pi)^H \mathbf{p}(\vartheta + \pi) > 0 \quad \forall \vartheta \in [0, 2\pi), \tag{IV.6}$$

$$\mathbf{p}(\vartheta)\mathbf{p}(\vartheta + \pi) = \mathbf{p}(\vartheta + \pi)\mathbf{p}(\vartheta) \quad \forall \vartheta \in [0, 2\pi) \tag{IV.7}$$

are fulfilled. Note that conditions (IV.5)–(IV.6) are the generalization of the scalar conditions (IV.1), while condition (IV.7) is new and it permits to simplify some expressions, as we explain in the following remark.

Remark 1. *By hypothesis we have that there exist a unitary transform $U(\cdot)$ and a diagonal matrix-valued function $D_p(\cdot)$ such that*

$$\mathbf{p}(\vartheta) = U(\vartheta)D_{\mathbf{p}}(\vartheta)U(\vartheta)^H \qquad \text{and} \qquad \mathbf{p}(\vartheta+\pi) = U(\vartheta+\pi)D_{\mathbf{p}}(\vartheta+\pi)U(\vartheta+\pi)^H.$$

Note that condition (IV.7) on the commutativity of $\mathbf{p}(\vartheta)$ and $\mathbf{p}(\vartheta+\pi)$ implies that they are simultaneously diagonalizable. Then,

$$\mathbf{p}(\vartheta+\pi) = U(\vartheta)D_{\mathbf{p}}(\vartheta+\pi)U(\vartheta)^H$$

and in particular, we have

$$(\mathbf{p}(\vartheta)^H\mathbf{p}(\vartheta) + \mathbf{p}(\vartheta+\pi)^H\mathbf{p}(\vartheta+\pi))^{-1} = U(\vartheta)(|D_{\mathbf{p}}(\vartheta)|^2 + |D_{\mathbf{p}}(\vartheta+\pi)|^2)^{-1}U(\vartheta)^H,$$

which ensures that $(\mathbf{p}(\vartheta)^H\mathbf{p}(\vartheta) + \mathbf{p}(\vartheta+\pi)^H\mathbf{p}(\vartheta+\pi))^{-1}$ commutes with $\mathbf{p}(\vartheta)$, $\mathbf{p}(\vartheta)^H$, $\mathbf{p}(\vartheta+\pi)$ and $\mathbf{p}(\vartheta+\pi)^H$.

IV.2.2 TGM Conditions: the General Case

Let $\mathcal{C}_n[\mathbf{f}]$ be the block-circulant matrix generated by a matrix-valued trigonometric polynomial $\mathbf{f} \geq 0$ and let us consider the grid transfer operator $P^s_{n,m} = \mathcal{C}_n[\mathbf{p}](K^{Odd}_{n,m} \otimes I_s)$, with $\mathbf{p}$ being a matrix-valued trigonometric polynomial. Suppose that there exist unique $\vartheta_0 \in [0, 2\pi)$ and $\bar{\jmath} \in \{1, \ldots, s\}$ such that

$$\begin{cases} \lambda_j(\mathbf{f}(\vartheta)) = 0, & \text{for } \vartheta = \vartheta_0 \text{ and } j = \bar{\jmath}, \\ \lambda_j(\mathbf{f}(\vartheta)) > 0, & \text{otherwise.} \end{cases} \tag{IV.8}$$

The latter assumption means that the matrix $\mathbf{f}(\vartheta)$ has exactly one zero eigenvalue in ϑ_0 and it is positive definite in $[0, 2\pi)\backslash\{\vartheta_0\}$.

As a consequence, the matrices $\mathcal{C}_n[\mathbf{f}]$ could be singular. On the other hand, the block-Toeplitz matrices $T_n[\mathbf{f}]$ are positive definite, they become ill-conditioned as n increases, and the ill-conditioned subspace is the eigenspace associated with $\lambda_{\bar{\jmath}}(\mathbf{f}(\vartheta_0))$.

Since $\mathbf{f}(\vartheta)$ is Hermitian, it can be diagonalized by an orthogonal matrix $Q(\vartheta)$. Hence,

$$\mathbf{f}(\vartheta) = Q(\vartheta)D(\vartheta)Q(\vartheta)^H =$$

$$\left[\, q_1(\vartheta) \,\middle|\, \ldots \,\middle|\, q_{\bar{\jmath}}(\vartheta) \,\middle|\, \ldots \,\middle|\, q_s(\vartheta) \,\right] \begin{bmatrix} \lambda_1(\mathbf{f}(\vartheta)) & & & & \\ & \ddots & & & \\ & & \lambda_{\bar{\jmath}}(\mathbf{f}(\vartheta)) & & \\ & & & \ddots & \\ & & & & \lambda_s(\mathbf{f}(\vartheta)) \end{bmatrix} \begin{bmatrix} \overline{q_1{}^H(\vartheta)} \\ \vdots \\ \overline{q_{\bar{\jmath}}^H(\vartheta)} \\ \vdots \\ \overline{q_s{}^H(\vartheta)} \end{bmatrix} \tag{IV.9}$$

where $q_{\bar{\jmath}}(\vartheta)$ is the eigenvector that generates the ill-conditioned subspace since $q_{\bar{\jmath}}(\vartheta_0)$ is the eigenvector of $\mathbf{f}(\vartheta_0)$ associated with $\lambda_{\bar{\jmath}}(\mathbf{f}(\vartheta_0)) = 0$.

Under the following assumptions, we show that there are sufficient conditions to ensure the linear convergence of the two-grid method. Indeed, it is sufficient to choose $\mathbf{p}$ such that

(i) condition (IV.6) is fulfilled, that is,

$$\mathbf{p}(\vartheta)^H\mathbf{p}(\vartheta) + \mathbf{p}(\vartheta+\pi)^H\mathbf{p}(\vartheta+\pi) > 0 \quad \forall\vartheta \in [0, 2\pi),$$

which implies that the trigonometric function

$$\mathbf{r}(\vartheta) = \mathbf{p}(\vartheta)\left(\mathbf{p}(\vartheta)^H\mathbf{p}(\vartheta) + \mathbf{p}(\vartheta+\pi)^H\mathbf{p}(\vartheta+\pi)\right)^{-1}\mathbf{p}(\vartheta)^H \tag{IV.10}$$

is well-defined for all $\vartheta \in [0, 2\pi)$;

(ii) the vector $q_{\bar{\jmath}}(\vartheta_0)$ defined in (IV.9) is an eigenvector for $\mathbf{r}(\vartheta_0)$ with eigenvalue 1, that is,

$$\mathbf{r}(\vartheta_0)q_{\bar{\jmath}}(\vartheta_0) = q_{\bar{\jmath}}(\vartheta_0);$$

(iii) it holds that

$$\limsup_{\vartheta\to\vartheta_0} \lambda_{\bar{\jmath}}(\mathbf{f}(\vartheta))^{-1}(1 - \lambda_{\bar{\jmath}}(\mathbf{r}(\vartheta))) = c,$$

where $c \in \mathbb{R}$ is a constant.

IV.3 Proofs of Convergence

The current section is outlined as follows. Firstly, we give a result on the validation of the smoothing property in a specific setting. Then, we focus on preliminary results concerning the grid transfer operators and the validation of the approximation property. Further, in Subsections IV.3.1–IV.3.2 we prove the convergence and optimality of the two-grid method in the setting of Subsections IV.2.1 and IV.2.2 respectively.

The smoothing property has been proven in [38] for the simple Richardson iteration considering both pre-smoothing and post-smoothing.

Lemma IV.3.1 ([38]). *Let $\mathcal{C}_n[\mathbf{f}]$ with $\mathbf{f} = [\mathbf{f}_{\ell,g}]_{\ell,g=1}^s \in \mathbb{C}^{s\times s}$ trigonometric polynomial, $\mathbf{f} \geq 0$, with $f_{j,j}$, $j = 1, \ldots, s$, not identically zero, and let $V_n := I_{sn} - \omega\mathcal{C}_n[\mathbf{f}]$. If we choose $\omega \in (0, 2/\|\mathbf{f}\|_\infty)$, then relation ($a$) in Theorem I.9.2 holds true.*

The iteration matrix of the relaxed Jacobi method is $V_n := I_{sn} - \omega D_n^{-1}\mathcal{C}_n[\mathbf{f}]$, where D_n is a diagonal matrix with the same diagonal as $\mathcal{C}_n[\mathbf{f}]$. We define the matrix $\tilde{D}_n := \min_{j=1,\ldots,s}\left[\hat{\mathbf{f}}_0\right]_{(j,j)} I_{sn}$ and we notice that $\tilde{D}_n^{-1} \geq D_n^{-1}$. Applying to the matrix $I_{sn} - \omega\tilde{D}_n^{-1}\mathcal{C}_n[\mathbf{f}]$ the same idea of the proof used for the Richardson method in [38, Proposition 4], we obtain that relation (a) in Theorem I.9.2 is satisfied if ω verifies the following inequality:

$$0 \leq \omega \leq \frac{\min_{j=1,\ldots,s}\left[\hat{\mathbf{f}}_0\right]_{(j,j)}}{\|\mathbf{f}\|_\infty}. \tag{IV.11}$$

Before discussing the details on the approximation property, we consider a crucial result both from a theoretical and a practical point of view.

Proposition IV.3.2. *Let $\mathbf{f}$ be a non-negative definite $s \times s$ matrix-valued function. Let $P^s_{n,m}$ and $K^{Odd}_{n,m}$ be defined as in Section IV.2, that is, $P^s_{n,m} = \mathcal{C}_n[\mathbf{p}](K^{Odd}_{n,m} \otimes I_s) \in \mathbb{C}^{sn \times sm}$, where $\mathbf{p}$ is a trigonometric polynomial. Then the matrix $(P^s_{n,m})^H \mathcal{C}_n[\mathbf{f}] P^s_{n,m} \in \mathbb{C}^{sm \times sm}$ coincides with $\mathcal{C}_m(\hat{\mathbf{f}})$ where $\hat{\mathbf{f}}$ is non-negative definite and*

$$\hat{\mathbf{f}}(\vartheta) = \frac{1}{2}\left(\mathbf{p}\left(\frac{\vartheta}{2}\right)^H \mathbf{f}\left(\frac{\vartheta}{2}\right)\mathbf{p}\left(\frac{\vartheta}{2}\right) + \mathbf{p}\left(\frac{\vartheta}{2}+\pi\right)^H \mathbf{f}\left(\frac{\vartheta}{2}+\pi\right)\mathbf{p}\left(\frac{\vartheta}{2}+\pi\right)\right). \tag{IV.12}$$

Proof. Using the definition of $P^s_{n,m}$, we have that

$$\begin{aligned}
(P^s_{n,m})^H \mathcal{C}_n[\mathbf{f}] P^s_{n,m} &= ((K^{Odd}_{n,m})^T \otimes I_s)\mathcal{C}_n[\mathbf{p}^H]\mathcal{C}_n[\mathbf{f}]\mathcal{C}_n[\mathbf{p}](K^{Odd}_{n,m} \otimes I_s) \\
&= ((K^{Odd}_{n,m})^T \otimes I_s)\mathcal{C}_n[\mathbf{p}^H\mathbf{f}\mathbf{p}](K^{Odd}_{n,m} \otimes I_s) \\
&= ((K^{Odd}_{n,m})^T \otimes I_s)(F_n \otimes I_s) \operatorname*{diag}_{i\in\mathcal{I}_n}(\mathbf{p}^H\mathbf{f}\mathbf{p}(\vartheta^{(n)}_i))(F_n^H \otimes I_s)(K^{Odd}_{n,m} \otimes I_s) \\
&= \frac{1}{2}(F_m \otimes I_s) \operatorname*{diag}_{i\in\mathcal{I}_m}\left(\mathbf{p}^H\mathbf{f}\mathbf{p}(\vartheta^{(n)}_i\right) + \mathbf{p}^H\mathbf{f}\mathbf{p}(\vartheta^{(n)}_{\tilde{i}}))(F_m^H \otimes I_s) \\
&= \frac{1}{2}(F_m \otimes I_s) \operatorname*{diag}_{i\in\mathcal{I}_m}\left(\mathbf{p}^H\mathbf{f}\mathbf{p}\left(\frac{\vartheta^{(m)}_i}{2}\right) + \mathbf{p}^H\mathbf{f}\mathbf{p}\left(\frac{\vartheta^{(m)}_i}{2}+\pi\right)\right)(F_m^H \otimes I_s) \\
&= \mathcal{C}_m\left[\hat{\mathbf{f}}\right],
\end{aligned}$$

where $\tilde{i} = i + m$; this is again a block-circulant matrix of size sm. From the structure of $\hat{\mathbf{f}}$ it is clear that if $\mathbf{f}$ is non-negative definite also $\hat{\mathbf{f}}$ is non-negative definite. □

In the following theorem we give the main result on the validation of the approximation property, which involves the boundedness of a matrix-valued function $R(\vartheta)$ that depends both on the problem and on the grid transfer operator.

Theorem IV.3.3. *Let $\mathcal{C}_n[\mathbf{f}]$, with $\mathbf{f}(\vartheta) \in \mathbb{C}^{s\times s}$ trigonometric polynomial, $\mathbf{f} \geq 0$, and define Θ_0 as the set of points ϑ such that $\lambda_j(\mathbf{f}(\vartheta)) = 0$ for some j and define $H = \{\eta | \eta \in \{\vartheta, (\vartheta+\pi) \bmod 2\pi\}, \vartheta \in \Theta_0\}$. Assume that, for $\vartheta \in \Theta_0$, $\lambda_j(\mathbf{f}(\vartheta+\pi)) \neq 0$ for all $j = 1, \ldots, s$. Let $P^s_{n,m} = \mathcal{C}_n[\mathbf{p}](K^{Odd}_n \otimes I_s)$ be a projecting operator. Suppose that $\mathbf{p}$ is a matrix-valued trigonometric polynomial that fulfils condition (IV.6) and there exists $c > 0$ such that for all $\vartheta \in [0, 2\pi)\backslash H$*

$$R(\vartheta) \leq cI_{2s} \tag{IV.13}$$

with

$$R(\vartheta) = \begin{bmatrix}\mathbf{f}(\vartheta) & \\ & \mathbf{f}(\vartheta+\pi)\end{bmatrix}^{-\frac{1}{2}} \left(I_{2s} - \begin{bmatrix}\mathbf{p}(\vartheta) \\ \mathbf{p}(\vartheta+\pi)\end{bmatrix}\mathbf{q}(\vartheta)\left[\mathbf{p}(\vartheta)^H\mathbf{p}(\vartheta+\pi)^H\right]\right)\begin{bmatrix}\mathbf{f}(\vartheta) & \\ & \mathbf{f}(\vartheta+\pi)\end{bmatrix}^{-\frac{1}{2}}$$

and

$$\mathbf{q}(\vartheta) = \left(\mathbf{p}(\vartheta)^H\mathbf{p}(\vartheta) + \mathbf{p}(\vartheta+\pi)^H\mathbf{p}(\vartheta+\pi)\right)^{-1}. \tag{IV.14}$$

Then, there exists a positive value γ independent of n such that inequality (b) in Theorem I.9.2 is satisfied.

Proof. In order to prove that there exists $\gamma > 0$ independent of n such that for any $x_n \in \mathbb{C}^{sn}$

$$\min_{y\in\mathbb{C}^{sm}} \|x_n - P^s_{n,m}y\|_2^2 \leq \gamma \|x_n\|^2_{\mathcal{C}_n[\mathbf{f}]}, \tag{IV.15}$$

we choose a special instance of y in such a way that the previous inequality is reduced to a matrix inequality in the sense of the partial ordering of the real space of the Hermitian matrices. For any $x_n \in \mathbb{C}^{sn}$, let $\overline{y} \equiv \overline{y}(x_n) \in \mathbb{C}^{sm}$ be defined as

$$\overline{y} = [(P^s_{n,m})^H P^s_{n,m}]^{-1} (P^s_{n,m})^H x_n.$$

We observe that $(P^s_{n,m})^H P^s_{n,m}$ is invertible, indeed, using the same arguments of Proposition IV.3.2 with $\mathbf{f} = I_s$, we have that $(P^s_{n,m})^H P^s_{n,m} = \mathcal{C}_m[\hat{\mathbf{p}}]$ with

$$\hat{\mathbf{p}}(\vartheta) = \frac{1}{2}\left(\mathbf{p}\left(\frac{\vartheta}{2}\right)^H \mathbf{p}\left(\frac{\vartheta}{2}\right) + \mathbf{p}\left(\frac{\vartheta}{2}+\pi\right)^H \mathbf{p}\left(\frac{\vartheta}{2}+\pi\right)\right)$$

and condition (IV.6) ensure that $\hat{\mathbf{p}} > 0$, that is $\mathcal{C}_m[\hat{\mathbf{p}}]$ is positive definite.

Therefore, (IV.15) is implied by

$$\|x_n - P^s_{n,m}\overline{y}\|_2^2 \le \gamma \|x_n\|^2_{\mathcal{C}_n[\mathbf{f}]},$$

where the latter is equivalent to the matrix inequality

$$W_n(\mathbf{p})^H W_n(\mathbf{p}) \le \gamma \mathcal{C}_n[\mathbf{f}].$$

with $W_n(\mathbf{p}) = I_{sn} - P^s_{n,m}[(P^s_{n,m})^H P^s_{n,m}]^{-1}(P^s_{n,m})^H$. Since, by construction, $W_n(\mathbf{p})$ is a Hermitian unitary projector, it holds that $W_n(\mathbf{p})^H W_n(\mathbf{p}) = W_n(\mathbf{p})^2 = W_n(\mathbf{p})$. As a consequence, the preceding matrix inequality can be rewritten as

$$W_n(\mathbf{p}) \le \gamma \mathcal{C}_n[\mathbf{f}]. \tag{IV.16}$$

Now, using the expression of the matrix $P^s_{n,m} = \mathcal{C}_n[\mathbf{p}](K_n^{Odd} \otimes I_s)$ and the relation (IV.2), we write $(P^s_{n,m})^H$ as

$$\begin{aligned}(P^s_{n,m})^H &= \frac{1}{\sqrt{2}}(F_m \otimes I_s)(I_{n,2} \otimes I_s) \operatorname*{diag}_{i\in\mathcal{I}_n}(\mathbf{p}(\vartheta_i^{(n)})^H)(F_n^H \otimes I_s) \\ &= \frac{1}{\sqrt{2}}(F_m \otimes I_s)\left[\operatorname*{diag}_{i\in\mathcal{I}_m}(\mathbf{p}(\vartheta_i^{(n)})^H) \mid \operatorname*{diag}_{i\in\mathcal{I}_m}(\mathbf{p}(\vartheta_{i+m}^{(n)})^H)\right](F_n^H \otimes I_s),\end{aligned}$$

where $I_{n,2} = [I_m | I_m]_{m\times n}$. Then,

$$(P^s_{n,m})^H P^s_{n,m} = \frac{1}{2}(F_m \otimes I_r)\left[\operatorname*{diag}_{i\in\mathcal{I}_m}\left(p(\vartheta_i^{(n)})^H p(\vartheta_i^{(n)})\right) + \operatorname*{diag}_{i\in\mathcal{I}_m}\left(p(\vartheta_{i+m}^{(n)})^H p(\vartheta_{i+m}^{(n)})\right)\right](F_m^H \otimes I_r).$$

Hence, the matrix $(F_n^H \otimes I_s)W_n(\mathbf{p})(F_n \otimes I_s)$ becomes

$$\begin{aligned}&(F_n^H \otimes I_s)W_n(\mathbf{p})(F_n \otimes I_s) \\ &\quad = I_{sn} - \operatorname*{diag}_{i\in\mathcal{I}_n}\left(\mathbf{p}(\vartheta_i^{(n)})\right)(I_{n,2}^T \otimes I_s)\left[\operatorname*{diag}_{i\in\mathcal{I}_m}\left(\mathbf{p}(\vartheta_i^{(n)})^H\mathbf{p}(\vartheta_i^{(n)}) + \mathbf{p}(\vartheta_{\tilde{i}}^{(n)})^H\mathbf{p}(\vartheta_{\tilde{i}}^{(n)})\right)\right]^{-1} \\ &\qquad\qquad \cdot (I_{n,2} \otimes I_s)\operatorname*{diag}_{i\in\mathcal{I}_n}\left(\mathbf{p}(\vartheta_i^{(n)})^H\right)\end{aligned} \tag{IV.17}$$

where $\tilde{i} = i + m$. Now, it is clear that there exists a suitable permutation by rows and columns of $(F_n^H \otimes I_s)W_n(\mathbf{p})(F_n \otimes I_s)$ such that we can obtain a $2s \times 2s$ block-diagonal matrix of the form

$$I_{sn} - \underset{i\in\mathcal{I}_m}{\operatorname{diag}} \begin{bmatrix} \mathbf{p}(\vartheta_i^{(n)}) \\ \mathbf{p}(\vartheta_{\tilde{i}}^{(n)}) \end{bmatrix} \begin{bmatrix} (\mathbf{p}(\vartheta_i^{(n)})^H\mathbf{p}(\vartheta_i^{(n)}) + \mathbf{p}(\vartheta_{\tilde{i}}^{(n)})^H\mathbf{p}(\vartheta_{\tilde{i}}^{(n)}))^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{p}(\vartheta_i^{(n)})^H & \mathbf{p}(\vartheta_{\tilde{i}}^{(n)})^H \end{bmatrix}.$$

Therefore, by considering the same permutation by rows and columns of $(F_n^H \otimes I_s)\mathcal{C}_n[\mathbf{f}].(F_n \otimes I_s) = \underset{i\in\mathcal{I}_n}{\operatorname{diag}}(\mathbf{f}(\vartheta_i^{(n)})$, condition (IV.16) is equivalent to requiring that there exists $c > 0$ independent of n such that, $\forall j = 0, \ldots, m-1$

$$I_{2s} - \begin{bmatrix} \mathbf{p}(\vartheta_i^{(n)}) \\ \mathbf{p}(\vartheta_{\tilde{i}}^{(n)}) \end{bmatrix} \begin{bmatrix} (\mathbf{p}(\vartheta_i^{(n)})^H\mathbf{p}(\vartheta_i^{(n)}) + \mathbf{p}(\vartheta_{\tilde{i}}^{(n)})^H\mathbf{p}(\vartheta_{\tilde{i}}^{(n)}))^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{p}(\vartheta_i^{(n)})^H & \mathbf{p}(\vartheta_{\tilde{i}}^{(n)})^H \end{bmatrix}$$
$$\leq c \begin{bmatrix} \mathbf{f}(\vartheta_i^{(n)}) & \\ & \mathbf{f}(\vartheta_{\tilde{i}}^{(n)}) \end{bmatrix}.$$

Due of the continuity of $\mathbf{p}$ and $\mathbf{f}$ it is clear that the preceding set of inequalities can be reduced to requiring that a unique inequality of the form

$$I_{2s} - \begin{bmatrix} \mathbf{p}(\vartheta) \\ \mathbf{p}(\vartheta+\pi) \end{bmatrix} \begin{bmatrix} (\mathbf{p}(\vartheta)^H\mathbf{p}(\vartheta) + \mathbf{p}(\vartheta+\pi)^H\mathbf{p}(\vartheta+\pi))^{-1} \end{bmatrix}$$
$$\begin{bmatrix} \mathbf{p}(\vartheta)^H & \mathbf{p}(\vartheta+\pi)^H \end{bmatrix} \leq c \begin{bmatrix} \mathbf{f}(\vartheta) & \\ & \mathbf{f}(\vartheta+\pi) \end{bmatrix}$$

holds for all $\vartheta \in [0, 2\pi)\backslash H$. By the Sylvester inertia law [65], the latter relation is satisfied if

$$R(\vartheta) \leq cI_{2s} \tag{IV.18}$$

for all $\vartheta \in [0, 2\pi)\backslash H$ and the proof is complete. □

IV.3.1 TGM Convergence and Optimality: the Diagonalizable Case

The current subsection is devoted to show that the setting in Subsection IV.2.1 is appropriate to obtain the TGM convergence and optimality. In particular, the following result shows that conditions (IV.5), (IV.6), and (IV.7) are sufficient in order to satisfy the approximation property.

Theorem IV.3.4. *Let $\mathcal{C}_n[\mathbf{f}]$, with $\mathbf{f}(\vartheta) \in \mathbb{C}^{s\times s}$ trigonometric polynomial, $\mathbf{f} \geq 0$, and Define Θ_0 as the set of points ϑ such that $\lambda_j(\mathbf{f}(\vartheta)) = 0$ for some j and define $H = \{\eta | \eta \in \{\vartheta, (\vartheta+\pi) \bmod 2\pi\}, \vartheta \in \Theta_0\}$. Assume that, for $\vartheta \in \Theta_0$, $\lambda_j(\mathbf{f}(\vartheta+\pi)) \neq 0$ for all $j = 1, \ldots, s$. Let $P_{n,m}^s = \mathcal{C}_n[\mathbf{p}](K_n^{Odd} \otimes I_s)$ be a projecting operator, where $\mathbf{p}(\vartheta)$ is a unitarily diagonalizable matrix-valued trigonometric polynomial satisfying conditions (IV.5), (IV.6) and (IV.7). Then, there exists a positive value γ independent of n such that inequality (b) in Theorem I.9.2 is satisfied.*

Proof. By Theorem IV.3.3, it is clear that it is sufficient to prove that there exists a constant $c > 0$ such that for all $\vartheta \in [0, 2\pi)\backslash H$

$$R(\vartheta) \leq cI_{2s}. \tag{IV.19}$$

By simple computations, using condition (IV.7) and Remark 1 the matrix-valued function $R(\vartheta)$ can be written as

$$R(\vartheta) = \begin{bmatrix} \mathbf{f}(\vartheta) & \\ & \mathbf{f}(\vartheta+\pi) \end{bmatrix}^{-\frac{1}{2}} \mathbf{q}(\vartheta) \begin{bmatrix} \mathbf{p}(\vartheta+\pi)^H\mathbf{p}(\vartheta+\pi) & -\mathbf{p}(\vartheta)\mathbf{p}(\vartheta+\pi)^H \\ -\mathbf{p}(\vartheta+\pi)\mathbf{p}(\vartheta)^H & \mathbf{p}(\vartheta)^H\mathbf{p}(\vartheta) \end{bmatrix} \begin{bmatrix} \mathbf{f}(\vartheta) & \\ & \mathbf{f}(\vartheta+\pi) \end{bmatrix}^{-\frac{1}{2}}. \tag{IV.20}$$

The relation (IV.19) is satisfied if the matrix-valued function $R(\vartheta)$ is uniformly bounded in the spectral norm, which can be obtained proving that all the components of $R(\vartheta)$ are uniformly bounded in the $1-$norm. Using again the commutativity hypothesis (IV.7), we can write $R(\vartheta)$ as

$$R(\vartheta) = \begin{bmatrix} \mathbf{f}^{-\frac{1}{2}}(\vartheta)\mathbf{p}(\vartheta+\pi)^H\mathbf{q}(\vartheta)\mathbf{p}(\vartheta+\pi)\mathbf{f}^{-\frac{1}{2}}(\vartheta) & -\mathbf{f}^{-\frac{1}{2}}(\vartheta)\mathbf{p}(\vartheta+\pi)^H\mathbf{q}(\vartheta)\mathbf{p}(\vartheta)\mathbf{f}^{-\frac{1}{2}}(\vartheta+\pi) \\ -\mathbf{f}^{-\frac{1}{2}}(\vartheta+\pi)\mathbf{p}(\vartheta)^H\mathbf{q}(\vartheta)\mathbf{p}(\vartheta+\pi)\mathbf{f}^{-\frac{1}{2}}(\vartheta) & \mathbf{f}^{-\frac{1}{2}}(\vartheta+\pi)\mathbf{p}(\vartheta)^H\mathbf{q}(\vartheta)\mathbf{p}(\vartheta)\mathbf{f}^{-\frac{1}{2}}(\vartheta+\pi) \end{bmatrix}.$$

For all $\vartheta \in [0, 2\pi)\backslash H$, we can write

$$\begin{aligned} \|R_{1,1}(\vartheta)\|_1 &= \left\|\mathbf{f}^{-\frac{1}{2}}(\vartheta)\mathbf{p}(\vartheta+\pi)^H\mathbf{q}(\vartheta)\mathbf{p}(\vartheta+\pi)\mathbf{f}^{-\frac{1}{2}}(\vartheta)\right\|_1 \\ &\leq \left\|\mathbf{f}^{-\frac{1}{2}}(\vartheta)\mathbf{p}(\vartheta+\pi)^H\right\|_1 \|\mathbf{q}(\vartheta)\|_1 \left\|\mathbf{p}(\vartheta+\pi)\mathbf{f}^{-\frac{1}{2}}(\vartheta)\right\|_1. \end{aligned}$$

Noticing that

$$\left\|\mathbf{p}(\vartheta+\pi)\mathbf{f}^{-\frac{1}{2}}(\vartheta)\right\|_1 = \left\|\mathbf{f}^{-\frac{1}{2}}(\vartheta)\mathbf{p}(\vartheta+\pi)^H\right\|_1$$

and using conditions (IV.5) and (IV.6), we can find δ such that $\|R_{1,1}(\vartheta)\|_1 < \delta$ for all $\vartheta \in [0, 2\pi)\backslash H$.

The uniform boundedness of the other components of $R(\vartheta)$ can be proven in an analogous way, recalling that if ϑ belongs to Θ_0, then $\mathbf{f}$ is non-singular in $\vartheta+\pi$. This implies that the matrix-valued function $R(\vartheta)$ is uniformly bounded in the 1-norm. Since the matrix dimension of $R(\vartheta)$ is fixed for all ϑ and equal to $2s$, the equivalence between the 1-norm and the spectral norm lets us to conclude the proof. □

We highlight that studying the conditions for which $R(\vartheta)$ is bounded can be useful to develop projection strategies for several applications, as we explain in the following chapter. In addition, since (IV.5-IV.7) are sufficient but not necessary conditions, they can be weakened in order to extend the choice of the trigonometric polynomial used to construct the projector.

IV.3.2 TGM Convergence and Optimality: the General Case

In the present subsection we prove the approximation property for a grid transfer operator with a matrix-valued symbol that might be non-diagonalizable. In particular, we focus on the setting of subsection IV.2.2.

Theorem IV.3.5. *Let $\mathcal{C}_n[\mathbf{f}]$, with $\mathbf{f}$ a matrix-valued trigonometric polynomial, $\mathbf{f} \geq 0$ such that condition (IV.8) is satisfied. Let $P^s_{n,m} = \mathcal{C}_n[\mathbf{p}](K_n^{Odd} \otimes I_s)$ be a projecting operator, where $\mathbf{p}$ is a matrix-valued trigonometric polynomial satisfying conditions (i)-(iii) of Section IV.2.2. Then, there exists a positive value γ independent of n such that inequality (b) in Theorem I.9.2 is satisfied.*

Proof. Analogously to Theorem IV.3.4, we note that using the result of Theorem IV.3.3, it is sufficient to prove that there exists a constant $c > 0$ such that for all $\vartheta \in [0, 2\pi)\backslash\Omega(\vartheta_0)$

$$R(\vartheta) \leq cI_{2s}, \tag{IV.21}$$

where

$$\Omega(\vartheta_0) = \{\vartheta_0, (\vartheta_0 + \pi) \mod 2\pi\}.$$

Note that we can write the matrix-valued function $R(\vartheta)$ as

$$R(\vartheta) = \begin{bmatrix} \mathbf{f}(\vartheta) & \\ & \mathbf{f}(\vartheta+\pi) \end{bmatrix}^{-\frac{1}{2}} \begin{bmatrix} I_s - \mathbf{p}(\vartheta)\mathbf{q}(\vartheta)\mathbf{p}(\vartheta)^H & -\mathbf{p}(\vartheta)\mathbf{q}(\vartheta)\mathbf{p}(\vartheta+\pi)^H \\ -\mathbf{p}(\vartheta+\pi)\mathbf{q}(\vartheta)\mathbf{p}(\vartheta)^H & I_s - \mathbf{p}(\vartheta+\pi)\mathbf{q}(\vartheta)\mathbf{p}(\vartheta+\pi)^H \end{bmatrix} \begin{bmatrix} \mathbf{f}(\vartheta) & \\ & \mathbf{f}(\vartheta+\pi) \end{bmatrix}^{-\frac{1}{2}}.$$

Hence, if we prove that for every $\vartheta \in [0, 2\pi)\backslash\Omega(\vartheta_0)$ the matrix $R(\vartheta)$ is uniformly bounded in the spectral norm, then we have that there exists $c > 0$ which bounds the spectral radius of $R(\vartheta)$ and then the latter implies inequality (IV.21). To show that the matrix $R(\vartheta)$ is uniformly bounded in the spectral norm, we can rewrite $R(\vartheta)$ in components as

$$R(\vartheta) = \begin{bmatrix} R_{1,1}(\vartheta) & R_{1,2}(\vartheta) \\ R_{2,1}(\vartheta) & R_{2,2}(\vartheta) \end{bmatrix} = \begin{bmatrix} \mathbf{f}^{-\frac{1}{2}}(\vartheta)(I_s - \mathbf{p}(\vartheta)\mathbf{q}(\vartheta)\mathbf{p}(\vartheta)^H)\mathbf{f}^{-\frac{1}{2}}(\vartheta) & -\mathbf{f}^{-\frac{1}{2}}(\vartheta)\mathbf{p}(\vartheta)\mathbf{q}(\vartheta)\mathbf{p}(\vartheta+\pi)^H\mathbf{f}^{-\frac{1}{2}}(\vartheta+\pi) \\ -\mathbf{f}^{-\frac{1}{2}}(\vartheta+\pi)\mathbf{p}(\vartheta+\pi)\mathbf{q}(\vartheta)\mathbf{p}(\vartheta)^H\mathbf{f}^{-\frac{1}{2}}(\vartheta) & \mathbf{f}^{-\frac{1}{2}}(\vartheta+\pi)(I_s - \mathbf{p}(\vartheta+\pi)\mathbf{q}(\vartheta)\mathbf{p}(\vartheta+\pi)^H)\mathbf{f}^{-\frac{1}{2}}(\vartheta+\pi) \end{bmatrix}.$$

The function $\|R(\vartheta)\|_2 : [0, 2\pi)\backslash\Omega(\vartheta_0) \to \mathbb{R}$ is continuous and, in order to show that $R(\vartheta)$ is uniformly bounded in the spectral norm, Weierstrass Theorem implies that it is sufficient to prove that the following limits exist and are finite:

$$\lim_{\vartheta\to\vartheta_0} \|R(\vartheta)\|_2, \qquad \lim_{\vartheta\to\vartheta_0+\pi} \|R(\vartheta)\|_2.$$

By definition, $R(\vartheta)$ is a Hermitian matrix for $\vartheta \in [0, 2\pi)\backslash\Omega(\vartheta_0)$. Moreover, by direct computation, one can verify that the matrix

$$\begin{bmatrix} I_s - \mathbf{p}(\vartheta)\mathbf{q}(\vartheta)\mathbf{p}(\vartheta)^H & -\mathbf{p}(\vartheta)\mathbf{q}(\vartheta)\mathbf{p}(\vartheta+\pi)^H \\ -\mathbf{p}(\vartheta+\pi)\mathbf{q}(\vartheta)\mathbf{p}(\vartheta)^H & I_s - \mathbf{p}(\vartheta+\pi)\mathbf{q}(\vartheta)\mathbf{p}(\vartheta+\pi)^H \end{bmatrix}$$

is a projector, then it has eigenvalues 0 and 1. Consequently, from the Sylvester inertia law, it follows that $R(\vartheta)$ is a non-negative definite matrix.

We remark that in order to bound the spectral norm of a non-negative definite matrix-valued function, it is sufficient to bound its trace. Hence, we check that the spectral norms of the elements on the block diagonal of $R(\vartheta)$ are bounded. The latter is equivalent to verify that the following limits

$$\lim_{\vartheta\to\vartheta_0} \|R_{1,1}(\vartheta)\|_2, \tag{IV.22}$$

$$\lim_{\vartheta\to\vartheta_0} \|R_{2,2}(\vartheta)\|_2, \tag{IV.23}$$

$$\lim_{\vartheta\to\vartheta_0+\pi} \|R_{1,1}(\vartheta)\|_2, \tag{IV.24}$$

$$\lim_{\vartheta\to\vartheta_0+\pi} \|R_{2,2}(\vartheta)\|_2 \tag{IV.25}$$

exist and they are finite, which in practice requires only the proof of (IV.22). Indeed, the finiteness of (IV.23) and (IV.24) is implied by the hypotheses on $\mathbf{f}$, which is non-singular in $\vartheta_0 + \pi$. The finiteness of (IV.25) can be proven as (IV.22) taking into account that $R(\vartheta)$ is 2π-periodic.

To prove (IV.22) we note that for all $\vartheta \in [0, 2\pi)\backslash\Omega(\vartheta_0)$, we can write

$$\begin{aligned}\|R_{1,1}(\vartheta)\|_2 = &\left\|\mathbf{f}^{-\frac{1}{2}}(\vartheta)(I_s - \mathbf{p}(\vartheta)\mathbf{q}(\vartheta)\mathbf{p}(\vartheta)^H)\mathbf{f}^{-\frac{1}{2}}(\vartheta)\right\|_2 = \\ &\left\|\mathbf{f}^{-1}(\vartheta) - \mathbf{f}^{-\frac{1}{2}}(\vartheta)\mathbf{r}(\vartheta)\mathbf{f}^{-\frac{1}{2}}(\vartheta)\right\|_2,\end{aligned}$$

with $\mathbf{r}(\vartheta)$ defined as in (IV.10).

Without loss of generality, we can assume that $\bar{\jmath} = 1$, that is $q_1(\vartheta_0)$ is the eigenvector of $\mathbf{f}(\vartheta_0)$ associated with the eigenvalue 0. Indeed, if $\bar{\jmath} \neq 1$, it is sufficient to permute rows and columns of $D(\vartheta_0)$ in the factorization in (IV.9) via a permutation matrix Π which brings the diagonalization of $\mathbf{f}(\vartheta_0)$ into the desired form. Moreover, we can assume that $\|q_1(\vartheta_0)\|_2 = 1$.

From condition (i) we have that the matrix-valued function $\mathbf{r}(\vartheta)$ is Hermitian for all $\vartheta \in [0, 2\pi)$. In addition, from condition (ii) and from the latter assumption on $\bar{\jmath}$, the matrix $\mathbf{r}(\vartheta)$ can be decomposed as $\mathbf{r}(\vartheta) = W_{\mathbf{r}}(\vartheta)D_{\mathbf{r}}(\vartheta)W_{\mathbf{r}}^H(\vartheta)$ and

$$\begin{aligned}\mathbf{r}(\vartheta_0) =&W_{\mathbf{r}}(\vartheta_0)D_{\mathbf{r}}(\vartheta_0)W_{\mathbf{r}}^H(\vartheta_0) = \\ &\left[\, q_1(\vartheta_0) \,\middle|\, w_2(\vartheta_0) \,\middle|\, \dots \,\middle|\, w_s(\vartheta_0) \,\right] \begin{bmatrix} 1 & & & \\ & \lambda_2(\mathbf{r}(\vartheta_0)) & & \\ & & \ddots & \\ & & & \lambda_s(\mathbf{r}(\vartheta_0)) \end{bmatrix} \begin{bmatrix} \frac{q_1{}^H(\vartheta_0)}{} \\ \hline \frac{w_2{}^H(\vartheta_0)}{} \\ \hline \vdots \\ \hline \frac{w_s{}^H(\vartheta_0)}{} \end{bmatrix}.\end{aligned}$$

Then, we can rewrite the quantity to bound as follows:

$$\begin{aligned}&\lim_{\vartheta\to\vartheta_0} \|\mathbf{q}(\vartheta)D^{-1}(\vartheta)Q^H(\vartheta) - \mathbf{q}(\vartheta)D^{-\frac{1}{2}}(\vartheta)Q^H(\vartheta)W_{\mathbf{r}}(\vartheta)D_{\mathbf{r}}(\vartheta)W_{\mathbf{r}}^H(\vartheta)Q^H(\vartheta)D^{-\frac{1}{2}}(\vartheta)Q^H(\vartheta)\|_2 = \\ &\lim_{\vartheta\to\vartheta_0} \|D^{-1}(\vartheta) - D^{-\frac{1}{2}}(\vartheta)Q^H(\vartheta)W_{\mathbf{r}}(\vartheta)D_{\mathbf{r}}(\vartheta)W_{\mathbf{r}}^H(\vartheta)Q^H(\vartheta)D^{-\frac{1}{2}}(\vartheta)\|_2.\end{aligned} \tag{IV.26}$$

By definition of $Q(\vartheta_0)$ and $W_{\mathbf{r}}(\vartheta_0)$, the vector $q_0(\vartheta_0)$ is orthogonal with respect to both $q_j(\vartheta_0)$, $w_j(\vartheta_0)$, $j = 2, \dots, s$. Denoting by $\mathbf{o}_{s-1}$ the null column vector of size $s-1$, we have

$$\lim_{\vartheta\to\vartheta_0} Q^H(\vartheta)W_{\mathbf{r}}(\vartheta) = \begin{bmatrix} q_1(\vartheta_0)^H q_1(\vartheta_0) & \mathbf{o}_{s-1}^T \\ \mathbf{o}_{s-1} & M(\vartheta_0) \end{bmatrix}, \tag{IV.27}$$

where $M(\vartheta)$ is a matrix-valued function which is well-defined and continuous on $[0, 2\pi]$. Then, since the eigenvalue functions $\lambda_i(\mathbf{f}(\vartheta))^{-1}$, for $i = 2, \dots, s$, are well-defined and continuous on $[0, 2\pi]$, see Lemma I.3.1, the quantity to bound

$$\left\| \begin{bmatrix} \lim_{\vartheta\to\vartheta_0} \lambda_1(\mathbf{f}(\vartheta)))^{-1}(1 - \lambda_1(\mathbf{r}(\vartheta))) & & \mathbf{o}_{s-1}^T \\ \mathbf{o}_{s-1} & & \begin{bmatrix} \lambda_2(\mathbf{f}(\vartheta_0))^{-1} & & \\ & \ddots & \\ & & \lambda_s(\mathbf{f}(\vartheta_0))^{-1} \end{bmatrix} (I_{s-1} - M(\vartheta_0)M^T(\vartheta_0)) \end{bmatrix} \right\|_2.$$

Consequently, the thesis follows from condition (iii) of Subsection IV.2.2. □

In practical applications choosing a trigonometric polynomial $\mathbf{p}$ such that condition (ii) is verified could not be trivial. Hence, in the following, assuming that $\mathbf{p}$ satisfies condition (i) so that the matrix-valued function $\mathbf{r}$ is well-defined, we provide a useful result that can be applied to construct $\mathbf{p}$ that fulfils condition (ii).

Lemma IV.3.6. *Let* $\mathbf{f}$ *be a matrix-valued trigonometric polynomial,* $\mathbf{f} \geq 0$ *that satisfies condition (IV.8). Assume* $\mathbf{p}$ *is a matrix-valued trigonometric polynomial such that condition* (i) *is fulfilled, so that the matrix-valued function* $\mathbf{r}$ *defined as in* (IV.10) *is well-defined. Assume that the eigenvector* $q_{\bar{\jmath}}(\vartheta_0)$ *associated with the ill-conditioned subspace of* $\mathbf{f}(\vartheta_0)$*, i.e.,* $\mathbf{f}(\vartheta_0)q_{\bar{\jmath}}(\vartheta_0) = 0q_{\bar{\jmath}}(\vartheta_0)$*, is such that:*

1. $q_{\bar{\jmath}}(\vartheta_0)$ *is an eigenvector of* $\mathbf{p}(\vartheta_0)$*, associated to* $\lambda_{\bar{\jmath}}^{(1)} \neq 0$ *that is*

$$\mathbf{p}(\vartheta_0)q_{\bar{\jmath}}(\vartheta_0) = \lambda_{\bar{\jmath}}^{(1)}q_{\bar{\jmath}}(\vartheta_0);$$

2. $q_{\bar{\jmath}}(\vartheta_0)$ *is an eigenvector of* $\mathbf{p}(\vartheta_0 + \pi)$ *associated with the zero eigenvalue, that is*

$$\mathbf{p}(\vartheta_0 + \pi)q_{\bar{\jmath}}(\vartheta_0) = 0q_{\bar{\jmath}}(\vartheta_0);$$

3. $q_{\bar{\jmath}}(\vartheta_0)$ *is an eigenvector of* $\mathbf{p}(\vartheta_0)^H$*, associated to* $\lambda_{\bar{\jmath}}^{(2)} \neq 0$*, that is*

$$\mathbf{p}(\vartheta_0)^H q_{\bar{\jmath}}(\vartheta_0) = \lambda_{\bar{\jmath}}^{(2)}q_{\bar{\jmath}}(\vartheta_0).$$

Then condition (ii) *is satisfied.*

Proof. From all the hypotheses on $q_{\bar{\jmath}}(\vartheta_0)$ and by direct computation, we have

$$\left(\mathbf{p}(\vartheta_0)^H\mathbf{p}(\vartheta_0) + \mathbf{p}(\vartheta_0 + \pi)^H\mathbf{p}(\vartheta_0 + \pi)\right) q_{\bar{\jmath}}(\vartheta_0) = \lambda_{\bar{\jmath}}^{(1)}\lambda_{\bar{\jmath}}^{(2)}q_{\bar{\jmath}}(\vartheta_0).$$

Then, by definition of $\mathbf{r}(\vartheta)$ in (IV.10), it holds that

$$\begin{aligned}
\mathbf{r}(\vartheta_0)q_{\bar{\jmath}}(\vartheta_0) &= \mathbf{p}(\vartheta_0)\left(\mathbf{p}(\vartheta_0)^H\mathbf{p}(\vartheta_0) + \mathbf{p}(\vartheta_0 + \pi)^H\mathbf{p}(\vartheta_0 + \pi)\right)^{-1}\mathbf{p}(\vartheta_0)^H q_{\bar{\jmath}}(\vartheta_0) \\
&= \lambda_{\bar{\jmath}}^{(2)}\mathbf{p}(\vartheta_0)\left(\mathbf{p}(\vartheta_0)^H\mathbf{p}(\vartheta_0) + \mathbf{p}(\vartheta_0 + \pi)^H\mathbf{p}(\vartheta_0 + \pi)\right)^{-1} q_{\bar{\jmath}}(\vartheta_0) \\
&= \lambda_{\bar{\jmath}}^{(2)}\frac{1}{\lambda_{\bar{\jmath}}^{(1)}\lambda_{\bar{\jmath}}^{(2)}}\mathbf{p}(\vartheta_0)q_{\bar{\jmath}}(\vartheta_0) = q_{\bar{\jmath}}(\vartheta_0).
\end{aligned}$$

□

IV.4 Extension to the Multidimensional Case

In the present section we give a possible extension of the convergence results in the multidimensional setting.

First, we define the objects of our analysis in more dimensions. Let $\mathbf{n} := (n_1, \ldots, n_k)$ be a multi-index in $\mathbb{N}^k$. We need to provide a generalized definition of the projector $P_{n,m}^s$ for the $k-$level block-circulant matrix $\mathcal{C}_{\mathbf{n}}[\mathbf{f}]$ of dimension $s\mathcal{N}(\mathbf{n})$ generated by a multivariate matrix-valued trigonometric polynomial $\mathbf{f}$.

Analogously to the scalar case, we want to construct the projectors from an arbitrary multi-level block-circulant matrix $\mathcal{C}_{\mathbf{n}}[\mathbf{p}]$, with $\mathbf{p}$ multivariate matrix-valued trigonometric polynomial. For the construction of the projector we can use a tensor product approach:

$$P_{\mathbf{n},\mathbf{m}} = \mathcal{C}_{\mathbf{n}}[\mathbf{p}]\left(K_{\mathbf{n},\mathbf{m}}^{Odd} \otimes I_s\right), \tag{IV.28}$$

where $K^{Odd}_{\mathbf{n},\mathbf{m}}$ is the $\mathcal{N}(\mathbf{n}) \times \frac{\mathcal{N}(\mathbf{n})}{2^k}$ matrix defined by $K^{Odd}_{\mathbf{n},\mathbf{m}} = K^{Odd}_{n_1,m_1} \otimes K^{Odd}_{n_2,m_2} \otimes \cdots \otimes K^{Odd}_{n_k,m_k}$ and $\mathcal{C}_{\mathbf{n}}[\mathbf{p}]$ is a multilevel block-circulant matrix generated by $\mathbf{p}$. The main goal is to combine the proof of Theorem IV.3.5 with the multilevel techniques in [119], in order to generalize conditions *(i)-(iii)* to the multilevel case.

In the $k-$level setting, we are assuming that $\boldsymbol{\vartheta}_0 \in [0, 2\pi)^k$ and $\bar{\jmath} \in \{1, \ldots, s\}$ such that

$$\begin{cases} \lambda_j(\mathbf{f}(\boldsymbol{\vartheta})) = 0 & \text{for } \boldsymbol{\vartheta} = \boldsymbol{\vartheta}_0 \text{ and } j = \bar{\jmath}, \\ \lambda_j(\mathbf{f}(\boldsymbol{\vartheta})) > 0 & \text{otherwise.} \end{cases} \tag{IV.29}$$

The latter assumption means that the matrix $\mathbf{f}(\boldsymbol{\vartheta})$ has exactly one zero eigenvalue in $\boldsymbol{\vartheta}_0$ and it is positive definite in $[0, 2\pi)^k \backslash \{\boldsymbol{\vartheta}_0\}$. Let us assume that, $q_{\bar{\jmath}}(\boldsymbol{\vartheta}_0)$ is the eigenvector of $\mathbf{f}(\boldsymbol{\vartheta}_0)$ associated with $\lambda_{\bar{\jmath}}(\mathbf{f}(\boldsymbol{\vartheta}_0)) = 0$. Moreover, define $\Omega(\boldsymbol{\vartheta}) = \left\{\boldsymbol{\vartheta} + \pi\boldsymbol{\eta},\, \boldsymbol{\eta} \in \{0, 1\}^k\right\}$. Under these hypotheses, the multilevel extension of conditions *(i)-(iii)* of Section IV.2.2, which are sufficient to ensure the optimal convergence of the TGM in the multilevel case, is the following. Choose $\mathbf{p}(\cdot)$ such that

- $$\sum_{\boldsymbol{\xi} \in \Omega(\boldsymbol{\vartheta})} \mathbf{p}(\boldsymbol{\xi})^H \mathbf{p}(\boldsymbol{\xi}) > 0, \quad \forall\, \boldsymbol{\vartheta} \in [0, 2\pi)^k, \tag{IV.30}$$

  which implies that the trigonometric function

  $$\mathbf{r}(\boldsymbol{\vartheta}) = \mathbf{p}(\boldsymbol{\vartheta}) \left(\sum_{\boldsymbol{\xi} \in \Omega(\boldsymbol{\vartheta})} \mathbf{p}(\boldsymbol{\xi})^H \mathbf{p}(\boldsymbol{\xi}) \right)^{-1} \mathbf{p}(\boldsymbol{\vartheta})^H$$

  is well-defined for all $\boldsymbol{\vartheta} \in [0, 2\pi)^k$.

- $$\mathbf{r}(\boldsymbol{\vartheta}_0) q_{\bar{\jmath}}(\boldsymbol{\vartheta}_0) = q_{\bar{\jmath}}(\boldsymbol{\vartheta}_0). \tag{IV.31}$$

- $$\lim_{\boldsymbol{\vartheta} \to \boldsymbol{\vartheta}_0} \lambda_{\bar{\jmath}}(\mathbf{f}(\boldsymbol{\vartheta}))^{-1} (1 - \lambda_{\bar{\jmath}}(\mathbf{r}(\boldsymbol{\vartheta}))) = c, \tag{IV.32}$$

  where $c \in \mathbb{R}$ is a constant.

In the following we want to construct a multilevel projector $P_{\mathbf{n},\mathbf{m}}$ such that the conditions (IV.30)-(IV.32) are satisfied and, then, the optimal convergence of the TGM is ensured in our multidimensional setting.

In particular, starting from $s_\ell \times s_\ell$ matrix-valued trigonometric polynomials $\mathbf{p}_\ell$, $\ell = 1, \ldots, k$, we aim at defining a multivariate polynomial $\mathbf{p}^{(k)}$ associated to the multilevel projector $P_{\mathbf{n},\mathbf{m}}$ such that the conditions (IV.30)-(IV.32) are satisfied.

In the following lemmas, we show that the aforementioned goal is achieved if we choose the multivariate matrix-valued trigonometric polynomial

$$\mathbf{p}^{(k)}(\vartheta_1, \vartheta_2, \ldots, \vartheta_k) = \bigotimes_{\ell=1}^{k} \mathbf{p}_\ell(\vartheta_\ell), \tag{IV.33}$$

where $\mathbf{p}_\ell(\vartheta_\ell) \in \mathbb{C}^{s_\ell \times s_\ell}$ are polynomials that satisfy conditions *(i)-(iii)* of Subsection IV.2.2.

Lemma IV.4.1. *Let* $\mathbf{p}^{(k)}(\vartheta_1, \vartheta_2, \ldots, \vartheta_k)$ *be defined as in (IV.33). Then,*

$$\sum_{\boldsymbol{\xi}\in\Omega(\boldsymbol{\vartheta})} \mathbf{p}^{(k)}(\boldsymbol{\xi})^H \mathbf{p}^{(k)}(\boldsymbol{\xi}) = \bigotimes_{\ell=1}^{k} \left(\mathbf{p}_\ell(\vartheta_\ell)^H \mathbf{p}_\ell(\vartheta_\ell) + \mathbf{p}_\ell(\vartheta_\ell+\pi)^H \mathbf{p}_\ell(\vartheta_\ell+\pi)\right).$$

Proof. By definition, $\mathbf{p}^{(k)}(\boldsymbol{\vartheta}) = \bigotimes_{\ell=1}^{k} \mathbf{p}_\ell(\vartheta_\ell)$, then

$$\begin{aligned}\sum_{\boldsymbol{\xi}\in\Omega(\boldsymbol{\vartheta})} \mathbf{p}^{(k)}(\boldsymbol{\xi})^H \mathbf{p}^{(k)}(\boldsymbol{\xi}) &= \sum_{\boldsymbol{\xi}\in\Omega(\boldsymbol{\vartheta})} \left(\bigotimes_{\ell=1}^{k} \mathbf{p}_\ell(\xi_\ell)^H\right)\left(\bigotimes_{\ell=1}^{k} \mathbf{p}_\ell(\xi_\ell)\right) \\ &= \sum_{\boldsymbol{\xi}\in\Omega(\boldsymbol{\vartheta})} \left(\bigotimes_{\ell=1}^{k} \left(\mathbf{p}_\ell(\xi_\ell)^H \mathbf{p}_{r_\ell}(\xi_\ell)\right)\right).\end{aligned}$$

The proof is then concluded once we prove by induction on k the following equality

$$\sum_{\boldsymbol{\xi}\in\Omega(\boldsymbol{\vartheta})} \left(\bigotimes_{\ell=1}^{k} \left(\mathbf{p}_\ell(\xi_\ell)^H \mathbf{p}_{r_\ell}(\xi_\ell)\right)\right) = \bigotimes_{\ell=1}^{k} \left(\mathbf{p}_\ell(\vartheta_\ell)^H \mathbf{p}_\ell(\vartheta_\ell) + \mathbf{p}_\ell(\vartheta_\ell+\pi)^H \mathbf{p}_\ell(\vartheta_\ell+\pi)\right). \tag{IV.34}$$

The equation above is clearly verified for $k = 1$, indeed, by definition

$$\begin{aligned}\sum_{\xi\in\Omega(\vartheta)} \left(\bigotimes_{\ell=1}^{1} \left(\mathbf{p}_\ell(\xi_\ell)^H \mathbf{p}_{r_\ell}(\xi_\ell)\right)\right) &= \sum_{\xi\in\{\vartheta_1,\vartheta_1+\pi\}} \left(\mathbf{p}_1(\xi_1)^H \mathbf{p}_1(\xi_1)\right) = \\ &\mathbf{p}_{s_1}(\vartheta_1)^H \mathbf{p}_1(\vartheta_1) + \mathbf{p}_1(\vartheta_1+\pi)^H \mathbf{p}_1(\vartheta_1+\pi) = \\ &\bigotimes_{\ell=1}^{1} \left(\mathbf{p}_\ell(\vartheta_\ell)^H \mathbf{p}_\ell(\vartheta_\ell) + \mathbf{p}_\ell(\vartheta_\ell+\pi)^H \mathbf{p}_\ell(\vartheta_\ell+\pi)\right).\end{aligned}$$

Let us assume that equality (IV.34) is true for $k-1$. We have that

$$\begin{aligned}&\bigotimes_{\ell=1}^{k} \left(\mathbf{p}_\ell(\vartheta_\ell)^H \mathbf{p}_\ell(\vartheta_\ell) + \mathbf{p}_\ell(\vartheta_\ell+\pi)^H \mathbf{p}_\ell(\vartheta_\ell+\pi)\right) = \\ &\left[\bigotimes_{\ell=1}^{k-1} \left(\mathbf{p}_\ell(\vartheta_\ell)^H \mathbf{p}_\ell(\vartheta_\ell) + \mathbf{p}_\ell(\vartheta_\ell+\pi)^H \mathbf{p}_\ell(\vartheta_\ell+\pi)\right)\right] \otimes \\ &\left(\mathbf{p}_k(\vartheta_k)^H \mathbf{p}_k(\vartheta_k) + \mathbf{p}_k(\vartheta_k+\pi)^H \mathbf{p}_k(\vartheta_k+\pi)\right)\end{aligned}$$

The left-hand side of the latter term is a function of $k-1$ variables $(\vartheta_1, \vartheta_2, \ldots, \vartheta_{k-1})$. Then,

by the inductive hypothesis and from the properties of the tensor product we have

$$\left[\bigotimes_{\ell=1}^{k-1}\left(\mathbf{p}_\ell(\vartheta_\ell)^H\mathbf{p}_\ell(\vartheta_\ell)+\mathbf{p}_\ell(\vartheta_\ell+\pi)^H\mathbf{p}_\ell(\vartheta_\ell+\pi)\right)\right]\otimes$$
$$\left(\mathbf{p}_k(\vartheta_k)^H\mathbf{p}_k(\vartheta_k)+\mathbf{p}_k(\vartheta_k+\pi)^H\mathbf{p}_k(\vartheta_k+\pi)\right)=$$
$$\left(\sum_{\substack{(\xi_1,\xi_2,\ldots,\xi_{k-1})\\ \in \\ \Omega(\vartheta_1,\vartheta_2,\ldots,\vartheta_{k-1})}}\bigotimes_{\ell=1}^{k-1}\mathbf{p}_\ell(\xi_\ell)^H\mathbf{p}_\ell(\xi_\ell)\right)\otimes$$
$$\left(\mathbf{p}_k(\vartheta_k)^H\mathbf{p}_k(\vartheta_k)+\mathbf{p}_k(\vartheta_k+\pi)^H\mathbf{p}_k(\vartheta_k+\pi)\right)=$$
$$\sum_{\substack{(\xi_1,\xi_2,\ldots,\xi_{k-1})\\ \in \\ \Omega(\vartheta_1,\vartheta_2,\ldots,\vartheta_{k-1})}}\left[\left(\bigotimes_{\ell=1}^{k-1}\mathbf{p}_\ell(\xi_\ell)^H\mathbf{p}_\ell(\xi_\ell)\right)\otimes\left(\mathbf{p}_k(\vartheta_k)^H\mathbf{p}_k(\vartheta_k)+\mathbf{p}_k(\vartheta_k+\pi)^H\mathbf{p}_k(\vartheta_k+\pi)\right)\right]=$$
$$\sum_{\substack{\boldsymbol{\xi}\in\{(\vartheta_1+l_1\pi,\ldots,\vartheta_{k-1}+l_{k-1}\pi\},\\ \mathbf{l}\in\{0,1\}^{k-1}}}\left[\left(\bigotimes_{\ell=1}^{k-1}\mathbf{p}_\ell(\xi_\ell)^H\mathbf{p}_\ell(\xi_\ell)\right)\otimes\mathbf{p}_k(\vartheta_k)^H\mathbf{p}_k(\vartheta_k)+\right.$$
$$\left.+\left(\bigotimes_{\ell=1}^{k-1}\mathbf{p}_\ell(\xi_\ell)^H\mathbf{p}_\ell(\xi_\ell)\right)\otimes\mathbf{p}_k(\vartheta_k+\pi)^H\mathbf{p}_k(\vartheta_k+\pi)\right]=$$
$$\sum_{\substack{\boldsymbol{\xi}\in\{(\vartheta_1+l_1\pi,\ldots,\vartheta_{k-1}+l_{k-1}\pi,\vartheta_k)\},\\ \mathbf{l}\in\{0,1\}^{k-1}}}\bigotimes_{\ell=1}^{k}\mathbf{p}_\ell(\xi_\ell)^H\mathbf{p}_\ell(\xi_\ell)+$$
$$\sum_{\substack{\boldsymbol{\xi}\in\{(\vartheta_1+l_1\pi,\ldots,\vartheta_{k-1}+l_{k-1}\pi,\vartheta_k+\pi)\},\\ \mathbf{l}\in\{0,1\}^{k-1}}}\bigotimes_{\ell=1}^{k}\mathbf{p}_\ell(\xi_\ell)^H\mathbf{p}_\ell(\xi_\ell)=$$
$$\sum_{\boldsymbol{\xi}\in\Omega(\boldsymbol{\vartheta})}\bigotimes_{\ell=1}^{k}\mathbf{p}_\ell(\xi_\ell)^H\mathbf{p}_\ell(\xi_\ell).$$

Then, relation (IV.34) is verified for k, and this concludes the proof. □

Lemma IV.4.2. *Let $\mathbf{p}^{(k)}(\vartheta_1,\vartheta_2,\ldots,\vartheta_k)$ defined as in (IV.33) where $\mathbf{p}_\ell$, for every $\ell=1,\ldots,k$, is a polynomial which verifies the positivity condition (i). Then, $\mathbf{p}^{(k)}$ is such that the positivity condition in the multilevel setting (IV.30) is satisfied.*

Proof. The thesis is a consequence of Lemma IV.4.1 and the matrix tensor product properties. Indeed, the eigenvalues of a tensor product of matrices are the product of the eigenvalues of the matrices. Then, condition (IV.30) is trivially implied from the fact that

$$\sum_{\boldsymbol{\xi}\in\Omega(\boldsymbol{\vartheta})}\mathbf{p}^{(k)}(\boldsymbol{\xi})^H\mathbf{p}^{(k)}(\boldsymbol{\xi})=\bigotimes_{\ell=1}^{k}\left(\mathbf{p}_\ell(\vartheta_\ell)^H\mathbf{p}_\ell(\vartheta_\ell)+\mathbf{p}_\ell(\vartheta_\ell+\pi)^H\mathbf{p}_\ell(\vartheta_\ell+\pi)\right),$$

and from the positivity condition in the unilevel case.

□

Lemma IV.4.3. *Let $\mathbf{p}^{(k)}(\vartheta_1, \vartheta_2, \ldots, \vartheta_k)$ be defined as in (IV.33) and it verifies (IV.30). Then, the trigonometric function*

$$\mathbf{r}(\boldsymbol{\vartheta}) = \mathbf{p}^{(k)}(\boldsymbol{\vartheta}) \left(\sum_{\boldsymbol{\xi} \in \Omega(\boldsymbol{\vartheta})} \mathbf{p}^{(k)}(\boldsymbol{\xi})^H \mathbf{p}^{(k)}(\boldsymbol{\xi}) \right)^{-1} \mathbf{p}^{(k)}(\boldsymbol{\vartheta})^H$$

is well-defined for all $\boldsymbol{\vartheta} \in [0, 2\pi)^k$. Moreover, it holds that

$$\mathbf{r}(\boldsymbol{\vartheta}) = \bigotimes_{\ell=1}^{k} \mathbf{r}_\ell(\vartheta_\ell), \tag{IV.35}$$

where, for $\ell = 1, \ldots, k$, $\mathbf{r}_\ell(\vartheta_\ell) = \mathbf{p}_\ell(\vartheta_\ell) \left(\mathbf{p}_\ell(\vartheta_\ell)^H \mathbf{p}_\ell(\vartheta_\ell) + \mathbf{p}_\ell(\vartheta_\ell + \pi)^H \mathbf{p}_\ell(\vartheta_\ell + \pi) \right)^{-1} \mathbf{p}_\ell(\vartheta_\ell)^H$.

Proof. From Lemma IV.4.2, we have that $\mathbf{r}(\boldsymbol{\vartheta})$ is well-defined for all $\boldsymbol{\vartheta} \in [0, 2\pi)^k$. From Lemma IV.4.1 and the properties of the tensor product, we have

$$\begin{aligned}
&\mathbf{r}(\boldsymbol{\vartheta}) = \mathbf{p}^{(k)}(\boldsymbol{\vartheta}) \left(\sum_{\boldsymbol{\xi} \in \Omega(\boldsymbol{\vartheta})} \mathbf{p}^{(k)}(\boldsymbol{\xi})^H \mathbf{p}^{(k)}(\boldsymbol{\xi}) \right)^{-1} \mathbf{p}^{(k)}(\boldsymbol{\vartheta})^H = \\
&\bigotimes_{\ell=1}^{k} \mathbf{p}_\ell(\vartheta_\ell) \left(\bigotimes_{\ell=1}^{k} \left[\mathbf{p}_\ell(\vartheta_\ell)^H \mathbf{p}_\ell(\vartheta_\ell) + \mathbf{p}_\ell(\vartheta_\ell + \pi)^H \mathbf{p}_\ell(\vartheta_\ell + \pi) \right]^{-1} \right) \bigotimes_{\ell=1}^{k} \mathbf{p}_\ell(\vartheta_\ell)^H = \\
&\bigotimes_{\ell=1}^{k} \left(\mathbf{p}_\ell(\vartheta_\ell) \left[\mathbf{p}_\ell(\vartheta_\ell)^H \mathbf{p}_\ell(\vartheta_\ell) + \mathbf{p}_\ell(\vartheta_\ell + \pi)^H \mathbf{p}_\ell(\vartheta_\ell + \pi) \right]^{-1} \mathbf{p}_\ell(\vartheta_\ell)^H \right) = \bigotimes_{\ell=1}^{k} \mathbf{r}_\ell(\vartheta_\ell).
\end{aligned} \tag{IV.36}$$

□

Lemma IV.4.4. *Let $\mathbf{p}^{(k)}(\vartheta_1, \vartheta_2, \ldots, \vartheta_k)$ be defined as in (IV.33), such that, for all $\ell = 1, \ldots, k$, $\mathbf{p}_\ell(\vartheta_\ell) \in \mathbb{C}^{s_\ell \times s_\ell}$ is a polynomial that satisfies conditions (i)-(iii) of Subsection IV.2.2. Let $q^{(k)} = \bigotimes_{\ell=1,\ldots,k} \mathrm{q}_\ell$, where q_ℓ is the column vector of length s_ℓ such that $\mathbf{r}_\ell(\vartheta_0^{(\ell)})\mathrm{q}_\ell = \mathrm{q}_\ell$, $\ell = 1, \ldots, k$. Then,*

$$\mathbf{r}(\boldsymbol{\vartheta_0}) q^{(k)} = q^{(k)},$$

where $\boldsymbol{\vartheta}_0 = \left(\vartheta_0^{(1)}, \ldots, \vartheta_0^{(k)} \right)$.

Proof. From Lemma IV.4.3, we have that $\mathbf{r}(\boldsymbol{\vartheta_0}) = \bigotimes_{\ell=1}^{k} \mathbf{r}_\ell \left(\vartheta_0^{(\ell)} \right)$, then, by definition and from the properties of the tensor product, it holds

$$\mathbf{r}(\boldsymbol{\vartheta_0}) q^{(k)} = \left(\bigotimes_{\ell=1}^{k} \mathbf{r}_\ell \left(\vartheta_0^{(\ell)} \right) \right) \left(\bigotimes_{\ell=1}^{k} \mathrm{q}_\ell \right) = \bigotimes_{\ell=1}^{k} \left(\mathbf{r}_\ell \left(\vartheta_0^{(\ell)} \right) \mathrm{q}_\ell \right) = \bigotimes_{\ell=1}^{k} \mathrm{q}_\ell = q^{(k)}. \tag{IV.37}$$

□

Lemma IV.4.5. *Let $\mathbf{p}^{(k)}(\vartheta_1, \vartheta_2, \ldots, \vartheta_k)$ be defined as in (IV.33) such that verifies (IV.30). Consider*

$$\mathbf{r}(\boldsymbol{\vartheta}) = \bigotimes_{\ell=1}^{k} \mathbf{r}_\ell(\vartheta_\ell),$$

where, for $\ell = 1, \ldots, k$, $\mathbf{r}_\ell(\vartheta_\ell) = \mathbf{p}_\ell(\vartheta_\ell) \left(\mathbf{p}_\ell(\vartheta_\ell)^H \mathbf{p}_\ell(\vartheta_\ell) + \mathbf{p}_\ell(\vartheta_\ell + \pi)^H \mathbf{p}_\ell(\vartheta_\ell + \pi) \right)^{-1} \mathbf{p}_\ell(\vartheta_\ell)^H$ and they verify condition (iii) of Subsection IV.2.2. Then, $\mathbf{r}(\boldsymbol{\vartheta})$ satisfies condition (IV.32).

Proof. Without loss of generality, suppose that the order of the zero of $\lambda_{\bar{\jmath}}(\mathbf{f}(\vartheta_\ell))$ in $\vartheta_0^{(\ell)}$ is $\varsigma \geq 2$ for $\ell = 1, \ldots, k$, then the functions $1 - \lambda_{\bar{\jmath}}(\mathbf{r}_\ell(\vartheta_\ell))$ have a zero in $\vartheta_0^{(\ell)}$ of order at least $\varsigma \in \mathbb{N}$ for all $\ell = 1, \ldots, k$ by condition *(iii)*. Hence, the $(\varsigma - 1)$-th derivative of $1 - \lambda_{\bar{\jmath}}(\mathbf{r}_\ell(\vartheta_\ell))$ in $\vartheta_0^{(\ell)}$ is equal to zero. Then we have, for $\ell = 1, \ldots, k$,

$$\lambda_{\bar{\jmath}}(\mathbf{r}_\ell(\vartheta_\ell))^{(\varsigma-1)}\Big|_{\vartheta_0^{(\ell)}} = 0.$$

The thesis follows by direct computation of the partial derivatives of $1-\lambda_{\bar{\jmath}}(\mathbf{r}(\boldsymbol{\vartheta}))$ in $\boldsymbol{\vartheta}_0$, exploiting the fact that

$$\mathbf{r}(\boldsymbol{\vartheta}) = \bigotimes_{\ell=1}^{k} \mathbf{r}_\ell(\vartheta_\ell),$$

and

$$\lambda_{\bar{\jmath}}(\mathbf{r}(\boldsymbol{\vartheta})) = \prod_{\ell=1}^{k} \lambda_{\bar{\jmath}}(\mathbf{r}_\ell(\vartheta_\ell)).$$

□

Chapter V

Multigrid for $\mathbb{Q}_s$ Finite Element Matrices Using Block-Toeplitz Symbol Approaches

In the current chapter we consider multigrid methods for the solution of large linear systems whose coefficient matrices arise from the $\mathbb{Q}_s$ approximation of the elliptic problem

$$\begin{cases} \operatorname{div}(-a(\mathbf{x})\nabla u(\mathbf{x})) = \psi(\mathbf{x}), & \mathbf{x} \in \Omega, \\ u(\mathbf{x}) = 0, & \mathbf{x} \in \partial\Omega \end{cases}, \tag{V.1}$$

with Ω being a bounded subset of $\mathbb{R}^k$ having smooth boundaries and with a being continuous and positive on $\overline{\Omega}$.

The multigrid techniques that we present are based both on the theoretical results of **Chapter IV** and on the spectral analysis of the involved matrix-sequences by means of the study of the associated spectral symbol provided in [64].

Indeed, in the systematic work in [64], tensor rectangular Finite Element approximations $\mathbb{Q}_s$ of any degree s and of any dimensionality k are considered and the spectral analysis of the stiffness matrix-sequences $\{A_n\}_n$ is provided in the sense of asymptotic distributions, spectral clustering, spectral localization, extremal eigenvalues, and conditioning.

We observe that the information obtained in [64] is strongly based on the notion of spectral symbol and it is studied from the perspective of multilevel block–Toeplitz operators and GLT sequences, which are all concepts that we introduced in **Chapter I**.

The first procedure that we propose is a classical multigrid strategy that follows a functional approach, that is, we define the prolongation operator as the inclusion operator between the coarser and finer involved functional spaces. Our aim is to analyse the prolongation matrix as a cut block-Toeplitz matrix so that the grid transfer operator fits in the setting of Section IV.2. Indeed, we provide a two-grid convergence and optimality analysis exploiting the results in Section IV.3.

We perform an analogous analysis also to a second multigrid strategy, where we choose the standard bisection prolongation operator. In this case, we employ the results in Subsection IV.3.2 to prove that the chosen grid transfer operator fulfils the approximation property for the linear systems associated to the $\mathbb{Q}_s$ discretization of the model problem (V.1).

Finally, we define a new class of grid transfer operators that satisfy the theoretical conditions of **Chapter IV**. In particular, we explain how to choose the trigonometric polynomial that generates the block-Toeplitz matrix used in the construction of the grid transfer operator focusing only on algebraic considerations on the symbol of the linear system matrix-sequence. We highlight that the presented procedure has a wide interest, since it might be applied to every matrix-sequence that falls into the theoretical setting.

The contents of this chapter are partly published in [55] and partly in the process of being published in [20, 39]. The chapter is outlined as follows. In Section V.1, we present the problem, the specific $\mathbb{Q}_s$ approximation, and the analysis of the structure and of the spectral features of the related matrices. Section V.2 is devoted to the multigrid strategy definition and analysis for the geometric projection operators. Moreover, we confirm the derived optimality results through numerical tests for different values of the function a, both in one dimension and in higher dimension. In Section V.3, we provide the analysis of the approximation property for the standard bisection grid transfer operator. Finally, Section V.4 is dedicated to the development of a class of grid transfer operators that are suitable for both the two-grid and the V-cycle convergence. A selection of numerical experiments confirms the effectiveness of the presented projection strategy for the two-grid method and indicates a heuristic technique for V-cycle optimality.

V.1 $\mathbb{Q}_s$ Lagrangian FEM Stiffness Matrices

In what follows, we present the details of the $\mathbb{Q}_s$ approximation of a simplified version of the problem in (V.1) as follows. We set the dimensionality k equal to 1, the function $a(x)$ identically equal to 1, and $\Omega = (0,1)$. In this context, the problem becomes:

find u such that

$$\begin{cases} -u''(x) = \psi(x) \quad \text{on } (0,1) \\ u(0) = u(1) = 0, \end{cases} \tag{V.2}$$

where $\psi(x) \in L^2((0,1))$.

We write the weak formulation of the problem as follows:

find $u \in H_0^1(0,1)$ such that

$$\alpha(u,v) =< \psi, v > \quad \forall v \in H_0^1(0,1), \tag{V.3}$$

where $\alpha(u,v) := \int_{(0,1)} u'(x)v'(x)\,dx$ and $< \psi, v >:= \int_{(0,1)} \psi(x)v(x)\,dx$. For $s, n \geq 1$, we define the space

$$\mathcal{V}_n^{(s)} := \left\{ \sigma \in C([0,1]) : \sigma_{|_{\left[\frac{i}{n}, \frac{i+1}{n}\right]}} \in \mathbb{P}_s,\ \forall i = 0, \ldots, n-1 \right\}, \tag{V.4}$$

where we denote by $\mathbb{P}_s$ the space of polynomials of degree less than or equal to s. Then, the space $\mathcal{V}_n^{(s)}$ represents the space of continuous piecewise polynomial functions. Starting from $\mathcal{V}_n^{(s)}$, we consider its subspace of functions that vanish on the boundary, defined by

$$\mathcal{W}_n^{(s)} := \left\{ \sigma \in \mathcal{V}_n^{(s)} : \sigma(0) = \sigma(1) = 0 \right\}. \tag{V.5}$$

Note that $\mathcal{W}_n^{(s)}$ is a finite $ns-1$ dimensional subspace of $H_0^1(0,1)$ and, following a Galerkin approach, we approximate the solution u of the variational problem by solving the problem:

find $u_{s,n} \in \mathcal{W}_n^{(s)}$ such that

$$\alpha(u_{s,n}, v) =< \psi, v > \quad \forall v \in \mathcal{W}_n^{(s)}. \tag{V.6}$$

We define the uniform knot sequence

$$\xi_i = \frac{i}{ns},\ i = 0, \dots, ns, \tag{V.7}$$

and the Lagrangian basis functions by

$$\varphi_j^{n,s}(\xi_i) = \delta_{i,j},\ i, j = 0, \dots, ns. \tag{V.8}$$

with $\delta_{i,j}$ being the Kronecker delta. It can be shown that the latter definition is well-posed and that $\{\varphi_1^{n,s}, \dots, \varphi_{ns-1}^{n,s}\}$ is a basis for $\mathcal{W}_n^{(s)}$. Then $u_{s,n}$ can be written as linear combination as

$$u_{s,n} = \sum_{j=1}^{ns-1} u_j \varphi_j^{n,s},$$

and solving the problem (V.6) reduces to the solution of the linear system

$$A_n^{(s)} \mathbf{u} = \mathbf{b},$$

with

$$A_n^{(s)} = [\alpha(\varphi_j^{n,s}, \varphi_i^{n,s})]_{i,j=1}^{ns-1}, \quad \mathbf{b} = [< \psi, \varphi_i^{n,s} >]_{i=1}^{ns-1}, \quad \mathbf{u} = [u_i]_{i=1}^{ns-1}.$$

The spectral properties of the Stiffness matrix-sequence $\left\{A_n^{(s)}\right\}_n$ were studied in [64] and, in the following, we report the essential features. Let us consider the Lagrange polynomials $L_0, \dots, L_s$ associated with the reference knots $t_j = j/s,\ j = 0, \dots, s$:

$$\begin{aligned} L_i(t) &= \prod_{\substack{j=0 \\ j \neq i}}^{s} \frac{t - t_j}{t_i - t_j} = \prod_{\substack{j=0 \\ j \neq i}}^{s} \frac{st - j}{i - j}, \quad i = 0, \dots, s, \\ L_i(t_j) &= \delta_{ij}, \quad i, j = 0, \dots, s. \end{aligned} \tag{V.9}$$

Then, the $\mathbb{Q}_s$ stiffness matrix for equation (V.2) equals the matrix $A_n^{(s)}$ in the next theorem.

Theorem V.1.1. *([64]) Let $s, n \geq 1$. Then,*

$$A_n^{(s)} = \begin{bmatrix} K_0 & K_1^T & & \\ K_1 & \ddots & \ddots & \\ & \ddots & \ddots & K_1^T \\ & & K_1 & K_0 \end{bmatrix}_-, \tag{V.10}$$

where the subscripts "$-$" mean that the last row and column of the of the whole matrices in square brackets are deleted, while K_0, K_1 are $s \times s$ blocks given by

$$K_0 = \left[\begin{array}{ccc|c} \langle L_1', L_1'\rangle & \cdots & \langle L_{s-1}', L_1'\rangle & \langle L_s', L_1'\rangle \\ \vdots & & \vdots & \vdots \\ \langle L_1', L_{s-1}'\rangle & \cdots & \langle L_{s-1}', L_{s-1}'\rangle & \langle L_s', L_{s-1}'\rangle \\ \hline \langle L_1', L_s'\rangle & \cdots & \langle L_{s-1}', L_s'\rangle & \langle L_s', L_s'\rangle + \langle L_0', L_0'\rangle \end{array}\right],$$
$$K_1 = \left[\begin{array}{ccc|c} 0 & \cdots & 0 & \langle L_0', L_1'\rangle \\ 0 & \cdots & 0 & \langle L_0', L_2'\rangle \\ \vdots & & \vdots & \vdots \\ 0 & \cdots & 0 & \langle L_0', L_s'\rangle \end{array}\right], \tag{V.11}$$

with $L_0, \ldots, L_s$ being the Lagrange polynomials in (V.9)*. In particular, $A_n^{(s)}$ is the $(ns-1)\times(ns-1)$ leading principal submatrix of the block-Toeplitz matrices $T_n[\mathbf{f}_{\mathbb{Q}_s}]$ and $\mathbf{f}_{\mathbb{Q}_s} : [-\pi, \pi] \to \mathbb{C}^{s\times s}$ is the Hermitian matrix-valued trigonometric polynomial given by*

$$\mathbf{f}_{\mathbb{Q}_s}(\vartheta) := K_0 + K_1 \mathrm{e}^{\hat{\imath}\vartheta} + K_1^T \mathrm{e}^{-\hat{\imath}\vartheta}. \tag{V.12}$$

An interesting property of the Hermitian matrix-valued functions $\mathbf{f}_{\mathbb{Q}_s}(\vartheta)$ defined in equation (V.12) is reported in the theorem below: in fact, from the point of view of asymptotic spectral distributions, the message is that, independently of the parameter s, the spectral symbol possesses the same character as $2-2\cos(\vartheta)$, which is the symbol of the basic linear Finite Elements and the most standard Finite Differences.

Theorem V.1.2. *([64]) Let $s \geq 1$. Then,*

$$\det(\mathbf{f}_{\mathbb{Q}_s}(\vartheta)) = d_s(2 - 2\cos(\vartheta)), \tag{V.13}$$

where $d_s = \det([\langle L_j', L_i'\rangle]_{i,j=1}^{s}) = \det([\langle L_j', L_i'\rangle]_{i,j=1}^{s-1}) > 0$ (with $d_1 = 1$, being the determinant of the empty matrix equal to 1 by convention) and $L_0, \ldots, L_s$ are the Lagrange polynomials in Equation (V.9)*.*

Furthermore, a generalization of the previous result in higher dimension is given in [106] and is reported in the subsequent theorem.

Theorem V.1.3. *([106]) Given the symbols $\mathbf{f}_{\mathbb{Q}_s}$ in dimension $k \geq 1$, the following statements hold true:*

1. $\mathbf{f}_{\mathbb{Q}_s}(0)\mathbf{e}_s = 0\mathbf{e}_s$, $s \geq 1$;
2. *there exist constants $C_2 \geq C_1 > 0$ (dependent on $\mathbf{f}_{\mathbb{Q}_s}$) such that*

$$C_1 \sum_{j=1}^{s}(2 - 2\cos(\vartheta_j)) \leq \lambda_1(\mathbf{f}_{\mathbb{Q}_s}(\vartheta)) \leq C_2 \sum_{j=1}^{s}(2 - 2\cos(\vartheta_j)); \tag{V.14}$$

3. *there exist constants $M \geq m > 0$ (dependent on $\mathbf{f}_{\mathbb{Q}_s}$) such that*

$$0 < m \leq \lambda_j(\mathbf{f}_{\mathbb{Q}_s}(\vartheta)) \leq M, \quad j = 2, \ldots, s^k. \tag{V.15}$$

V.2 A Geometric Multigrid Strategy: Definition, Symbol Analysis, and Numerics

Let us consider the following family of Finite Element functional spaces of the form (V.5):

$$\left\{\mathcal{W}_{2^t}^{(s)}\right\}_{t=0,\ldots,\bar{t}}.$$

From the definition of $\mathcal{V}_n^{(s)}$ and $\mathcal{W}_n^{(s)}$, it is clear that the following inclusion property holds

$$\mathcal{W}_1^{(s)} \subseteq \mathcal{W}_2^{(s)} \subseteq \ldots \subseteq \mathcal{W}_{2^{\bar{t}-1}}^{(s)} \subseteq \mathcal{W}_{2^{\bar{t}}}^{(s)}.$$

Therefore, to formulate a multigrid strategy, it is quite natural to follow a functional approach and to impose the prolongation operator $\mathcal{P}_{t,t+1} : \mathcal{W}_{2^t}^{(s)} \to \mathcal{W}_{2^{t+1}}^{(s)}$ to be defined as the identity operator, that is

$$\mathcal{P}_{t,t+1} v_t = v_t, \quad \text{for all } v_t \in \mathcal{W}_{2^t}^{(s)}.$$

Thus, the matrix representing the prolongation operator is formed, column by column, by representing each function of the basis of the coarse space $\mathcal{W}_{2^t}^{(s)}$ as linear combination of the basis of the fine space $\mathcal{W}_{2^{t+1}}^{(s)}$, the coefficients being the values of the functions $\varphi_i^{(2^t),s}$ on the fine mesh grid points, that is,

$$\varphi_i^{(2^t),s}(x) = \sum_{j=0}^{s2^{t+1}} \varphi_i^{(2^t),s}\left(\frac{j}{2^{t+1}s}\right) \varphi_j^{(2^{t+1}),s}(x). \tag{V.16}$$

In the following subsections, we consider in detail the case of $\mathbb{Q}_s$ Finite Element approximation with $s = 2$ and $s = 3$, the case $s = 1$ being reported in short just for the sake of completeness.

V.2.1 $\mathbb{Q}_1$ Case

Firstly, let us consider the case of $\mathbb{Q}_1$ Finite Elements, where, as it is well known, the stiffness matrix is the scalar Toeplitz matrix generated by $\mathbf{f}_{\mathbb{Q}_1}(\vartheta) = 2 - 2\cos(\vartheta)$, and, for the sake of simplicity, let us consider the spaces $\mathcal{W}_4^{(1)}$ and $\mathcal{W}_8^{(1)}$ with respective dimension 3 and 7. In the standard geometric multigrid, the prolongation operator matrix is defined as

$$P_{n,m}^s = P_{7,3}^1 = \begin{bmatrix} \frac{1}{2} & & \\ 1 & & \\ \frac{1}{2} & \frac{1}{2} & \\ & 1 & \\ & \frac{1}{2} & \frac{1}{2} \\ & & 1 \\ & & \frac{1}{2} \end{bmatrix}. \tag{V.17}$$

Indeed, for polynomial degree equal to 1, the basis functions with respect to the reference interval $[0, 1]$ are $\hat{\varphi}_1(\hat{x}) = 1 - \hat{x}$, $\hat{\varphi}_2(\hat{x}) = \hat{x}$, and, according to Equation (V.16), the $\varphi_i^{4,1}$ coefficients are

$$\hat{\varphi}_2(1/2) = 1/2, \quad \hat{\varphi}_2(1) = 1, \quad \hat{\varphi}_1(1/2) = 1/2,$$

Table V.1: Number of iterations needed for the convergence of the two-grid and V-cycle methods for $s = 1, 2, 3$ in one dimension with $a(x) \equiv 1$ and $\varepsilon = 1 \times 10^{-6}$.

	$s=1$		$s=2$		$s=3$	
# Subintervals	TGM	V-Cycle	TGM	V-Cycle	TGM	V-Cycle
16	6	7	7	7	9	9
32	7	7	7	7	9	9
64	7	7	7	7	9	9
128	6	7	7	7	9	9
256	6	7	7	7	9	9
512	6	7	7	7	9	9

giving the columns of the matrix in Equation (V.17).

However, our aim is to write $P^s_{n,m}$ in the form of equation (IV.4). For the latter purpose, we think of the prolongation matrix above as the product of the Toeplitz matrix generated by the polynomial $p_{G_1}(\vartheta) = 1 + \cos(\vartheta)$, where the subscript G_1 stands for "Geometric for polynomial degree 1", and the cutting matrix $K^{Even}_{n,m}$, that is, $P^1_{n,m} = T_n\left[p_{G_1}\right] K^{Even}_{n,m}$.

The two-grid and multigrid convergence with the above defined restriction/prolongation operators and a simple smoother (for instance, a Gauss–Seidel iteration) is a classical result, both from the point of view of the literature of approximated differential operators [69] and from the point of view of the literature of structured matrices [4, 56].

In the first panel of Table V.1, we report the number of iterations needed for achieving the predefined tolerance $\varepsilon = 10^{-6}$, when increasing the matrix size in the setting of the current subsection. Indeed, for the two-grid method we use $P^1_{n,m} = T_n\left[p_{G_1}\right] K^{Even}_{n,m}$ and its transpose as prolongation and restriction operators and Gauss–Seidel as a smoother. We highlight that only one iteration of pre-smoothing and only one iteration of post-smoothing are employed in the current numerics. In this scalar setting, it is straightforward to see that the conditions in (IV.1) are fulfilled, and hence there is no surprise in observing that the number of iterations needed for the two-grid remains almost constant when we increase the matrix size, numerically confirming the predicted optimality of the methods. Moreover, we obtain an analogous optimal behaviour also for the V-cycle method implemented with the same prolongation, restriction, and smoothing strategies at each level and this is expected from the analysis in [4].

We remark that we consider the one-dimensional case for the theoretical development of the method, which can be extended to more dimensions through a tensor argument, as we detail in Subsection V.4.3.

V.2.2 $\mathbb{Q}_2$ Case

Let us consider the case of $\mathbb{Q}_2$ Finite Elements, where we have that the basis functions with respect to the reference interval $[0, 1]$ are

$$\begin{aligned}\hat{\varphi}_1(\hat{x}) &= 2\hat{x}^2 - 3\hat{x} + 1,\\ \hat{\varphi}_2(\hat{x}) &= -4\hat{x}^2 + 4\hat{x},\\ \hat{\varphi}_3(\hat{x}) &= 2\hat{x}^2 - \hat{x}.\end{aligned}$$

For the sake of simplicity, let us consider the spaces $\mathcal{W}_2^{(2)}$ and $\mathcal{W}_4^{(2)}$ with respective dimension 3 and 7. Thus, with respect to Equation (V.16), the $\varphi_1^{2,2}$ coefficients are

$$\hat{\varphi}_2(1/4) = 3/4, \quad \hat{\varphi}_2(1/2) = 1, \quad \hat{\varphi}_2(3/4) = 3/4, \quad \hat{\varphi}_2(1) = 0,$$

while the $\varphi_2^{2,2}$ coefficients are

$$\begin{aligned} &\hat{\varphi}_3(1/4) = -1/8, \quad \hat{\varphi}_3(1/2) = 0, \quad \hat{\varphi}_3(3/4) = 3/8, \quad \hat{\varphi}_3(1) = 1, \\ &\hat{\varphi}_1(1/4) = 3/8, \quad \hat{\varphi}_1(1/2) = 0, \quad \hat{\varphi}_1(3/4) = -1/8, \quad \hat{\varphi}_1(1) = 0, \end{aligned}$$

and so on again as for that first couple of basis functions. Notice also that, to evaluate the coefficients, for the sake of simplicity, we are referring to the basis functions on the reference interval, as depicted in Figure V.1. Summarizing, the obtained prolongation matrix is as follows

$$P_{n,m}^s = P_{7,3}^2 = \begin{bmatrix} \frac{3}{4} & -\frac{1}{8} & \\ 1 & 0 & \\ \frac{3}{4} & \frac{3}{8} & \\ 0 & 1 & \\ & \frac{3}{8} & \frac{3}{4} \\ & 0 & 1 \\ & -\frac{1}{8} & \frac{3}{4} \end{bmatrix}. \tag{V.18}$$

Hereafter, we are interested in setting such a geometrical multigrid strategy, proposed in [26, 69, 70], in the framework of the more general algebraic multigrid theory and in particular in the one driven by the matrix symbol analysis. To this end, we represent the prolongation operator quoted above as the product of a Toeplitz matrix generated by a polynomial $\mathbf{p}_{G_2}$ and a suitable cutting matrix, following the theory in **Chapter IV**. We recall that the Finite Element stiffness matrix could be thought as a principal submatrix of a Toeplitz matrix generated by the matrix-valued symbol that, from Equation (V.12), has the compact form

$$\mathbf{f}_{\mathbb{Q}_2}(\vartheta) = \begin{bmatrix} \frac{16}{3} & -\frac{8}{3}(1+\mathrm{e}^{\hat{\imath}\vartheta}) \\ -\frac{8}{3}(1+\mathrm{e}^{-\hat{\imath}\vartheta}) & \frac{14}{3}+\frac{1}{3}(\mathrm{e}^{\hat{\imath}\vartheta}+\mathrm{e}^{\hat{\imath}\vartheta}) \end{bmatrix}. \tag{V.19}$$

Then, it is quite natural to look for a matrix-valued symbol for the polynomial $\mathbf{p}_{G_2}$ as well. In addition, the cutting matrix is also formed through the Kronecker product of the scalar cutting matrix $K_{n,m}^{Even}$ and the identity matrix of order 2, so that

$$P_{n,m}^2 = T_n\left[\mathbf{p}_{G_2}\right]\left(K_{n,m}^{Even} \otimes I_2\right).$$

Taking into account the action of the cutting matrix $K_{n,m}^{Even} \otimes I_2$, we can easily identify from Equation (V.18) the generating polynomial as

$$\mathbf{p}_{G_2}(\vartheta) = K_0 + K_1\mathrm{e}^{\hat{\imath}\vartheta} + K_{-1}\mathrm{e}^{-\hat{\imath}\vartheta} + K_2\mathrm{e}^{2\hat{\imath}\vartheta} + K_{-2}\mathrm{e}^{-2\hat{\imath}\vartheta}. \tag{V.20}$$

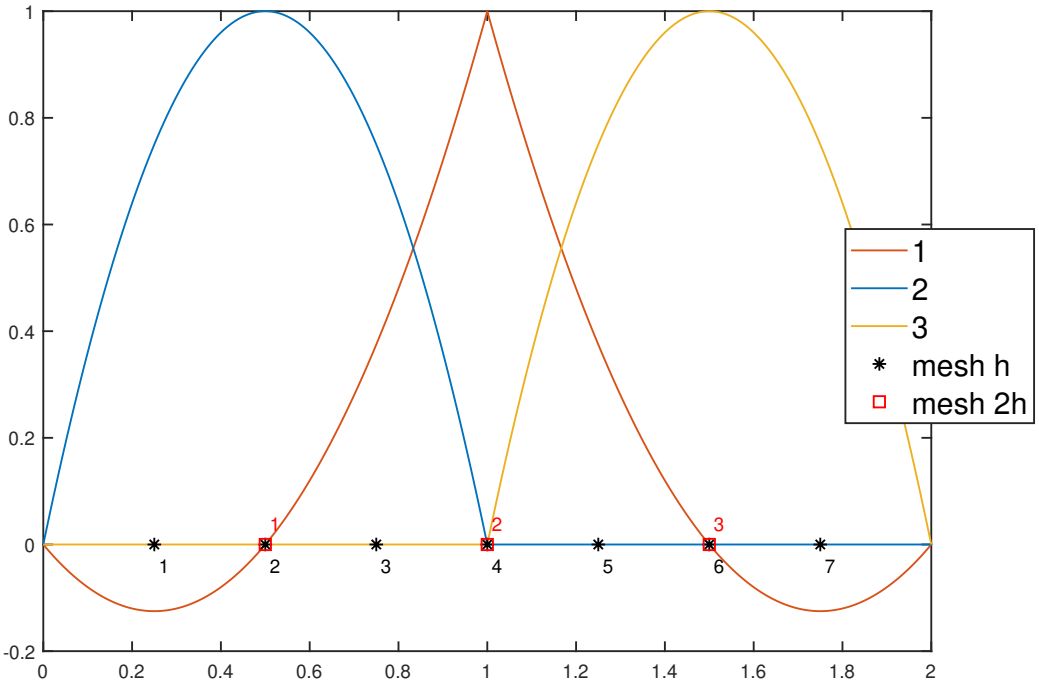

Figure V.1: Construction of the $\mathbb{Q}_2$ prolongation operator: basis functions on the reference element.

where

$$K_0 = \begin{bmatrix} \frac{3}{4} & \frac{3}{8} \\ 0 & 1 \end{bmatrix},\ K_1 = \begin{bmatrix} 0 & \frac{3}{8} \\ 0 & 0 \end{bmatrix},\ K_{-1} = \begin{bmatrix} \frac{3}{4} & -\frac{1}{8} \\ 1 & 0 \end{bmatrix},\ K_2 = \begin{bmatrix} 0 & -\frac{1}{8} \\ 0 & 0 \end{bmatrix},\ K_{-2} = O_{2,2},$$

that is

$$\mathbf{p}_{G_2}(\vartheta) = \begin{bmatrix} \frac{3}{4}(1+\mathrm{e}^{-\hat{\imath}\vartheta}) & \frac{3}{8}(1+\mathrm{e}^{\hat{\imath}\vartheta}) - \frac{1}{8}(\mathrm{e}^{-\hat{\imath}\vartheta}+\mathrm{e}^{2\hat{\imath}\vartheta}) \\ \mathrm{e}^{-\hat{\imath}\vartheta} & 1 \end{bmatrix}.$$

A very preliminary analysis, just by computing the determinant of $\mathbf{p}_{G_2}(\vartheta)$ shows there is a zero of third order in the mirror point $\vartheta = \pi$, being

$$\det(\mathbf{p}_{G_2}(\vartheta)) = \frac{1}{8}\mathrm{e}^{-2\hat{\imath}\vartheta}(\mathrm{e}^{\hat{\imath}\vartheta}+1)^3.$$

Moreover, we can provide a more rigorous convergence analysis if we recall Theorem IV.3.3. To this end, we have explicitly formed the matrices involved in equations (IV.6) and (IV.13) and computed their eigenvalues for $\vartheta \in [0, 2\pi]$. The results are reported in Figure V.2 and are in perfect agreement with the theoretical requirements. Indeed, by Theorem IV.3.3, our projection strategy for the $\mathbb{Q}_2$ FEM linear systems is such that the approximation property is fulfilled.

In the second panel of Table V.1, we report the number of iterations needed for achieving the predefined tolerance $\varepsilon = 10^{-6}$, when increasing the matrix size in the setting of the current subsection. Indeed, we use $T_n\left[\mathbf{p}_{G_2}\right]\left(K_{n,m}^{Even} \otimes I_2\right)$ and its transpose as prolongation and restriction operators and Gauss–Seidel as a smoother. Again, we remind that only one iteration

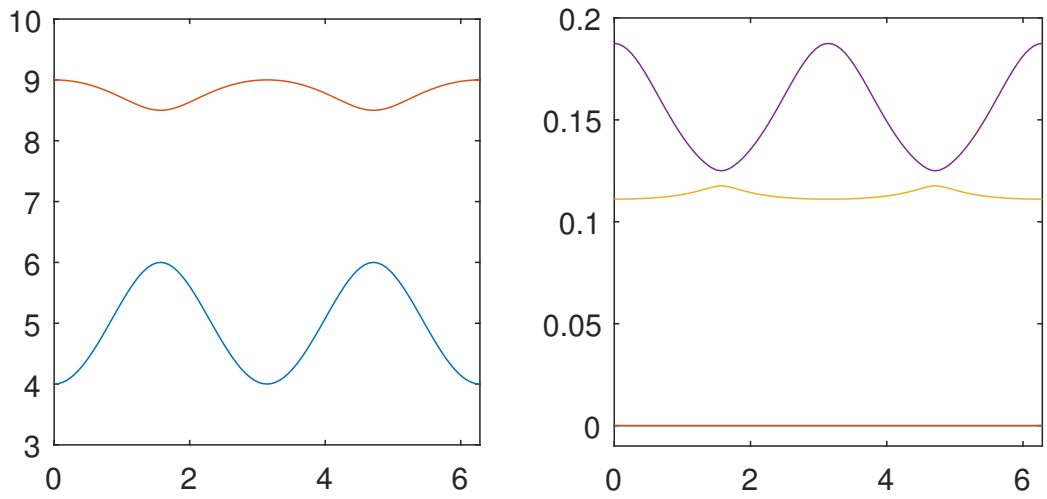

Figure V.2: Check of conditions for $\mathbb{Q}_2$ geometric prolongation: (**left**) the plot of the eigenvalues of $\mathbf{p}_{G_2}(\vartheta)^H\mathbf{p}_{G_2}(\vartheta) + \mathbf{p}_{G_2}(\vartheta+\pi)^H\mathbf{p}_{G_2}(\vartheta+\pi)$ for $\vartheta \in [0, 2\pi]$; and (**right**) the plot of the eigenvalues of $R(\vartheta)$ for $\vartheta \in [0, 2\pi]$.

of pre-smoothing and only one iteration of post-smoothing are employed in our numerical setting. As expected, we observe that the number of iterations needed for the two-grid convergence remains constant when we increase the matrix size, numerically confirming the optimality of the method.

Moreover, we notice that also the V-cycle method possesses optimal convergence properties. Although this behaviour is expected from the point of view of differential approximated operators, it is of particular interest in the setting of algebraic multigrid methods. Indeed, constructing an optimal V-cycle method for matrices in this block setting requires a further analysis of the spectral properties of the restricted operators, as we see in Section V.4.

Furthermore, we highlight that the presented analysis for $a \equiv 1$ can be easily extended to the case of non-constant coefficients $a(x) \neq 1$, since, following a geometric approach, the prolongation operators for the general variable coefficients remain unchanged. In Table V.2, we show the number of iterations needed for the convergence of the two-grid and V-cycle methods for $k = 2$ for different values of $a \not\equiv 1$.

We remark that we consider the one-dimensional case for the theoretical development of the method, which can be extended to more dimensions through a tensor argument.

V.2.3 $\mathbb{Q}_3$ Case

Hereafter, we briefly summarize the case of $\mathbb{Q}_3$ Finite Elements, following the very same path we already considered in the previous section for $\mathbb{Q}_2$ Finite Elements. The basis functions with

Table V.2: Number of iterations needed for the convergence of the two-grid and V-cycle methods for $s = 2$ in one dimension with $a(x) = \mathrm{e}^x$, $a(x) = 10x + 1$, $a(x) = |x - 1/2| + 1$, and $\varepsilon = 1 \times 10^{-6}$.

	$a(x) = \mathrm{e}^x$		$a(x) = 10x + 1$		$a(x) = \lvert x - 1/2\rvert + 1$	
# Subintervals	TGM	V-Cycle	TGM	V-Cycle	TGM	V-Cycle
16	7	7	9	12	7	7
32	7	8	7	14	7	7
64	7	8	7	14	7	7
128	7	8	7	15	7	7
256	7	8	7	15	7	7
512	7	8	7	14	7	7

respect to the reference interval $[0, 1]$ are

$$\begin{aligned}
\hat{\varphi}_1(\hat{x}) &= -\frac{9}{2}\hat{x}^3 + 9\hat{x}^2 - \frac{11}{2}\hat{x} + 1, \\
\hat{\varphi}_2(\hat{x}) &= \frac{27}{2}\hat{x}^3 - \frac{45}{2}\hat{x}^2 + 9\hat{x}, \\
\hat{\varphi}_3(\hat{x}) &= -\frac{27}{2}\hat{x}^3 + 18\hat{x}^2 - \frac{9}{2}\hat{x}, \\
\hat{\varphi}_4(\hat{x}) &= \frac{9}{2}\hat{x}^3 - \frac{9}{2}\hat{x}^2 + \hat{x}.
\end{aligned} \tag{V.21}$$

For the sake of simplicity, we consider the functional spaces $\mathcal{W}_2^{(3)}$ and $\mathcal{W}_4^{(3)}$ with respective dimension 5 and 11. Thus, with respect to equation (V.16) (see also Figure V.3), the $\varphi_1^{2,3}$ coefficients are

$$\begin{aligned}
&\hat{\varphi}_2(1/6) = 15/16, \quad \hat{\varphi}_2(1/3) = 1, \quad \hat{\varphi}_2(1/2) = 9/16, \\
&\hat{\varphi}_2(2/3) = 0, \quad \hat{\varphi}_2(5/6) = -5/16, \quad \hat{\varphi}_2(1) = 0,
\end{aligned}$$

while, the $\varphi_2^{2,3}$ coefficients are

$$\begin{aligned}
&\hat{\varphi}_3(1/6) = -5/16, \quad \hat{\varphi}_3(1/3) = 0, \quad \hat{\varphi}_3(1/2) = 9/16, \\
&\hat{\varphi}_3(2/3) = 1, \quad \hat{\varphi}_3(5/6) = 15/16, \quad \hat{\varphi}_3(1) = 0,
\end{aligned}$$

and the $\varphi_3^{2,3}$ coefficients are

$$\begin{aligned}
&\hat{\varphi}_4(1/6) = 1/16, \quad \hat{\varphi}_4(1/3) = 0, \quad \hat{\varphi}_4(1/2) = -1/16, \\
&\hat{\varphi}_4(2/3) = 0, \quad \hat{\varphi}_4(5/6) = 5/16, \quad \hat{\varphi}_4(1) = 1, \\
&\hat{\varphi}_1(1/6) = 5/16, \quad \hat{\varphi}_1(1/3) = 0, \quad \hat{\varphi}_1(1/2) = -1/16, \\
&\hat{\varphi}_1(2/3) = 0, \quad \hat{\varphi}_1(5/6) = 1/16, \quad \hat{\varphi}_1(1) = 0.
\end{aligned}$$

Consequently, the obtained prolongation matrix is as follows:

$$P^s_{n,m} = P^3_{11,5} = \begin{bmatrix} \frac{15}{16} & -\frac{5}{16} & \frac{1}{16} & & & \\ 1 & 0 & 0 & & & \\ \frac{9}{16} & \frac{9}{16} & -\frac{1}{16} & & & \\ 0 & 1 & 0 & & & \\ -\frac{5}{16} & \frac{15}{16} & \frac{5}{16} & & & \\ 0 & 0 & 1 & & & \\ & & \frac{5}{16} & \frac{15}{16} & -\frac{5}{16} & \\ & & 0 & 1 & 0 & \\ & & -\frac{1}{16} & \frac{9}{16} & \frac{9}{16} & \\ & & 0 & 0 & 1 & \\ & & \frac{1}{16} & -\frac{5}{16} & \frac{15}{16} & \end{bmatrix}. \tag{V.22}$$

Thus, taking into consideration that the stiffness matrix is a principal submatrix of the Toeplitz matrix generated by the matrix-valued function

$$\mathbf{f}_{\mathbb{Q}_3}(\vartheta) = \begin{bmatrix} \frac{54}{5} & -\frac{297}{40} & \frac{27}{20} - \frac{189}{40}\mathrm{e}^{\hat{\imath}\vartheta} \\ -\frac{297}{40} & \frac{54}{5} & -\frac{189}{40} + \frac{27}{20}\mathrm{e}^{\hat{\imath}\vartheta} \\ \frac{27}{20} - \frac{189}{40}\mathrm{e}^{-\hat{\imath}\vartheta} & -\frac{189}{40} + \frac{27}{20}\mathrm{e}^{-\hat{\imath}\vartheta} & \frac{37}{5} - \frac{13}{40}(\mathrm{e}^{\hat{\imath}\vartheta} + \mathrm{e}^{-\hat{\imath}\vartheta}) \end{bmatrix}, \tag{V.23}$$

we are looking for the matrix-valued trigonometric polynomial $\mathbf{p}_{G_3}$ as well. By defining

$$P^3_{n,m} = T_n\left[\mathbf{p}_{G_3}\right]\left(K^{Even}_{n,m} \otimes I_3\right),$$

it is straightforward to identify the generating polynomial as

$$\mathbf{p}_{G_3}(\vartheta) = K_0 + K_1\mathrm{e}^{\hat{\imath}\vartheta} + K_{-1}\mathrm{e}^{-\hat{\imath}\vartheta} + K_2\mathrm{e}^{2\hat{\imath}\vartheta} + K_{-2}\mathrm{e}^{-2\hat{\imath}\vartheta}, \tag{V.24}$$

where

$$K_0 = \begin{bmatrix} 0 & 1 & 0 \\ -\frac{5}{16} & \frac{15}{16} & \frac{5}{16} \\ 0 & 0 & 1 \end{bmatrix},\ K_1 = \begin{bmatrix} 0 & 0 & \frac{5}{16} \\ 0 & 0 & 0 \\ 0 & 0 & -\frac{1}{16} \end{bmatrix},\ K_{-1} = \begin{bmatrix} \frac{15}{16} & -\frac{5}{16} & \frac{1}{16} \\ 1 & 0 & 0 \\ \frac{9}{16} & \frac{9}{16} & -\frac{1}{16} \end{bmatrix},$$

$$K_2 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & \frac{1}{16} \\ 0 & 0 & 0 \end{bmatrix},\quad K_{-2} = O_{3,3},$$

that is

$$\mathbf{p}_{G_3}(\vartheta) = \begin{bmatrix} \frac{15}{16}\mathrm{e}^{-\hat{\imath}\vartheta} & 1 - \frac{5}{16}\mathrm{e}^{-\hat{\imath}\vartheta} & \frac{1}{16}\mathrm{e}^{-\hat{\imath}\vartheta} + \frac{5}{16}\mathrm{e}^{\hat{\imath}\vartheta} \\ \mathrm{e}^{-\hat{\imath}\vartheta} - \frac{5}{16} & \frac{15}{16} & \frac{5}{16} + \frac{1}{16}\mathrm{e}^{2\hat{\imath}\vartheta} \\ \frac{9}{16}\mathrm{e}^{-\hat{\imath}\vartheta} & \frac{9}{16}\mathrm{e}^{-\hat{\imath}\vartheta} & 1 - \frac{1}{16}(\mathrm{e}^{\hat{\imath}\vartheta} + \mathrm{e}^{-\hat{\imath}\vartheta}) \end{bmatrix}. \tag{V.25}$$

Table V.3: Number of iterations needed for the convergence of the two-grid and V-cycle methods for $s = 2$ in one dimension with $a(x) \equiv 1$ and $\varepsilon = 1 \times 10^{-2}, 1 \times 10^{-4}$, and 1×10^{-8}.

	$\varepsilon = 1 \times 10^{-2}$		$\varepsilon = 1 \times 10^{-4}$		$\varepsilon = 1 \times 10^{-8}$	
# Subintervals	TGM	V-Cycle	TGM	V-Cycle	TGM	V-Cycle
16	3	3	5	5	9	9
32	3	3	5	5	9	10
64	3	3	5	5	9	10
128	3	3	5	5	9	10
256	3	3	5	5	9	10
512	3	3	5	5	9	10

A trivial computation shows that the determinant of $\mathbf{p}_{G_3}(\vartheta)$ has a zero of fourth order in the mirror point $\vartheta = \pi$, being

$$\det(\mathbf{p}_{G_3}(\vartheta)) = \frac{1}{64} \mathrm{e}^{-3\hat{\imath}\vartheta} (\mathrm{e}^{\hat{\imath}\vartheta} + 1)^4.$$

However, the main goal is to verify the conditions in equations (IV.6) and (IV.13): we have explicitly formed the matrices involved and computed their eigenvalues for $\vartheta \in [0, 2\pi]$. The results are in perfect agreement with the theoretical requirements (see Figure V.4). We remark again that the purpose of this analysis is to link the geometric approach proposed in [26, 69, 70] to the novel algebraic multigrid methods for block-Toeplitz matrices.

In the third panel of Table V.1, we report the number of iterations needed for achieving the predefined tolerance $\varepsilon = 10^{-6}$, when increasing the matrix size in the setting of the current subsection. Indeed, we use $T_n\left[\mathbf{p}_{G_3}\right]\left(K_{n,m}^{Even} \otimes I_3\right)$ and its transpose as prolongation and restriction operators and Gauss–Seidel as a smoother (one iteration of pre-smoothing and one iteration of post-smoothing).

As expected, we observe that the number of iterations needed for the two-grid convergence remains constant when we increase the matrix size, numerically confirming the optimality of the method. As in the $\mathbb{Q}_2$ case, we also notice that the V-cycle method possesses the same optimal convergence properties.

Comparing the three panels in Table V.1, we also notice a mild dependency of the number of iterations on the polynomial degree s. In addition, we can see in Tables V.3 and V.4 that the optimal behaviour of the two-grid and V-cycle methods for $s = 2, 3$ remains unchanged if we test different tolerance values.

It is worth stressing that the results also hold in dimension $k = 2$, as well shown in Table V.5. The same optimal behaviour in the sense of the convergence rate is present also in the case of non-constant coefficients $a(x, y) \neq 1$. Indeed, in Table V.6, we show the number of iterations needed for the convergence of the two-grid and V-cycle for different values of $a \not\equiv 1$.

We finally remind that the tensor structure of the resulting matrices highly facilitates the generalization and extension to the case of $k \geq 2$. Indeed, from a theoretical point of view, the tensor structure permits to exploit the results of Section IV.4. Moreover, from the practical point of view, the prolongation operators in the multilevel case are constructed by a proper tensorization of those in 1D, as we detail in Subsection V.4.3.

Table V.4: Number of iterations needed for the convergence of the two-grid and V-cycle methods for $s = 3$ in one dimension with $a(x) \equiv 1$ and $\varepsilon = 1 \times 10^{-2}, 1 \times 10^{-4}$, and 1×10^{-8}.

	$\varepsilon = 1 \times 10^{-2}$		$\varepsilon = 1 \times 10^{-4}$		$\varepsilon = 1 \times 10^{-8}$	
# Subintervals	TGM	V-Cycle	TGM	V-Cycle	TGM	V-Cycle
16	3	3	6	6	12	12
32	3	3	6	6	12	12
64	3	3	6	6	12	12
128	3	3	6	6	12	12
256	3	3	6	6	12	12
512	3	3	6	6	12	12

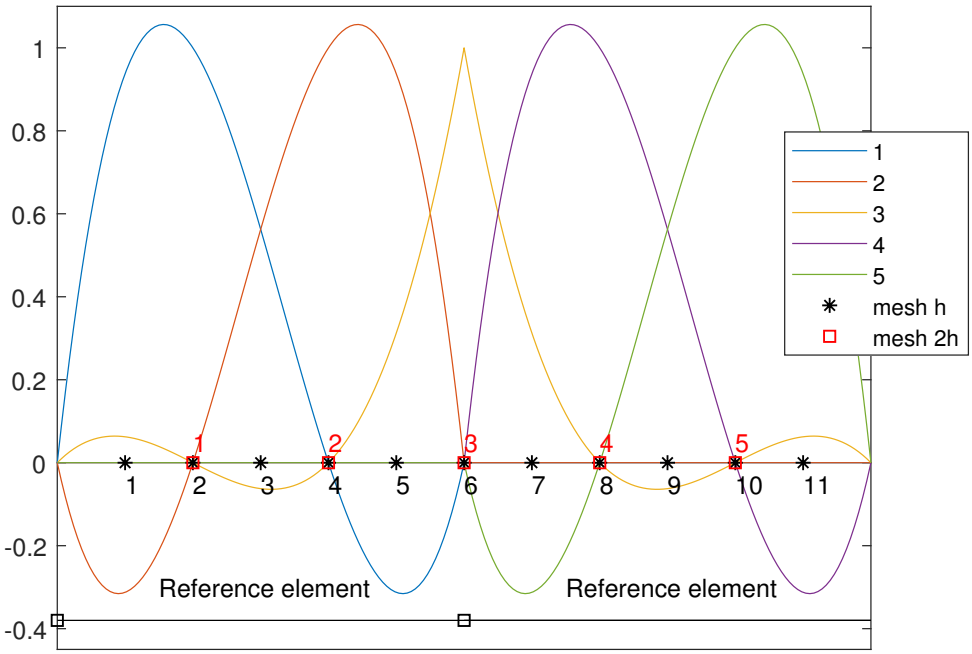

Figure V.3: Construction of the $\mathbb{Q}_3$ geometric prolongation operator: basis functions on the reference element.

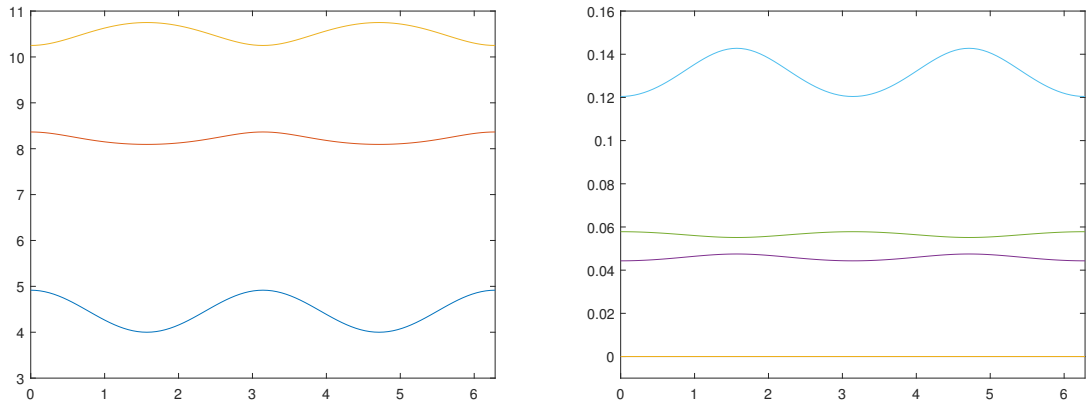

Figure V.4: Check of conditions for $\mathbb{Q}_3$ geometric prolongation: (**left**) the plot of the eigenvalues of $\mathbf{p}_{G_3}(\vartheta)^H \mathbf{p}_{G_3}(\vartheta) + \mathbf{p}_{G_3}(\vartheta + \pi)^H \mathbf{p}_{G_3}(\vartheta + \pi)$ for $\vartheta \in [0, 2\pi]$; and (**right**) the plot of the eigenvalues of $R(\vartheta)$ for $\vartheta \in [0, 2\pi]$.

Table V.5: Number of iterations needed for the convergence of the two-grid and V-cycle methods for $s = 1, 2, 3$ in dimension $k = 2$ with $a(\mathbf{x}) \equiv 1$.

$s=1$			$s=2$			$s=3$		
# Nodes	TGM	V-Cycle	# Nodes	TGM	V-Cycle	#	TGM	V-Cycle
7^2	5	5	15^2	6	6	23^2	7	7
15^2	5	6	31^2	6	6	47^2	7	7
31^2	5	6	63^2	6	6	95^2	7	7
63^2	5	6	127^2	6	6	191^2	7	7
127^2	5	6	255^2	6	6	383^2	7	7

Table V.6: Number of iterations needed for the convergence of the two-grid and V-cycle methods for $s = 2$ in two dimensions with $a(x, y) = \mathrm{e}^{(x+y)}$, $a(x, y) = 10(x + y) + 1$, $a(x, y) = |x - 1/2| + |y - 1/2| + 1$, $a(x, y) = 1$ if $x \leq 1/2$ and $y \leq 1/2$, 5000 otherwise, and $\varepsilon = 1 \times 10^{-6}$.

	$a(x,y) = \mathrm{e}^{(x+y)}$		$10(x+y)+1$		$\lvert x-1/2\rvert + \lvert y-1/2\rvert + 1$		$\begin{cases} 1 & x, y \leq 1/2 \\ 5000 & \text{otherwise} \end{cases}$	
# Nodes	TGM	V-Cycle	TGM	V-Cycle	TGM	V-Cycle	TGM	V-Cycle
7^2	6	6	6	6	6	6	6	6
15^2	6	6	6	6	6	6	6	6
31^2	6	6	6	6	6	6	6	6
63^2	6	6	6	6	6	6	6	6
127^2	6	6	6	6	6	6	6	6

V.3 Symbol Analysis of the Standard Bisection Grid Transfer Operator

In the present section we consider a standard grid transfer operator, that is, the bisection operator, which coincides with the interpolation operator of Subsection V.2.1. Indeed, for the $\mathbb{Q}_1$ stiffness matrices the approaches of the current and of the previous sections coincide. The new element here is that we use the same bisection operator for every polynomial degree, only paying attention to the choice of the correct matrix-sizes.

For our purposes, we can write the bisection operator in the following form

$$P^s_{n,m} = T_{sn}[2 + 2\cos(\vartheta)]K^{Even}_{sn,\frac{s(n-1)}{2}}. \tag{V.26}$$

We want to show that the latter operator fits into the block setting of the previous chapter. Hence, first we have to rewrite $P^s_{n,m}$ in the desired block form

$$P^s_{n,m} = T_n\left[\mathbf{p}_{L_s}\right](K^{Even}_{n,m} \otimes I_s),$$

which means that we want to find a matrix-valued trigonometric polynomial $\mathbf{p}_{L_s}$ such that the latter equation is true, with $P^s_{n,m}$ defined as in (V.26).

Recalling the action of the cutting matrices $(K_{n,m} \otimes I_s)$ seen in Section IV.2, we can see that $P^s_{n,m}$ can be rewritten in the desired block form with associated matrix-valued trigonometric polynomial $\mathbf{p}_{L_s}$ of the form

$$\mathbf{p}_{L_s}(\vartheta) = \hat{p}_0 + \hat{p}_{-1}\mathrm{e}^{-\hat{\imath}\vartheta} + \hat{p}_1\mathrm{e}^{\hat{\imath}\vartheta}, \tag{V.27}$$

where the expressions of the Fourier coefficients $\hat{p}_0, \hat{p}_{-1}, \hat{p}_1$ depend on whether the degree is even or odd. Indeed, we have the following two cases.

1. In the case of even degree s, we define

$$A_1 = T_s[2 + 2\cos(\vartheta)]K^{Even}_{s,\frac{s}{2}} = \begin{bmatrix} 1 & & & & \\ 2 & & & & \\ 1 & 1 & & & \\ & 2 & & & \\ & & & & \\ & & & 1 & 1 \\ & & & & 2 \end{bmatrix}_{s\times\frac{s}{2}}.$$

and we have

$$\hat{p}_0 = \left[\begin{array}{c|c} \mathbf{o}^T_{\frac{s}{2}-1} & 1 \\ \hline O_{\frac{s}{2}-1,\frac{s}{2}-1} & \mathbf{o}_{\frac{s}{2}-1} \end{array} \middle| A_1 \right], \quad \hat{p}_{-1} = \begin{bmatrix} A_1 & | & O_{s,\frac{s}{2}} \end{bmatrix} \tag{V.28}$$

$$\hat{p}_1 = \left[\begin{array}{c|c} \mathbf{o}^T_{s-1} & 1 \\ \hline O_{s-1,s-1} & \mathbf{o}_{s-1} \end{array}\right]. \tag{V.29}$$

Note that

$$\sum_{j=1}^{\frac{s}{2}} [A_1]_{1,j} = 1, \tag{V.30}$$

and for $i = 2, \ldots, s$

$$\sum_{j=1}^{\frac{s}{2}} [A_1]_{i,j} = 2. \tag{V.31}$$

2. In the case of odd degree s, we define

$$A_2 = T_s[2 + 2\cos(\vartheta)]K^{Odd}_{s,\frac{s+1}{2}} = \begin{bmatrix} 2 & & & & \\ 1 & 1 & & & \\ & 2 & & & \\ & & & & \\ & & & 1 & 1 \\ & & & & 2 \end{bmatrix}_{s\times\frac{s+1}{2}},$$

$$A_3 = T_s[2 + 2\cos(\vartheta)]K^{Even}_{s,\frac{s+1}{2}} = \begin{bmatrix} 1 & & & & \\ 2 & & & & \\ 1 & 1 & & & \\ & 2 & & & \\ & & & & \\ & & & 2 & \\ & & & 1 & 1 \end{bmatrix}_{s\times\frac{s+1}{2}}.$$

and we have

$$\hat{p}_0 = \begin{bmatrix} O_{s,\frac{s-1}{2}} & | & A_2 \end{bmatrix},$$

$$\hat{p}_{-1} = \begin{bmatrix} A_3 & | & O_{s,\frac{s-1}{2}} \end{bmatrix},$$

$$\hat{p}_1 = \left[\begin{array}{c|c} \mathbf{o}_{s-1}^T & 1 \\ \hline O_{s-1,s-1} & \mathbf{o}_{s-1} \end{array}\right].$$

Note that

$$\sum_{j=1}^{\frac{s+1}{2}} [A_3]_{1,j} = 1, \quad \sum_{j=1}^{\frac{s+1}{2}} [A_2]_{1,j} = 2, \tag{V.32}$$

and for $i = 2, \ldots, s$

$$\sum_{j=1}^{\frac{s+1}{2}} [A_2]_{i,j} = \sum_{j=1}^{\frac{s+1}{2}} [A_3]_{i,j} = 2. \tag{V.33}$$

In what follows we prove that $\mathbf{p}_{L_s}$ fulfils conditions *(i)–(iii)* of Subsection IV.2.2 and, in particular, in the following lemma we show that $\mathbf{p}_{L_r}$ satisfies the hypotheses of Lemma IV.3.6 for $\mathbf{f} = \mathbf{f}_{\mathbb{Q}_s}$.

Lemma V.3.1. *Let $\mathbf{p}_{L_s}$ be the $s \times s$ trigonometric polynomial defined in (V.27), and $\mathbf{e}_s = [1, \ldots, 1]^T$. Then*

1. $\mathbf{p}_{L_s}(0)\,\mathbf{e}_s = 4\,\mathbf{e}_s$.
2. $\mathbf{p}_{L_s}(\pi)\,\mathbf{e}_s = 0\,\mathbf{e}_s$.
3. $\mathbf{p}_{L_s}(0)^H\,\mathbf{e}_s = 4\,\mathbf{e}_s$.

Proof. The first two items are equivalent to require that the sum of the elements in each row of the matrices $\mathbf{p}_{L_s}(0)$ and $\mathbf{p}_{L_s}(\pi)$ is 4 and 0, respectively.

Hence, to prove *1.* it is sufficient to show that for every $i = 1, \ldots, s$ it holds

$$\sum_{j=1}^{s} [\mathbf{p}_{L_s}(0)]_{i,j} = 4.$$

From the expression of $\mathbf{p}_{L_s}(\vartheta)$ in (V.27) we have

$$\sum_{j=1}^{s} [\mathbf{p}_{L_s}(0)]_{i,j} = \sum_{j=1}^{s} [\hat{p}_0 + \hat{p}_1 + \hat{p}_{-1}]_{i,j}.$$

Then, we can exploit the structure of the Fourier coefficients $\hat{p}_{-1}$, $\hat{p}_0$, and $\hat{p}_1$ for even and odd degree. In particular, looking at the structure of the matrices A_1, A_2, A_3, and at relations (V.30)-(V.33), we have, for even degree s and for $i = 1, \ldots, s$,

$$\sum_{j=1}^{\frac{s}{2}} [\mathbf{p}_{L_s}(0)]_{i,j} = \sum_{j=1}^{\frac{s}{2}} [\hat{p}_0 + \hat{p}_1 + \hat{p}_{-1}]_{i,j} = \begin{cases} 1 + \left(2\sum_{j=1}^{\frac{s}{2}} [A_1]_{1,j}\right) + 1 = 4, & \text{for } i = 1 \\ \left(2\sum_{j=1}^{\frac{s}{2}} [A_1]_{i,j}\right) = 4, & \text{for } i = 2, \ldots, s \end{cases},$$

and, for odd degree s and for $i = 1, \dots, s$,

$$\sum_{j=1}^{\frac{s+1}{2}} [\mathbf{p}_{L_s}(0)]_{i,j} = \sum_{j=1}^{\frac{s+1}{2}} [\hat{p}_0 + \hat{p}_1 + \hat{p}_{-1}]_{i,j} = \begin{cases} \left(\sum_{j=1}^{\frac{s+1}{2}} [A_3]_{1,j} + [A_2]_{1,j}\right) + 1 = 4, & \text{for } i = 1 \\ \left(\sum_{j=1}^{\frac{s}{2}} [A_3]_{i,j} + [A_2]_{i,j}\right) = 4, & \text{for } i = 2, \dots, s \end{cases}.$$

The proof of *2.* can be repeated following the idea in *1.* and noting that

$$\mathbf{p}_{L_s}(\pi) = \hat{p}_0 - \hat{p}_1 - \hat{p}_{-1}.$$

Analogously, the third item can be proven following the same idea of *1.*, showing that the sum of the elements in each column of the matrices $\mathbf{p}_{L_s}(0)$ is 4. Since it is a straightforward computation, we omit the details. □

The latter result, together with Lemma IV.3.6 permits to conclude that $\mathbf{p}_{L_s}$ satisfies condition (ii), once that we prove that it satisfies condition (i), so that the matrix-valued function $\mathbf{r}$ is well-defined. By direct computation of the quantity $\mathbf{p}_{L_s}(\vartheta)^H \mathbf{p}_{L_s}(\vartheta) + \mathbf{p}_{L_s}(\vartheta+\pi)^H \mathbf{p}_{L_s}(\vartheta+\pi)$, we find that for both even and odd s we have

$$\mathbf{p}_{L_s}(\vartheta)^H \mathbf{p}_{L_s}(\vartheta) + \mathbf{p}_{L_s}(\vartheta+\pi)^H \mathbf{p}_{L_s}(\vartheta+\pi) = \begin{bmatrix} 12 & 2 & 0 & \dots & 2\mathrm{e}^{2\hat{\imath}\vartheta} \\ 2 & 12 & 2 & \dots & 0 \\ & & \ddots & & \\ 0 & & & 12 & 2 \\ 2\mathrm{e}^{-2\hat{\imath}\vartheta} & 0 & \dots & 2 & 12 \end{bmatrix},$$

which is clearly a definite positive matrix for all $\vartheta \in [0, 2\pi)$, so $\mathbf{p}_{L_s}$ satisfies condition (i). Then $\left(\mathbf{p}_{L_s}(\vartheta)^H \mathbf{p}_{L_s}(\vartheta) + \mathbf{p}_{L_s}(\vartheta+\pi)^H \mathbf{p}_{L_s}(\vartheta+\pi)\right)^{-1}$ is well-defined, for all $\vartheta \in [0, 2\pi)$ and the function $\mathbf{r}(\vartheta)$ defined in (IV.10) is well-defined as well.

We have to verify the limit condition *(iii)*, in order to ensure that the bisection operator fulfils the approximation property for the $\mathbb{Q}_s$ linear system by Theorem IV.3.5. For this purpose, it is sufficient to show that the function $1 - \lambda_{\bar{\jmath}}(\mathbf{r}(\vartheta))$ has a zero at least of the same order of $\lambda_{\bar{\jmath}}(\mathbf{f}_{\mathbb{Q}_s}(\vartheta))$.

For even polynomial degree s, we have that $\mathbf{r}(\vartheta)$ is a projector, since it can be easily verified by direct computation that

$$\mathbf{r}^2(\vartheta) - \mathbf{r}(\vartheta) = O_{s,s}, \qquad \text{for all } \vartheta \in [0, 2\pi).$$

Hence, from condition *(ii)*, we have $\lambda_{\bar{\jmath}}(\mathbf{r}(0)) = 1$, and, from the continuity of the eigenvalue functions (Lemma I.3.1), we have that $\lambda_{\bar{\jmath}}(\mathbf{r}(\vartheta)) \equiv 1$. Hence, it is straightforward to see that condition *(iii)* is verified.

The proof of condition *(iii)* for odd polynomial degree s is under investigation in [20].

As a numerical confirmation of the validity of the projection strategy that we presented in the current section, we report the results of the two-grid procedure applied to the $\mathbb{Q}_s$ linear systems in Table V.7. In particular, we report the number of iterations needed for achieving the predefined tolerance 10^{-6} when increasing the matrix–size, using Gauss–Seidel as a smoother, with only one iteration of pre-smoothing and only one iteration of post-smoothing. Even though we do not present the theoretical convergence analysis of the V-cycle method, we report also the V-cycle tests. The results are comparable to those of Table V.1.

Table V.7: Number of iterations needed for the convergence of the two-grid and V-cycle methods for $s = 1, 2, 3$ in one dimension with $a(x) \equiv 1$ and $\varepsilon = 1 \times 10^{-6}$, using the standard bisection grid transfer operator.

	$s=1$		$s=2$		$s=3$	
# Subintervals	TGM	V-Cycle	TGM	V-Cycle	TGM	V-Cycle
8	5	5	8	8	9	10
16	6	7	8	9	9	10
32	7	7	8	9	9	10
64	7	7	8	9	9	10
128	6	7	8	9	9	10
256	6	7	8	9	9	10
512	6	7	8	9	9	10

V.4 A New Multigrid Strategy: Construction, Analysis, and Numerics

The scope of the current section is twofold. On one hand, we intend to give further numerical evidence of the results proven in Section IV.3. On the other hand, we present a new multigrid strategy for the $\mathbb{Q}_s$ Finite Element matrices. Our final goal is not to improve the convergence results of the geometric and standard bisection projection strategies, but to provide a general method to construct suitable grid transfer operators on the base of algebraic considerations related to **Chapter IV**.

Contrary to the theoretical analysis that we performed in the last chapter, here we deal with block-Toeplitz matrices generated by a matrix-valued trigonometric polynomial, instead of block-circulant matrices. As in Sections V.2 and V.3, we expect that the theoretical results of Section IV.3 still hold since block Toeplitz matrices with polynomial generating functions are a low rank correction of block circulant matrices with the same generating function. The only difference could be a slight deterioration of the convergence in the case of block Toeplitz matrices with respect to block circulant matrices.

As far as the choice of the right-hand side is concerned, we impose that the solution x of the linear system $T_n[\mathbf{f}]x = b$ is a uniform sampling of the sine function on $[0, \pi]$ and we compute the right-hand side b as $b = T_n[\mathbf{f}]x$. Moreover, in all the considered examples the used stopping criterion is a standard one, that is $\frac{\|r^{(k)}\|_2}{\|b\|_2} < \varepsilon$, where $r^{(k)} = b - T_n[\mathbf{f}]x^{(k)}$. As we mentioned in Section IV.2, the structure of the projector slightly changes for block-Toeplitz matrices, in order to preserve the structure at coarser levels. The dimension of the problem at level t becomes ns, with n of the form $2^t - 1$. The cutting matrix $K_{n,m}^{Even}$ is defined as in Section IV.2 and, for a matrix-valued trigonometric polynomial $\mathbf{p}$, the projector $P_{n,m}^s$ is of the form in (IV.4).

The section is outlined as follows. In Subsection V.4.1, we give a general methodology to construct a grid transfer operator for problems of Laplacian type. In Subsection V.4.2 we present strategies for an implementation of both the two-grid and V-cycle methods for $\mathbb{Q}_s$ stiffness matrices for the considered second order elliptic differential problem on $[0, 1]$. In Subsection V.4.3 we consider the two-dimensional problem, that is, we study multigrid methods for the $\mathbb{Q}_s$ stiffness matrices for the second order elliptic differential problem on the unit square.

Apart from the first example, we use the Gauss-Seidel method as a smoother. The method

damps the high frequencies, which makes it a suitable smoother for our problems.

V.4.1 A Strategy to Achieve Optimality of the V-Cycle

In the current section we consider a problem of Laplacian type, that is, we deal with matrices $T_n[\mathbf{f}] \geq 0$ generated by a trigonometric polynomial $\mathbf{f} : [-\pi, \pi] \to \mathbb{C}^{s\times s}$, $\mathbf{f} \geq 0$, that has a non-negative minimal eigenvalue function $\lambda_{\min}(\mathbf{f})$ with a unique zero in the origin of order two. Moreover, it holds $\mathbf{f}(0)\mathbf{e}_s = 0\mathbf{e}_s$, where $\mathbf{e}_s$ is the vector of all ones of length s.

We recall that the V-cycle method involves a set of coarser linear operators $T_{m_\ell}[\mathbf{f}_\ell]$, where ℓ represents the level. In order to define a robust projector $P_{n,m}$ that ensures a linear convergence rate also for the V-cycle applied to $T_n[\mathbf{f}]$, we study the quantity

$$\kappa(\mathbf{f}_\ell) = \frac{\|\lambda_{\max}(\mathbf{f}_\ell)\|_\infty}{\lambda''_{\min}(\mathbf{f}_\ell)\big|_0},$$

which gives an estimate of the ill-conditioning of the coarse problem at level ℓ, see [21]. Indeed the conditioning of the matrix $T_{m_\ell}[\mathbf{f}_\ell]$ depends on $\|\lambda_{\max}(\mathbf{f}_\ell)\|_\infty$ and $\lambda''_{\min}(\mathbf{f}_\ell)|_0$, which measure the magnitude of the maximum eigenvalue function $\lambda_{\max}(\mathbf{f}_\ell)$ and how flat the minimal eigenvalue function is around the origin, respectively. Note that the flatness of the minimal eigenvalue function is crucial to identify how large is the ill-conditioned subspace where the smoother cannot be effective. Therefore, the projector should be defined in order to keep $\kappa(\mathbf{f}_\ell)$ as small as possible. Note that the V-cycle convergence analysis is not straightforward, even using recent results based on the two-grid analysis like in [97], and it is under investigation in [20].

We select a class of projectors $P_{m_\ell,m_{\ell+1}}(z) = T_{m_\ell}[\mathbf{p}_z](K^{Even}_{m_\ell,m_{\ell+1}} \otimes I_s)$ according to the theoretical analysis of Section IV.2 with $\mathbf{p}_z(\cdot)$ of form

$$\mathbf{p}_z(\vartheta) = (1+\cos\vartheta)\left(I_s + \frac{z-1}{s}\mathbf{e}_s\mathbf{e}_s^T\right), \quad z > 0. \tag{V.34}$$

Note that $\mathbf{e}_s\mathbf{e}_s^T$ is the $s \times s$ matrix of all ones. Then, we have

$$\begin{aligned}
\mathbf{p}_z(\vartheta) &= \begin{bmatrix} 1+\cos\vartheta & & \\ & \ddots & \\ & & 1+\cos\vartheta \end{bmatrix} + (1+\cos(\vartheta))(z-1)\begin{bmatrix} \frac{1}{s} & \cdots & \cdots & \frac{1}{s} \\ \vdots & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ \frac{1}{s} & \cdots & \cdots & \frac{1}{s} \end{bmatrix} = \\
&= F_s\left(\begin{bmatrix} 1+\cos\vartheta & & \\ & \ddots & \\ & & 1+\cos\vartheta \end{bmatrix} + (1+\cos(\vartheta))(z-1)\begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 0 \end{bmatrix}\right)F_s^H = \\
&= F_s\begin{bmatrix} z+z\cos\vartheta & & & \\ & 1+\cos\vartheta & & \\ & & \ddots & \\ & & & 1+\cos\vartheta \end{bmatrix}F_s^H.
\end{aligned} \tag{V.35}$$

Hence the eigenvalue functions of $\mathbf{p}_z(\cdot)$ have a zero at π of order two for all $z > 0$ and it is straightforward to prove that the matrix-valued function $\mathbf{r}$ defined in Subsection IV.2.2 fulfils conditions *(i)-(iii)*.

In the next section we study the conditioning $\kappa(\mathbf{f}_{z,\ell})$, where $\mathbf{f}_{z,\ell}$ is the generating function at level ℓ obtained using $\mathbf{p}_z(\cdot)$, proving as it influences the V-cycle convergence. In particular, we look for a $z > 0$ such that

$$\lim_{\ell\to\infty} \lambda''_{\min}(\mathbf{f}_{z,\ell})\big|_0 > 0 \tag{V.36}$$

that guarantees that the behaviour of the minimal eigenvalue function around the origin remains unchanged at the coarser levels.

V.4.2 The One-Dimensional Case

Consider the $\mathbb{Q}_s$ approximation of the second order elliptic differential problem (V.2). As we outlined in Section V.1, the resulting stiffness matrix of size $(s \cdot n - 1) \times (s \cdot n - 1)$ is $nA_n^{(s)}$, where $A_n^{(s)}$ is the block matrix

$$A_n^{(s)} = T_n[\mathbf{f}]_-,$$

with the subscript $-$ denoting that the last row and column of $T_n[\mathbf{f}]$ are removed. This is because of the homogeneous boundary conditions. Moreover, we recall that the $s \times s$ matrix-valued generating function of $T_n[\mathbf{f}]$ has the form

$$\mathbf{f}(\vartheta) = \hat{f}_0 + \hat{f}_1 \mathrm{e}^{\hat{\imath}\vartheta} + \hat{f}_1^{\mathrm{T}} \mathrm{e}^{-\hat{\imath}\vartheta}.$$

TGM in the $s = 2$ setting

In the case of polynomial degree $s = 2$, the explicit expression of the generating function $\mathbf{f}$ can be seen in equation (V.19). In particular, the Fourier coefficients $\hat{f}_0$, $\hat{f}_1$ are given by

$$\hat{f}_0 = \frac{1}{3}\begin{bmatrix} 16 & -8 \\ -8 & 14 \end{bmatrix}, \quad \hat{f}_1 = \frac{1}{3}\begin{bmatrix} 0 & -8 \\ 0 & 1 \end{bmatrix}. \tag{V.37}$$

Moreover, it is possible to diagonalize $\mathbf{f}$ as

$$\mathbf{f}(\vartheta) = U(\vartheta) \begin{bmatrix} \lambda_1(\mathbf{f}(\vartheta)) & \\ & \lambda_2(\mathbf{f}(\vartheta)) \end{bmatrix} U^H(\vartheta),$$

where the eigenvalue functions $\lambda_1(\mathbf{f}(\vartheta)), \lambda_2(\mathbf{f}(\vartheta))$ of $\mathbf{f}$ are given explicitly by

$$\begin{aligned} \lambda_1(\mathbf{f}(\vartheta)) &= 5 + \frac{1}{3}\cos(\vartheta) - \frac{1}{3}\sqrt{129 + 126\cos(\vartheta) + \cos^2(\vartheta)}, \\ \lambda_2(\mathbf{f}(\vartheta)) &= 5 + \frac{1}{3}\cos(\vartheta) + \frac{1}{3}\sqrt{129 + 126\cos(\vartheta) + \cos^2(\vartheta)} \end{aligned}$$

and $U : [0, 2\pi) \to \mathbb{C}^{sn\times sn}$ is the matrix-valued function containing the eigenvectors of $\mathbf{f}$.

It is straightforward to verify that hypotheses requested in Section IV.2.1 that ensure the convergence and optimality of the TGM for $T_n[\mathbf{f}]$ are satisfied using the trigonometric polynomial

Table V.8: Number of iterations for the TGM using the $\mathbf{p}_z$ projection strategy for the $\mathbb{Q}_2$, using as pre- and post-smoother one iteration of Jacobi method with $\omega_{\text{pre}} = 7/8$, $\omega_{\text{post}} = 7/12$ and tolerance $\varepsilon = 10^{-7}$.

t	$n = 2^t - 1$	$2n$	$z = 1$	$z = 2$	$z = 3$	$z = 4$	$z = 5$
7	127	254	33	33	33	33	33
8	255	510	33	33	33	33	33
9	511	1022	33	33	33	33	33
10	1023	2046	33	33	33	33	33
11	2047	4094	33	33	33	33	33

$\mathbf{p}_z$ defined in (V.34) in the construction of the projector. Moreover, $\mathbf{f}(0)\mathbf{p}_z(0) = \mathbf{p}_z(0)\mathbf{f}(0)$ for every choice of $z > 0$ and hence $\mathbf{f}(0)$ and $\mathbf{p}_z(0)$ are simultaneously diagonalized by the same unitary transform. Therefore, we can control the ill-conditioning of the coarser problems in the subspace associated to $\vartheta = 0$ by taking different values of z, which is useful for the study of the V-cycle method as shown in Section V.4.1.

Now we implement a TGM for $T_n[\mathbf{f}]$ and we study the number of iterations that the method requires to reach the desired tolerance varying n and for different choices of z. In order to find the relaxation parameters for the Jacobi method we should compute the quantities in inequality (IV.11). We see from formula (V.19) that $\min_{j=1,\dots,s}\left([\hat{f}_0]_{(j,j)}\right)$ is equal to 14/3. For the computation of the quantity $\|\mathbf{f}\|_\infty = \max_{\vartheta\in[0,2\pi)} \|\mathbf{f}(\vartheta)\|_2$ we can write

$$\|\mathbf{f}\|_\infty = \max_{\vartheta\in[0,2\pi)} \lambda_2(\mathbf{f}(\vartheta)) = 5 + \frac{1}{3}\cos(0) + \frac{1}{3}\sqrt{129 + 126\cos(0) + \cos^2(0)} = \frac{32}{3}.$$

So, according to inequality (IV.11), our Jacobi relaxation parameter ω should be smaller than or equal to 7/8. In order to damp the error both in the middle and in the high frequencies, we take a different parameter for the pre-smoother and the post-smoother. For the pre-smoother, we take the greatest admissible value, $\omega_{\text{pre}} = 7/8$, and for the post-smoother we take $\omega_{\text{post}} = 2\omega_{\text{pre}}/3 = 7/12$.

In Tables V.8-V.9 we report for $z = 1, \dots, 5$ the number of iterations needed for achieving the tolerance $\varepsilon = 10^{-7}$ when increasing the matrix size and using $\mathbf{p}_z$ in the construction of the projector and with two different smoothers. Table V.8 shows the results using as pre- and post-smoother one iteration of the Jacobi method with relaxation parameters $\omega_{\text{pre}} = 7/8$ and $\omega_{\text{post}} = 7/12$. Table V.9 shows the results using as pre- and post-smoother one iteration of the Gauss-Seidel method with $\omega_{\text{pre,post}} = 1$.

As expected, in both cases we can observe that for all $z = 1, \dots, 5$ the number of iterations needed for the TGM convergence remains almost constant, when increasing the size sn, confirming the optimality of the method for every choice of z. Moreover, the number of iterations is essentially halved when using Gauss-Seidel instead of Jacobi.

V-cycle in the $s = 2$ setting

In order to maintain the optimality of the iterations also for the V-cycle we look for the best choice of the parameter z such that the behaviour of $\lambda_{\min}(\mathbf{f}_{z,\ell})$ around the origin remains

Table V.9: Number of iterations for the TGM using the $\mathbf{p}_z$ projection strategy for the $\mathbb{Q}_2$ stiffness matrix, using as pre- and post-smoother one iteration of Gauss-Seidel method with $\omega_{\text{pre,post}} = 1$ and tolerance $\varepsilon = 10^{-7}$.

t	$n = 2^t - 1$	$2n$	$z = 1$	$z = 2$	$z = 3$	$z = 4$	$z = 5$
7	127	254	15	15	15	15	15
8	255	510	15	15	15	15	15
9	511	1022	15	15	15	15	15
10	1023	2046	15	15	15	15	15
11	2047	4094	15	15	15	15	15

Table V.10: Condition numbers of $\mathbf{f}_{z,\ell}$ for $z = 1, 2, 3, 4$ and $\ell = 1, 2, 3, 4$.

j	$\kappa(\mathbf{f}_{1,\ell})$	$\kappa(\mathbf{f}_{2,\ell})$	$\kappa(\mathbf{f}_{3,\ell})$	$\kappa(\mathbf{f}_{4,\ell})$
1	43	11	4.7	4.7
2	171	11	4.7	4.7
3	683	11	4.7	4.7
4	2731	11	4.7	4.7

unchanged at the coarser levels, that is, for different choices of z, we check if $\lambda_{\min}(\mathbf{f}_{z,\ell})$ satisfies condition (V.36).

By direct computation, we derive the formula

$$\lambda''_{\min}(\mathbf{f}_{z,\ell})\big|_0 = \left(\frac{z^2}{2}\right)^j.$$

The latter implies that for values of z smaller than $\sqrt{2}$, the quantity $\lambda''_{\min}(\mathbf{f}_{z,\ell})\big|_0$ tends to zero as ℓ tends to ∞. This suggests that for $z < \sqrt{2}$ the conditioning becomes worse as the levels get coarser. This is numerically confirmed in Table V.10 where the condition numbers $\kappa(\mathbf{f}_{z,\ell})$ are listed for $z = 1, 2, 3, 4$ and $\ell = 1, 2, 3, 4$. Therefore we should avoid the choice $P^s_{n,m}(\mathbf{p}_1)$ as projector.

Indeed, Tables V.11-V.12 highlight that the number of iterations needed for the V-cycle convergence, with the desired tolerance, depends on the matrix size with $z = 1$, whereas it remains almost constant for $z > \sqrt{2}$ as n increases.

TGM and V-cycle in the $s > 2$ setting

We implemented the analogous TGM for polynomial degrees 3 and 4. From Tables V.13-V.14 we see that the number of iterations to achieve the desired tolerance still remains constant as the matrix size increases. However, we notice that this constant depends on the polynomial degree s. Achieving optimality from this point of view will be object of future research.

The analysis on the condition number that we exploited for $s = 2$ can be repeated assuming that Conjecture V.4.1 (numerically verified for $s = 3, 4$) holds.

Table V.11: Number of iterations for the V-cycle method using the $\mathbf{p}_z$ projection strategy for the $\mathbb{Q}_2$ stiffness matrix, pre- and post-smoother 1 iteration of Jacobi with $\omega_{\text{pre}} = 7/8$ and $\omega_{\text{post}} = 7/12$, tolerance $\varepsilon = 10^{-7}$.

t	$n = 2^t - 1$	$2n$	$z = 1$	$z = 2$	$z = 3$	$z = 4$	$z = 5$
7	127	254	1144	42	34	35	38
8	255	510	3365	45	35	35	37
9	511	1022	4000+	48	35	35	37
10	1023	2046	4000+	50	35	35	37
11	2047	4094	4000+	52	35	35	38
12	4095	8190	4000+	54	35	36	38
13	8191	16382	4000+	55	35	36	38

Table V.12: Number of iterations for the V-cycle method using the $\mathbf{p}_z$ projection strategy for the $\mathbb{Q}_2$ stiffness matrix, pre- and post-smoother 1 iteration of Gauss-Seidel with $\omega_{\text{pre,post}} = 1$, tolerance $\varepsilon = 10^{-7}$.

t	$n = 2^t - 1$	$2n$	$z = 1$	$z = 2$	$z = 3$	$z = 4$	$z = 5$
7	127	254	467	26	22	23	26
8	255	510	1343	29	23	26	28
9	511	1022	3992	31	24	28	30
10	1023	2046	4000+	33	27	29	32
11	2047	4094	4000+	35	28	30	33
12	4095	8190	4000+	36	29	31	34
13	8191	16382	4000+	38	29	32	34

Table V.13: Number of iterations for the TGM using the $\mathbf{p}_z$ projection strategy for the $\mathbb{Q}_3$ stiffness matrix, pre- and post-smoother 1 iteration of Gauss-Seidel with $\omega_{\text{pre,post}} = 1$, tolerance $\varepsilon = 10^{-7}$.

t	$n = 2^t - 1$	$3n$	$z = 1$	$z = 2$	$z = 3$	$z = 4$	$z = 5$
7	127	381	38	38	38	38	38
8	255	765	38	38	38	38	38
9	511	1533	38	38	38	38	38
10	1023	3069	38	38	38	38	38
11	2047	6141	38	38	38	38	38

Table V.14: Number of iterations for the TGM using the $\mathbf{p}_z$ projection strategy for the $\mathbb{Q}_4$ stiffness matrix, pre- and post-smoother 1 iteration of Gauss-Seidel with $\omega_{\text{pre,post}} = 1$, tolerance $\varepsilon = 10^{-7}$.

t	$n = 2^t - 1$	$4n$	$z = 1$	$z = 2$	$z = 3$	$z = 4$	$z = 5$
7	127	508	87	87	87	87	87
8	255	1020	87	87	87	87	87
9	511	2044	87	87	87	87	87
10	1023	4092	87	87	87	87	87
11	2047	8188	87	87	87	87	87

Conjecture V.4.1. *For every $s > 0$ and $z > 0$ there exists a constant $c_{z,s} > 0$ independent from j such that the following equality holds for all $j > 0$:*

$$\lambda''_{\min}(\mathbf{f}_{z,\ell})\big|_0 = c_{z,s}\left(\frac{z^2}{2}\right)^j.$$

The numerical experiments confirm the theoretical analysis associated with the previous conjecture, as we can see from the number of iterations obtained for $s = 3, 4$ in Tables V.15-V.16. Indeed, analogously to the case $s = 2$, we observe that we should avoid to take $z = 1$, for which $\lambda''_{\min}(\mathbf{f}_{z,\ell})\big|_0$ tends to 0 as ℓ tends to ∞.

Finally, we highlight that the slightly change of iterations, increasing t, is expected from the theory, since the exact constant which bounds the number of iterations can be reached for larger matrix-size values or studying the best choice of the smoother method and the relaxation parameters $\omega_{\text{pre,post}}$ in order to decrease such constant.

V.4.3 The Two-Dimensional Case

Consider the uniform $\mathbb{Q}_s$ approximation of the following particular case of problem (V.1):

$$\begin{cases} -\Delta u = \psi, & \text{in } \Omega := (0,1)^2, \\ u = 0, & \text{on } \partial\Omega, \end{cases} \tag{V.38}$$

where $\psi \in L^2(\Omega)$. Taking n elements in each direction, the resulting stiffness matrix of size $(s \cdot n - 1)^2 \times (s \cdot n - 1)^2$ is

$$\mathcal{A}_{\mathbf{n}}^{(s)} = A_n^{(s)} \otimes M_n^{(s)} + M_n^{(s)} \otimes A_n^{(s)}, \qquad N = (s \cdot n - 1)^2,$$

Table V.15: Number of iterations for the V-cycle method using the $\mathbf{p}_z$ projection strategy for the $\mathbb{Q}_3$ stiffness matrix, pre- and post-smoother 1 iteration of Gauss-Seidel with $\omega_{\text{pre,post}} = 1$, tolerance $\varepsilon = 10^{-7}$.

t	$n = 2^t - 1$	$3n$	$z = 1$	$z = 2$	$z = 3$	$z = 4$	$z = 5$
7	127	381	1180	51	43	44	46
8	255	765	3375	55	44	47	50
9	511	1533	4000+	59	45	51	52
10	1023	3069	4000+	63	47	52	54
11	2047	6141	4000+	66	50	54	56
12	4095	12285	4000+	69	53	55	57
13	8191	24573	4000+	72	53	57	59

Table V.16: Number of iterations for the V-cycle method using the $\mathbf{p}_z$ projection strategy for the $\mathbb{Q}_4$ stiffness matrix, pre- and post-smoother 1 iteration of Gauss-Seidel with $\omega_{\text{pre,post}} = 1$, tolerance $\varepsilon = 10^{-7}$.

t	$n = 2^t - 1$	$4n$	$z = 1$	$z = 2$	$z = 3$	$z = 4$	$z = 5$
7	127	508	2693	103	92	94	96
8	255	1020	4000+	108	94	96	97
9	511	2044	4000+	114	95	97	99
10	1023	4092	4000+	120	96	99	100
11	2047	8188	4000+	125	98	100	100
12	4095	16380	4000+	129	99	101	101
13	8191	32764	4000+	133	101	101	101

where $A_n^{(s)}$ and $M_n^{(s)}$ are the block-Toeplitz matrices

$$A_n^{(s)} = T_n[\mathbf{f}]_-, \qquad M_n^{(s)} = T_n[\mathbf{h}]_-,$$

with the subscript $-$ denoting again that the last row and column of $T_n[\mathbf{f}]$ are removed. Explicit formulae for the matrix-valued trigonometric polynomials $\mathbf{f}$ and $\mathbf{h}$ and the spectral distribution of the matrices are given in [64].

In the following we want to apply V-cycle strategy to the matrix $\mathcal{A}_{\mathbf{n}}^{(s)}$ which has a multilevel block structure, for different choices of s. In the one dimensional case, we took the block-Toeplitz matrix with block size s. In the two dimensional case, we take the actual matrices arising from the considered FEM approximation of problem (V.38), which are not pure block-Toeplitz matrices with block size s^2. However, we can still apply our multigrid procedure due to its spectral properties given in [64].

Since the matrices are cut and are the permutation of multilevel block-Toeplitz matrices, the projector slightly changes accordingly. In fact, we use the projectors

$$P_{\mathbf{n},\mathbf{m}}^s = [T_n[\mathbf{p}_z](K_{n,m} \otimes I_s)]_- \otimes [T_n[\mathbf{p}_z](K_{n,m} \otimes I_s)]_-,$$

where $\mathbf{p}_z$ is the univariate matrix-valued trigonometric polynomial defined in (V.34).

Extending the considerations that we made for the univariate case, we numerically look for the best choices of z to obtain the optimality of the V-cycle method.

In Tables V.17-V.18 we report for $z = 1, \dots, 5$ the number of iterations needed for achieving the tolerance $\varepsilon = 10^{-7}$ when increasing the matrix size and using $\mathbf{p}_z$ in the construction of the projector. Table V.17 shows the results for the $\mathbb{Q}_2$ stiffness matrix and Table V.18 for the $\mathbb{Q}_3$ stiffness matrix. In both cases, we used as pre-smoother and post-smoother one iteration of Gauss-Seidel with $\omega_{\text{pre,post}} = 1$. Moreover, we can see that the choice $z = 1$ does not yield optimality. For the other choices of z, conversely, the number of iterations needed for the V-cycle convergence remains almost constant, when increasing the size N. We numerically see that the best choice of z is around 3 for both $s = 2$ and $s = 3$.

As a concluding note, we stress the fact that a crucial role for the optimality of a multigrid method is also played by the choice of the smoothing strategy and in particular of the relaxation parameters $\omega_{\text{pre,post}}$. One could set different parameters for pre- and post-smoother, and study numerically the best values in order to make the constant which bounds from above the number of iterations smaller. Moreover, it is also possible to compute the optimal smoother methods and the associated relaxation parameters ω_ℓ at each level ℓ in order to reduce the number of iterations for the convergence of the V-cycle procedure. All these aspects will be object of future research.

Table V.17: Number of iterations for the V-cycle method using the $\mathbf{p}_z$ projection strategy for the $\mathbb{Q}_2$ stiffness matrix for the two-dimensional problem, pre- and post-smoother 1 iteration of Gauss-Seidel with $\omega_{\text{pre,post}} = 1$, tolerance $\varepsilon = 10^{-7}$.

t	$n = 2^t - 1$	$(2n-1)^2$	$z = 1$	$z = 2$	$z = 3$	$z = 4$	$z = 5$
7	127	64009	2724	63	26	25	25
8	255	259081	4000+	73	27	23	22
9	511	1042441	4000+	80	27	23	24
10	1023	4182025	4000+	84	27	24	25

Table V.18: Number of iterations for the V-cycle method using the $\mathbf{p}_z$ projection strategy for the $\mathbb{Q}_3$ stiffness matrix for the two-dimensional problem, pre- and post-smoother 1 iteration of Gauss-Seidel with $\omega_{\text{pre,post}} = 1$, tolerance $\varepsilon = 10^{-7}$.

t	$n = 2^t - 1$	$(2n-1)^2$	$z = 1$	$z = 2$	$z = 3$	$z = 4$	$z = 5$
6	63	35344	2719	69	57	59	60
7	127	144400	4000+	83	71	73	74
8	255	583696	4000+	90	60	60	60
9	511	2347024	4000+	94	59	60	61

Chapter VI

Fast Parallel Solver for the Space-Time IgA-DG Discretization of the Anisotropic Diffusion Equation

In the present chapter, we focus on the numerical solution of the time-dependent linear anisotropic diffusion equation

$$\begin{cases} \partial_t u(t,\mathbf{x}) - \nabla \cdot \mathcal{K}(\mathbf{x}) \nabla u(t,\mathbf{x}) = \psi(t,\mathbf{x}), & (t,\mathbf{x}) \in (0,\mathrm{T}) \times (0,1)^k, \\ u(t,\mathbf{x}) = 0, & (t,\mathbf{x}) \in (0,\mathrm{T}) \times \partial((0,1)^k), \\ u(t,\mathbf{x}) = 0, & (t,\mathbf{x}) \in \{0\} \times (0,1)^k, \end{cases} \tag{VI.1}$$

where $\mathcal{K}(\mathbf{x}) \in \mathbb{R}^{k \times k}$ is the matrix of diffusion coefficients and $\psi(t,\mathbf{x})$ is a source term. We impose homogeneous Dirichlet initial and boundary conditions both for simplicity and because the inhomogeneous case reduces to the homogeneous case by considering a lifting of the boundary data [105]. As far as the discretization techniques are concerned, we consider for (VI.1) the same space-time approximation as in [16], involving a $\boldsymbol{p}$-degree $C^{\boldsymbol{r}}$ finite element (FE) discretization in space and a q-degree discontinuous Galerkin (DG) discretization in time. Here, $\boldsymbol{p} = (p_1, \dots, p_k)$ and $\boldsymbol{r} = (r_1, \dots, r_k)$ are multi-indices such that $\mathbf{0} \leq \boldsymbol{r} \leq \boldsymbol{p} - \mathbf{1}$ and the parameters p_i and r_i represent, respectively, the polynomial degree and the smoothness of the FE basis functions in direction x_i.

We highlight that space-time approximations of dynamic problems, in contrast to standard time-stepping techniques, enable full space-time parallelism on modern massively parallel architectures [59]. Moreover, they can naturally deal with moving domains [86, 126, 128, 129, 137] and allow for space-time adaptivity [1, 48, 60, 87, 95, 98, 131]. The main idea of space-time formulations is to consider the temporal dimension as an additional spatial one and assemble a large space-time system to be solved in parallel as in [50]. Space-time methods have been used in combination with various numerical techniques, including finite differences [2, 17, 81], finite elements [9, 85, 88], isogeometric analysis [76, 89], and discontinuous Galerkin methods [1, 68, 85, 86, 96, 126, 137, 139]. Moreover, they have been considered for a variety of applications, such as mechanics [18], fluid dynamics [17, 86, 125], fluid-structure interaction [130], and many others. When dealing with space-time finite elements, the time direction needs special care. To ensure that the information flows in the positive time direction, a particular choice of

the basis in time is often used. The discontinuous Galerkin formulation with an "upwind" flow is a common choice in this context; see, for example, [86, 110, 126, 132].

In many cases, the overall discretization process leads to solving a large space-time linear system, for the solution of which a selection of specialized parallel solvers has been recently developed. We mention in particular the space-time parallel multigrid proposed by Gander and Neumüller [61] and the parallel preconditioners for space-time isogeometric analysis proposed by Hofer et al. [76].

In this chapter, we propose a fast solver for the system resulting from the discretization of (VI.1) through the space-time method mentioned above in the case of maximal smoothness $\boldsymbol{r} = \boldsymbol{p} - \mathbf{1}$, which is a standard approach [5, 14, 33, 84]. The solver is a preconditioned GMRES (PGMRES) method [108] whose preconditioner $\tilde{P}$ is obtained as an approximation of another preconditioner P inspired by the spectral analysis carried out in [16]. Informally speaking, the preconditioner $\tilde{P}$ is a standard multigrid, which is applied only in space and not in time, and which involves, at all levels, a single Gauss-Seidel post-smoothing step and standard bisection for the interpolation and restriction operators (following the Galerkin assembly). The proposed solver is then a multigrid preconditioned GMRES (MG-GMRES).

The solver's performance is illustrated through numerical experiments and turns out to be highly satisfactory in terms of iteration count and computational times. In addition, the solver is suited for parallel computation as it shows remarkable scaling properties with respect to the number of cores. Comparisons with other benchmark solvers are also presented and reveal the actual competitiveness of our proposal.

The current chapter reports the results in [15] and is organized as follows. In Section VI.1, we briefly recall the space-time FE-DG discretization of (VI.1) and we report the main result of [16] concerning the spectral distribution of the associated discretization matrix C. In Section VI.2, we present a fast PGMRES method for the matrix C, which is the root from which the proposed solver originated. In Section VI.4, we describe the proposed solver, in Section VI.5 we describe its parallel version, and in Section VI.6 we illustrate its performance in terms of iteration count, run-times and scaling.

VI.1 Space-Time FE-DG Discretization of Anisotropic Diffusion

Let $N \in \mathbb{N}$ and $\boldsymbol{n} = (n_1, \dots, n_k) \in \mathbb{N}^k$, and define the following uniform partitions in time and space:

$$t_i = i\Delta t, \qquad i = 0, \dots, N, \qquad \Delta t = \mathrm{T}/N,$$
$$\mathbf{x}_{\boldsymbol{i}} = \boldsymbol{i}\Delta\mathbf{x} = (i_1\Delta x_1, \dots, i_k\Delta x_k), \qquad \boldsymbol{i} = \mathbf{0}, \dots, \boldsymbol{n}, \qquad \Delta\mathbf{x} = (\Delta x_1, \dots, \Delta x_k) = (1/n_1, \dots, 1/n_k).$$

We consider for the differential problem (VI.1) the same space-time discretization as in [16], that is, we use a $\boldsymbol{p}$-degree $C^{\boldsymbol{r}}$ FE approximation in space based on the uniform mesh $\{\mathbf{x}_{\boldsymbol{i}},\ \boldsymbol{i} = \mathbf{0}, \dots, \boldsymbol{n}\}$ and a q-degree DG approximation in time based on the uniform mesh $\{t_i,\ i = 0, \dots, N\}$. Here, $\boldsymbol{p} = (p_1, \dots, p_k)$ and $\boldsymbol{r} = (r_1, \dots, r_k)$ are multi-indices, with p_i and $0 \leq r_i \leq p_i - 1$ representing, respectively, the polynomial degree and the smoothness of the FE basis functions in direction x_i. As carefully explained in [16, Section 3], the overall discretization

process leads to the following, normally large, linear system:

$$C_{N,\boldsymbol{n}}^{[q,\boldsymbol{p},\boldsymbol{r}]}(\mathcal{K})u = b, \tag{VI.2}$$

where:

- $C_{N,\boldsymbol{n}}^{[q,\boldsymbol{p},\boldsymbol{r}]}(\mathcal{K})$ is the $N \times N$ block matrix given by

$$C_{N,\boldsymbol{n}}^{[q,\boldsymbol{p},\boldsymbol{r}]}(\mathcal{K}) = \begin{bmatrix} A_{\boldsymbol{n}}^{[q,\boldsymbol{p},\boldsymbol{r}]}(\mathcal{K}) & & & \\ B_{\boldsymbol{n}}^{[q,\boldsymbol{p},\boldsymbol{r}]} & A_{\boldsymbol{n}}^{[q,\boldsymbol{p},\boldsymbol{r}]}(\mathcal{K}) & & \\ & \ddots & \ddots & \\ & & B_{\boldsymbol{n}}^{[q,\boldsymbol{p},\boldsymbol{r}]} & A_{\boldsymbol{n}}^{[q,\boldsymbol{p},\boldsymbol{r}]}(\mathcal{K}) \end{bmatrix};$$

- the blocks $A_{\boldsymbol{n}}^{[q,\boldsymbol{p},\boldsymbol{r}]}(\mathcal{K})$ and $B_{\boldsymbol{n}}^{[q,\boldsymbol{p},\boldsymbol{r}]}$ are $(q+1)\bar{n} \times (q+1)\bar{n}$ matrices given by

$$A_{\boldsymbol{n}}^{[q,\boldsymbol{p},\boldsymbol{r}]}(\mathcal{K}) = K_{[q]} \otimes M_{\boldsymbol{n},[\boldsymbol{p},\boldsymbol{r}]} + \frac{\Delta t}{2} M_{[q]} \otimes K_{\boldsymbol{n},[\boldsymbol{p},\boldsymbol{r}]}(\mathcal{K}), \tag{VI.3}$$

$$B_{\boldsymbol{n}}^{[q,\boldsymbol{p},\boldsymbol{r}]} = -J_{[q]} \otimes M_{\boldsymbol{n},[\boldsymbol{p},\boldsymbol{r}]}, \tag{VI.4}$$

  where $\bar{n} = \prod_{i=1}^{k}(n_i(p_i - r_i) + r_i - 1)$ is the number of degrees of freedom (DoFs) in space (the total number of DoFs is equal to the size $N(q+1)\bar{n}$ of the matrix $C_{N,\boldsymbol{n}}^{[q,\boldsymbol{p},\boldsymbol{r}]}(\mathcal{K})$);

- $M_{\boldsymbol{n},[\boldsymbol{p},\boldsymbol{r}]}$ and $K_{\boldsymbol{n},[\boldsymbol{p},\boldsymbol{r}]}(\mathcal{K})$ are the $\bar{n} \times \bar{n}$ mass and stiffness matrices in space, which are given by

$$M_{\boldsymbol{n},[\boldsymbol{p},\boldsymbol{r}]} = \left[\int_{[0,1]^k} B_{\boldsymbol{j}+\boldsymbol{1},[\boldsymbol{p},\boldsymbol{r}]}(\mathbf{x}) B_{\boldsymbol{i}+\boldsymbol{1},[\boldsymbol{p},\boldsymbol{r}]}(\mathbf{x}) \mathrm{d}\mathbf{x}\right]_{\boldsymbol{i},\boldsymbol{j}=\boldsymbol{1}}^{\boldsymbol{n}(\boldsymbol{p}-\boldsymbol{r})+\boldsymbol{r}-\boldsymbol{1}}, \tag{VI.5}$$

$$K_{\boldsymbol{n},[\boldsymbol{p},\boldsymbol{r}]}(\mathcal{K}) = \left[\int_{[0,1]^k} \left[\mathcal{K}(\mathbf{x}) \nabla B_{\boldsymbol{j}+\boldsymbol{1},[\boldsymbol{p},\boldsymbol{r}]}(\mathbf{x})\right] \cdot \nabla B_{\boldsymbol{i}+\boldsymbol{1},[\boldsymbol{p},\boldsymbol{r}]}(\mathbf{x}) \mathrm{d}\mathbf{x}\right]_{\boldsymbol{i},\boldsymbol{j}=\boldsymbol{1}}^{\boldsymbol{n}(\boldsymbol{p}-\boldsymbol{r})+\boldsymbol{r}-\boldsymbol{1}}, \tag{VI.6}$$

  where $B_{\boldsymbol{1},[\boldsymbol{p},\boldsymbol{r}]}, \ldots, B_{\boldsymbol{n}(\boldsymbol{p}-\boldsymbol{r})+\boldsymbol{r}+\boldsymbol{1},[\boldsymbol{p},\boldsymbol{r}]}$ are the tensor-product B-splines defined by

$$B_{\boldsymbol{i},[\boldsymbol{p},\boldsymbol{r}]}(\mathbf{x}) = \prod_{j=1}^{k} B_{i_j,[p_j,r_j]}(x_j), \quad \boldsymbol{i} = \boldsymbol{1}, \ldots, \boldsymbol{n}(\boldsymbol{p}-\boldsymbol{r}) + \boldsymbol{r} + \boldsymbol{1},$$

  with $B_{1,[p_j,k_j]}, \ldots, B_{n_j(p_j-r_j)+r_j+1,[p_j,r_j]}$ being the B-splines of degree p_j and smoothness C^{r_j} defined on the knot sequence

$$\left\{\underbrace{0,\ldots,0}_{p_j+1}, \underbrace{\frac{1}{n_j},\ldots,\frac{1}{n_j}}_{p_j-r_j}, \underbrace{\frac{2}{n_j},\ldots,\frac{2}{n_j}}_{p_j-r_j}, \ldots, \underbrace{\frac{n_j-1}{n_j},\ldots,\frac{n_j-1}{n_j}}_{p_j-r_j}, \underbrace{1,\ldots,1}_{p_j+1}\right\}.$$

- $M_{[q]}$, $K_{[q]}$, $J_{[q]}$ are the $(q+1) \times (q+1)$ blocks given by

$$M_{[q]} = \left[\int_{-1}^{1} \ell_{j,[q]}(\tau) \ell_{i,[q]}(\tau) \mathrm{d}\tau\right]_{i,j=1}^{q+1}, \tag{VI.7}$$

$$K_{[q]} = \left[\ell_{j,[q]}(1)\ell_{i,[q]}(1) - \int_{-1}^{1} \ell_{j,[q]}(\tau) \ell'_{i,[q]}(\tau) \mathrm{d}\tau\right]_{i,j=1}^{q+1}, \tag{VI.8}$$

$$J_{[q]} = \left[\ell_{j,[q]}(1)\ell_{i,[q]}(-1)\right]_{i,j=1}^{q+1}, \tag{VI.9}$$

where $\{\ell_{1,[q]},\dots,\ell_{q+1,[q]}\}$ is a fixed basis for the space $\mathbb{P}_q$ of polynomials of degree $\le q$. In the context of (nodal) DG methods [72], $\ell_{1,[q]},\dots,\ell_{q+1,[q]}$ are often chosen as the Lagrange polynomials associated with $q+1$ fixed points $\{\tau_1,\dots,\tau_{q+1}\}\subseteq[-1,1]$, such as, for example, the Gauss–Lobatto or the right Gauss–Radau nodes in $[-1,1]$.

The solution of system (VI.2) yields the approximate solution of problem (VI.1); see [16] for details. The main result of [16] is reported in the following theorem.

Theorem VI.1.1. *Let $q\ge 0$ be an integer, let $\boldsymbol{p}\in\mathbb{N}^k$ and $\mathbf{0}\le\boldsymbol{r}\le\boldsymbol{p}-\mathbf{1}$. Suppose that $\mathcal{K}:(0,1)^k\to\mathbb{R}^{k\times k}$ is a symmetric matrix-valued function in $L^\infty((0,1)^k,k)$ and that the following two conditions are met:*

- *$\boldsymbol{n}=\boldsymbol{\alpha}n$, where $\boldsymbol{\alpha}=(\alpha_1,\dots,\alpha_k)$ is a vector with positive components in $\mathbb{Q}^k$ and n varies in some infinite subset of $\mathbb{N}$ such that $\boldsymbol{n}=\boldsymbol{\alpha}n\in\mathbb{N}^k$;*
- *$N=N(n)$ is such that $N\to\infty$ and $N/n^2\to 0$ as $n\to\infty$.*

Then, for the sequence of normalized space-time matrices $\{2Nn^{k-2}C_{N,\boldsymbol{n}}^{[q,\boldsymbol{p},\boldsymbol{r}]}(\mathcal{K})\}_n$ we have the spectral distribution relation

$$\{2Nn^{k-2}C_{N,\boldsymbol{n}}^{[q,\boldsymbol{p},\boldsymbol{r}]}(\mathcal{K})\}_n\sim_\lambda \mathbf{f}_{[q,\boldsymbol{p},\boldsymbol{r}]}^{[\boldsymbol{\alpha},\mathcal{K}]},$$

where:

- *the spectral symbol $\mathbf{f}_{[q,\boldsymbol{p},\boldsymbol{r}]}^{[\boldsymbol{\alpha},\mathcal{K}]}:[0,1]^k\times[-\pi,\pi]^k\to\mathbb{C}^{(q+1)\prod_{i=1}^k(p_i-r_i)\times(q+1)\prod_{i=1}^k(p_i-r_i)}$ is defined as*

$$\mathbf{f}_{[q,\boldsymbol{p},\boldsymbol{r}]}^{[\boldsymbol{\alpha},\mathcal{K}]}(\mathbf{x},\boldsymbol{\vartheta})=\mathbf{f}_{[\boldsymbol{p},\boldsymbol{r}]}^{[\boldsymbol{\alpha},\mathcal{K}]}(\mathbf{x},\boldsymbol{\vartheta})\otimes \mathrm{T}M_{[q]};\tag{VI.10}$$

- *$\mathbf{f}_{[\boldsymbol{p},\boldsymbol{r}]}^{[\boldsymbol{\alpha},\mathcal{K}]}:[0,1]^k\times[-\pi,\pi]^k\to\mathbb{C}^{\prod_{i=1}^k(p_i-r_i)\times\prod_{i=1}^k(p_i-r_i)}$ is defined as*

$$\mathbf{f}_{[\boldsymbol{p},\boldsymbol{r}]}^{[\boldsymbol{\alpha},\mathcal{K}]}(\mathbf{x},\boldsymbol{\vartheta})=\frac{1}{\prod_{i=1}^k\alpha_i}\sum_{i,j=1}^k\alpha_i\alpha_j\mathcal{K}_{ij}(\mathbf{x})(H_{[\boldsymbol{p},\boldsymbol{r}]})_{ij}(\boldsymbol{\vartheta});\tag{VI.11}$$

- *$H_{[\boldsymbol{p},\boldsymbol{r}]}$ is a $k\times k$ block matrix whose (i,j) entry is a $\prod_{i=1}^k(p_i-r_i)\times\prod_{i=1}^k(p_i-r_i)$ block defined as in* [16, *Eq.* (5.12)];
- *T is the final time in* (VI.1) *and $M_{[q]}$ is given in* (VI.7).

With the same argument used in [16] to prove Theorem VI.1.1, it is not difficult to prove the following result.

Theorem VI.1.2. *Suppose the hypotheses of Theorem* VI.1.1 *are valid, and let*

$$X_{N,\boldsymbol{n}}^{[q,\boldsymbol{p},\boldsymbol{r}]}(\mathcal{K})=2Nn^{k-2}\Big(I_N\otimes\frac{\Delta t}{2}M_{[q]}\otimes K_{\boldsymbol{n},[\boldsymbol{p},\boldsymbol{r}]}(\mathcal{K})\Big)=\mathrm{T}n^{k-2}I_N\otimes M_{[q]}\otimes K_{\boldsymbol{n},[\boldsymbol{p},\boldsymbol{r}]}(\mathcal{K}).$$

Then,

$$\{2Nn^{k-2}(I_N\otimes A_{\boldsymbol{n}}^{[q,\boldsymbol{p},\boldsymbol{r}]}(\mathcal{K}))\}_n\sim_\lambda \mathbf{f}_{[q,\boldsymbol{p},\boldsymbol{r}]}^{[\boldsymbol{\alpha},\mathcal{K}]},$$
$$\{X_{N,\boldsymbol{n}}^{[q,\boldsymbol{p},\boldsymbol{r}]}(\mathcal{K})\}_n\sim_\lambda \mathbf{f}_{[q,\boldsymbol{p},\boldsymbol{r}]}^{[\boldsymbol{\alpha},\mathcal{K}]}.$$

VI.2 Fast PGMRES for the Space-Time FE-DG Matrix $C_{N,\boldsymbol{n}}^{[q,\boldsymbol{p},\boldsymbol{r}]}(\mathcal{K})$

Suppose the hypotheses of Theorem VI.1.1 are valid. Then, on the basis of Theorem VI.1.2 and the theory of block GLT sequences (see Section I.7 and references therein), we expect that the sequence of preconditioned matrices

$$(I_N \otimes A_{N,\boldsymbol{n}}^{[q,\boldsymbol{p},\boldsymbol{r}]}(\mathcal{K}))^{-1} C_{N,\boldsymbol{n}}^{[q,\boldsymbol{p},\boldsymbol{r}]}(\mathcal{K}), \tag{VI.12}$$

as well as the sequence of preconditioned matrices

$$(X_{N,\boldsymbol{n}}^{[q,\boldsymbol{p},\boldsymbol{r}]}(\mathcal{K}))^{-1}(2Nn^{k-2}C_{N,\boldsymbol{n}}^{[q,\boldsymbol{p},\boldsymbol{r}]}(\mathcal{K})) = \frac{2}{\Delta t}\,(I_N \otimes M_{[q]} \otimes K_{\boldsymbol{n},[\boldsymbol{p},\boldsymbol{r}]}(\mathcal{K}))^{-1} C_{N,\boldsymbol{n}}^{[q,\boldsymbol{p},\boldsymbol{r}]}(\mathcal{K}), \tag{VI.13}$$

has an asymptotic spectral distribution described by the preconditioned symbol

$$\big(\mathbf{f}_{[q,\boldsymbol{p},\boldsymbol{r}]}^{[\boldsymbol{\alpha},\mathcal{K}]}\big)^{-1}\mathbf{f}_{[q,\boldsymbol{p},\boldsymbol{r}]}^{[\boldsymbol{\alpha},\mathcal{K}]} = I_{(q+1)\prod_{i=1}^{k}(p_i-r_i)}.$$

This means that the eigenvalues of the two sequences of matrices (VI.12) and (VI.13) are (weakly) clustered at 1; see [13, Section 2.4.2]. Therefore, in view of the convergence properties of the GMRES method [108], we may expect that the PGMRES with preconditioner

$$I_N \otimes A_{N,\boldsymbol{n}}^{[q,\boldsymbol{p},\boldsymbol{r}]}(\mathcal{K}) \tag{VI.14}$$

or

$$P_{N,\boldsymbol{n}}^{[q,\boldsymbol{p},\boldsymbol{r}]}(\mathcal{K}) = \frac{\Delta t}{2}\, I_N \otimes M_{[q]} \otimes K_{\boldsymbol{n},[\boldsymbol{p},\boldsymbol{r}]}(\mathcal{K}) \tag{VI.15}$$

for solving a linear system with coefficient matrix $C_{N,\boldsymbol{n}}^{[q,\boldsymbol{p},\boldsymbol{r}]}(\mathcal{K})$ has an optimal convergence rate, i.e., the number of iterations for reaching a preassigned accuracy ε is independent of (or only weakly dependent on) the matrix size.

To show that this expectation is realized, we solve the system (VI.2) in two space dimensions ($k = 2$), up to a precision $\varepsilon = 10^{-8}$, by means of the GMRES and the PGMRES with preconditioner $P_{N,\boldsymbol{n}}^{[q,\boldsymbol{p},\boldsymbol{r}]}(\mathcal{K})$, using $\psi(t,\mathbf{x}) = 1$, $\mathrm{T} = 1$, $\boldsymbol{\alpha} = (1,1)$, $\boldsymbol{n} = \boldsymbol{\alpha}n = (n,n)$, $\boldsymbol{p} = (p,p)$, $\boldsymbol{r} = (r,r)$, and varying $\mathcal{K}(\mathbf{x})$, N, n, q, p, r. The resulting number of iterations are collected in Tables VI.1–VI.3. We see from the tables that the unpreconditioned GMRES solver rapidly deteriorates with increasing n, and it is not robust with respect to p, r. On the other hand, the convergence rate of the proposed PGMRES is robust with respect to all the spatial parameters n, p, r. However, as it is known, each PGMRES iteration requires solving a linear system with coefficient matrix given by the preconditioner $P_{N,\boldsymbol{n}}^{[q,\boldsymbol{p},\boldsymbol{r}]}(\mathcal{K})$, and this is not required in a GMRES iteration. Thus, in order to prove that the proposed PGMRES is fast, in Section VI.4 we show that we are able to solve efficiently a linear system associated with $P_{N,\boldsymbol{n}}^{[q,\boldsymbol{p},\boldsymbol{r}]}(\mathcal{K})$.

VI.3 Fast Tensor Solver for the PGMRES Preconditioner $P_{N,\boldsymbol{n}}^{[q,\boldsymbol{p},\boldsymbol{r}]}(\mathcal{K})$

The main observation of this section is that, thanks to the tensor structure of $P_{N,\boldsymbol{n}}^{[q,\boldsymbol{p},\boldsymbol{r}]}(\mathcal{K})$ (see (VI.15)), the solution of a linear system with coefficient matrix $P_{N,\boldsymbol{n}}^{[q,\boldsymbol{p},\boldsymbol{r}]}(\mathcal{K})$ reduces to the solution

Table VI.1: Number of iterations GM$[p]$ and PGM$[p]$ needed by, respectively, the GMRES and the PGMRES with preconditioner $P_{N,\boldsymbol{n}}^{[q,\boldsymbol{p},\boldsymbol{r}]}(\mathcal{K})$, for solving the linear system (VI.2), up to a precision $\varepsilon = 10^{-8}$, in the case where $k = 2$, $\mathcal{K}(\mathbf{x}) = I_2$, $\psi(t, \mathbf{x}) = 1$, T $= 1$, $q = 0$, $\boldsymbol{n} = (n, n)$, $\boldsymbol{p} = (p, p)$, $\boldsymbol{r} = (p-1, p-1)$, $N = n$. The total size of the space-time system (number of DoFs) is given by $n\bar{n} = n(n+p-2)^2$.

$n = N$	GM[3]	PGM[3]	GM[4]	PGM[4]	GM[5]	PGM[5]	GM[6]	PGM[6]	GM[7]	PGM[7]
20	66	21	85	21	170	21	269	21	532	21
40	168	40	178	40	235	40	380	40	572	40
60	295	59	314	59	360	59	477	59	611	59
80	443	77	473	77	506	77	621	77	720	77
100	609	94	652	94	699	94	780	94	879	94
120	790	111	847	111	909	111	971	111	1025	111

Table VI.2: Number of iterations GM$[p, r]$ and PGM$[p, r]$ needed by, respectively, the GMRES and the PGMRES with preconditioner $P_{N,\boldsymbol{n}}^{[q,\boldsymbol{p},\boldsymbol{r}]}(\mathcal{K})$, for solving the linear system (VI.2), up to a precision $\varepsilon = 10^{-8}$, in the case where $k = 2$, $\mathcal{K}(x_1, x_2) = \begin{bmatrix} \cos(x_1) + x_2 & 0 \\ 0 & x_1 + \sin(x_2) \end{bmatrix}$, $\psi(t, \mathbf{x}) = 1$, T $= 1$, $q = 1$, $\boldsymbol{n} = (n, n)$, $\boldsymbol{p} = (p, p)$, $\boldsymbol{r} = (r, r)$, $N = 20$. The number of DoFs is given by $40\bar{n} = 40(n(p-r)+r-1)^2$.

n	GM[1, 0]	PGM[1, 0]	GM[2, 0]	PGM[2, 0]	GM[2, 1]	PGM[2, 1]	GM[3, 1]	PGM[3, 1]
20	244	42	383	42	156	42	276	42
40	502	42	778	42	314	42	560	42
60	763	42	1174	42	474	42	842	42
80	1026	42	1570	42	635	42	1146	42
100	1275	42	1966	42	796	42	1894	42
120	1608	42	2374	42	954	42	1898	42
n	GM[4, 1]	PGM[4, 1]	GM[4, 2]	PGM[4, 2]	GM[5, 2]	PGM[5, 2]	GM[5, 3]	PGM[5, 3]
20	444	42	390	42	522	42	514	42
40	759	42	565	42	721	42	643	42
60	1148	42	771	42	953	42	831	42
80	1536	42	1035	42	1337	42	1026	42
100	1909	42	1299	42	2232	42	1226	42
120	2329	42	1564	42	2390	42	1831	42

Table VI.3: Number of iterations GM$[p,r]$ and PGM$[p,r]$ needed by, respectively, the GMRES and the PGMRES with preconditioner $P_{N,\boldsymbol{n}}^{[q,\boldsymbol{p},\boldsymbol{r}]}(\mathcal{K})$, for solving the linear system (VI.2), up to a precision $\varepsilon = 10^{-8}$, in the case where $k=2$, $\mathcal{K}(x_1,x_2) = \begin{bmatrix} (2+\cos x_1)(1+x_2) & \cos(x_1+x_2)\sin(x_1+x_2) \\ \cos(x_1+x_2)\sin(x_1+x_2) & (2+\sin x_2)(1+x_1) \end{bmatrix}$, $\psi(t,\mathbf{x})=1$, T $=1$, $q=2$, $\boldsymbol{n}=(n,n)$, $\boldsymbol{p}=(p,p)$, $\boldsymbol{r}=(r,r)$, $N=20$. The number of DoFs is given by $60\bar{n} = 60(n(p-r)+r-1)^2$.

n	GM$[2,0]$	PGM$[2,0]$	GM$[2,1]$	PGM$[2,1]$	GM$[3,0]$	PGM$[3,0]$	GM$[3,2]$	PGM$[3,2]$
20	286	40	112	40	400	40	123	40
40	579	40	228	40	809	40	224	40
60	874	40	345	40	1218	40	339	40
80	1170	40	463	40	1716	40	456	40
100	1466	40	580	40	2204	40	573	40
120	1757	40	697	40	2487	40	690	40
n	GM$[4,0]$	PGM$[4,0]$	GM$[4,3]$	PGM$[4,3]$	GM$[5,0]$	PGM$[5,0]$	GM$[5,4]$	PGM$[5,4]$
20	779	40	208	40	1460	40	396	40
40	1070	40	270	40	1982	40	419	40
60	1580	40	361	40	2376	40	466	40
80	2176	40	487	40	2733	40	531	40
100	2668	40	613	40	3559	40	657	40
120	3284	40	738	40	4565	40	791	40

of three linear systems with coefficient matrices I_N, $M_{[q]}$, $K_{\boldsymbol{n},[\boldsymbol{p},\boldsymbol{r}]}(\mathcal{K})$. Indeed, using the canonical algorithm for tensor-product matrices to solve the system $P_{N,\boldsymbol{n}}^{[q,\boldsymbol{p},\boldsymbol{r}]}(\mathcal{K})\mathbf{x}=\mathbf{y}$, we obtain

$$
\begin{aligned}
\mathbf{x} &= (P_{N,\boldsymbol{n}}^{[q,\boldsymbol{p},\boldsymbol{r}]}(\mathcal{K}))^{-1}\mathbf{y} \\
&= \Big(\frac{2}{\Delta t} I_N \otimes M_{[q]}^{-1} \otimes K_{\boldsymbol{n},[\boldsymbol{p},\boldsymbol{r}]}(\mathcal{K})^{-1}\Big)\mathbf{y} \\
&= (\widetilde{M}_{N,[q]} \otimes K_{\boldsymbol{n},[\boldsymbol{p},\boldsymbol{r}]}(\mathcal{K})^{-1})\mathbf{y} \\
&= (\widetilde{M}_{N,[q]} \otimes I_{\bar{n}})(I_{N(q+1)} \otimes K_{\boldsymbol{n},[\boldsymbol{p},\boldsymbol{r}]}(\mathcal{K})^{-1})\mathbf{y} \\
&= (\widetilde{M}_{N,[q]} \otimes I_{\bar{n}}) \begin{bmatrix} K_{\boldsymbol{n},[\boldsymbol{p},\boldsymbol{r}]}(\mathcal{K})^{-1} & & & \\ & K_{\boldsymbol{n},[\boldsymbol{p},\boldsymbol{r}]}(\mathcal{K})^{-1} & & \\ & & \ddots & \\ & & & K_{\boldsymbol{n},[\boldsymbol{p},\boldsymbol{r}]}(\mathcal{K})^{-1} \end{bmatrix} \mathbf{y} \qquad \text{(VI.16)} \\
&= \mathrm{vec}(K_{\boldsymbol{n},[\boldsymbol{p},\boldsymbol{r}]}(\mathcal{K})^{-1} Y \widetilde{M}_{N,[q]}), \qquad \text{(VI.17)}
\end{aligned}
$$

where:

- $\widetilde{M}_{N,[q]} = \frac{2}{\Delta t} I_N \otimes M_{[q]}^{-1}$ can be pre-computed with a negligible cost, because $M_{[q]}$ is a small $(q+1)\times(q+1)$ matrix (if Gauss–Radau nodes are used, $M_{[q]}$ is also diagonal and hence $\widetilde{M}_{N,[q]}$ is diagonal as well);
- $\mathrm{vec}(X)$ is the column-wise form of X, that is the vector obtained by stacking the columns of X;
- Y is the $\bar{n}\times N(q+1)$ matrix such that $\mathrm{vec}(Y)=\mathbf{y}$.

It is then clear that the computation of the solution $\mathbf{x}$ reduces to solving the $N(q+1)$ linear systems $K_{\boldsymbol{n},[\boldsymbol{p},\boldsymbol{r}]}(\mathcal{K})\mathbf{x}_i = \mathbf{y}_i$, $i = 1, \ldots, N(q+1)$, where $\mathbf{y}_i$ is the ith column of Y, and multiply the resulting matrix $K_{\boldsymbol{n},[\boldsymbol{p},\boldsymbol{r}]}(\mathcal{K})^{-1}Y$ by $\widetilde{M}_{N,[q]}$. Note that the various $\mathbf{x}_i$ can be computed in parallel as the computation of $\mathbf{x}_i$ is independent of the computation of $\mathbf{x}_j$ whenever $i \neq j$. Depending on the implementation and the parallel setting, it can be advantageous to express $\mathbf{x}$ using vec$(\cdot)$ as in (VI.17) or tensor products as in (VI.16).

VI.4 Solver for the Space-Time IgA-DG Matrix $C_{N,\boldsymbol{n}}^{[q,\boldsymbol{p}]}(\mathcal{K})$

The solver suggested in Section VI.2 for a linear system with coefficient matrix $C_{N,\boldsymbol{n}}^{[q,\boldsymbol{p},\boldsymbol{p-1}]}(\mathcal{K}) = C_{N,\boldsymbol{n}}^{[q,\boldsymbol{p}]}(\mathcal{K})$ is a PGMRES with preconditioner $P_{N,\boldsymbol{n}}^{[q,\boldsymbol{p},\boldsymbol{p-1}]}(\mathcal{K}) = P_{N,\boldsymbol{n}}^{[q,\boldsymbol{p}]}(\mathcal{K})$; the solution of a linear system associated with $P_{N,\boldsymbol{n}}^{[q,\boldsymbol{p}]}(\mathcal{K})$, which is required at each PGMRES iteration, is performed via the tensor solver described in Section VI.3 coupled with a suitable multigrid method for the space stiffness matrix $K_{\boldsymbol{n},[\boldsymbol{p}]}(\mathcal{K})$.

Actually, it was discovered experimentally that the PGMRES method converges faster if the linear system with coefficient matrix $P_{N,\boldsymbol{n}}^{[q,\boldsymbol{p}]}(\mathcal{K})$ occurring at each PGMRES iteration is not solved exactly. More precisely, when applying to $K_{\boldsymbol{n},[\boldsymbol{p}]}(\mathcal{K})$ a multigrid method involving, at all levels, a single Gauss-Seidel post-smoothing step and standard bisection for the interpolation and restriction operators, it is enough to perform only a few multigrid iterations in order to achieve an excellent PGMRES run-time and, in fact, only one multigrid iteration is sufficient.

In view of these experimental discoveries, we propose to solve a linear system with coefficient matrix $C_{N,\boldsymbol{n}}^{[q,\boldsymbol{p}]}(\mathcal{K})$ in the following way:

- apply to the given system the PGMRES algorithm with preconditioner $P_{N,\boldsymbol{n}}^{[q,\boldsymbol{p}]}(\mathcal{K})$;
- apply to the linear system with coefficient matrix $P_{N,\boldsymbol{n}}^{[q,\boldsymbol{p}]}(\mathcal{K})$ occurring at each PGMRES iteration the tensor solver described in Section VI.3;
- the tensor solver would require solving $q(N+1)$ linear systems with coefficient matrix $K_{\boldsymbol{n},[\boldsymbol{p}]}(\mathcal{K})$ as per Eq. (VI.16) or (VI.17); instead of solving exactly these systems, apply to each of them μ multigrid iterations involving, at all levels, a single Gauss-Seidel post-smoothing step and standard bisection for the interpolation and restriction operators at all levels (following the Galerkin approach).

VI.5 Parallel Solver for the Space-Time IgA-DG Matrix $C_{N,\boldsymbol{n}}^{[q,\boldsymbol{p}]}(\mathcal{K})$

In Section VI.4, we have described the sequential version of the proposed solver to be used when only one processor is available. If $\rho > 1$ processors are available, we use a modification of the solver, which is suited for parallel computation. It consists in solving $Cu = b$ (with $C = C_{N,\boldsymbol{n}}^{[q,\boldsymbol{p}]}(\mathcal{K})$) by a standard block Jacobi iterative method in which the block diagonal preconditioner D is formed by exactly ρ diagonal blocks $C_1, \ldots, C_\rho$, explicitly

$$D = \begin{bmatrix} C_1 & & \\ & \ddots & \\ & & C_\rho \end{bmatrix}.$$

It is implicitly assumed that $C_1, \ldots, C_\rho$ are invertible and that the sum of their sizes $m_1, \ldots, m_\rho$ equals the size $N(q+1)\bar{n}$ of C. The resulting block Jacobi method can be written as

$$u^{(j+1)} = u^{(j)} + D^{-1} r^{(j)},$$

where $r^{(j)} = b - Cu^{(j)}$ is the jth residual. Considering the block structure of D, it can also be written as

$$u_i^{(j+1)} = u_i^{(j)} + C_i^{-1} r_i^{(j)}, \qquad i = 1, \ldots, \rho, \tag{VI.18}$$

where for any vector y of length $N(q+1)\bar{n}$ we write $y^T = [y_1^T, \ldots, y_\rho^T]$ with y_i of length m_i. The ith processor takes care of solving the ith system (with matrix C_i) in (VI.18). It only remains to clarify how the blocks $C_1, \ldots, C_\rho$ are chosen and how the ρ systems in (VI.18) are solved. We distinguish two cases.

- $\rho \leq N$ (see Figure VI.1, left). In this case, each block C_i is chosen so that the block row of C corresponding to C_i consists of one or more time slabs (i.e., a positive integer number of time slabs). In this scenario, no time slab is shared between different processors. Moreover, each block C_i is just a smaller version of the matrix C and the ith processor solves the ith system in (VI.18) by simply applying the solver proposed in Section VI.4 to the matrix C_i.
- $\rho > N$ (see Figure VI.1, right). In this case, after partitioning C into N block rows (the N time slabs), we subdivide them into further block rows until exhaustion of the available processors, and we choose $C_1, \ldots, C_\rho$ as the diagonal blocks corresponding to the resulting row-wise partition. In this way, every processor owns at most one time slab. Moreover, each block C_i is a diagonal block of $A = A_{N,\boldsymbol{n}}^{[q,\boldsymbol{p}]}(\mathcal{K})$ that may coincide with A itself. The ith processor solves the ith system in (VI.18) according to the following procedure.
  - Apply to the ith system in (VI.18) the PGMRES with preconditioner given by the ith diagonal block of $P = P_{N,\boldsymbol{n}}^{[q,\boldsymbol{p}]}(\mathcal{K})$, that is, the diagonal block P_i occupying in P the same position as the diagonal block C_i in C. Note that P_i is a diagonal block of $I_{q+1} \otimes K$ (with $K = K_{\boldsymbol{n},[\boldsymbol{p}]}(\mathcal{K})$).
  - The exact solution of the linear system with matrix P_i occurring at each PGMRES iteration would require solving $\eta \leq q+1$ linear systems with a matrix given by a principal submatrix of K. For example, assuming $\bar{n}$ is even, if P_i is the leading principal submatrix of $I_{q+1} \otimes K$ of order $\bar{n} + \bar{n}/2$, then the solution of a linear system with matrix P_i requires solving 2 linear systems with matrices K and $K_{\bar{n}/2}$, respectively, where $K_{\bar{n}/2}$ is the leading principal submatrix of K of order $\bar{n}/2$.
  - Instead of solving exactly these η linear systems, apply to each of them, starting from the zero vector as initial guess, μ multigrid (V-cycle) iterations involving, at all levels, a single Gauss-Seidel post-smoothing step and standard bisection for the interpolation and restriction operators (following the Galerkin approach).

We remark that, when choosing the diagonal blocks $C_1, \ldots, C_\rho$, a load balancing principle is applied. This means that, as far as possible, the ρ systems in (VI.18) should have the same size. For example, as shown in Figure VI.1 (right), the first time slab is subdivided into two block rows of essentially the same thickness (exactly the same thickness if the size of A is even). Similarly, if we have $N = 2$ time slabs and $\rho = 6$ processors, then each time slab is subdivided into three block rows of essentially the same thickness; if we have $N = 30$ time slabs and $\rho = 4$

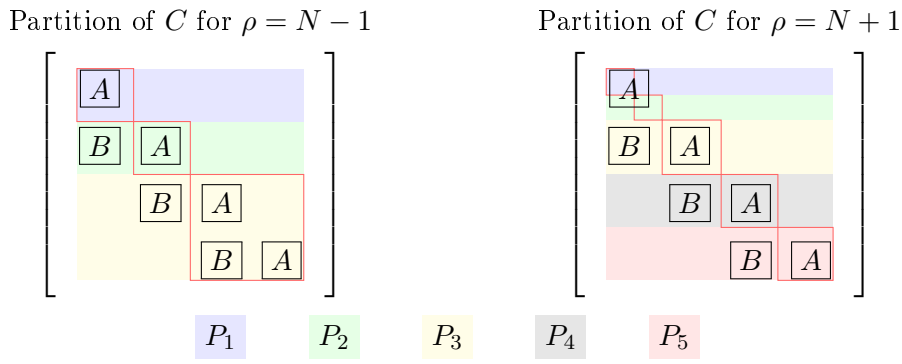

Figure VI.1: Row-wise partitions of the space-time matrix C using $\rho = N - 1$ processors (left) and $\rho = N + 1$ processors (right) with $N = 4$. For each partition, the corresponding diagonal blocks $C_1, \dots, C_\rho$ of the block diagonal preconditioner D are delimited by red squares. For simplicity, we write "A" instead of "$A_{\boldsymbol{n}}^{[q,\boldsymbol{p}]}(\mathcal{K})$" and "$B$" instead of $B_{\boldsymbol{n}}^{[q,\boldsymbol{p}]}$.

processors, then we assign 7 time slabs to one processor, 7 time slabs to another processor, 8 time slabs to another processor, and 8 time slabs to the last processor, and so on.

VI.6 Numerical Experiments: Iteration Count, Timing and Scaling

In this section, we illustrate through numerical experiments the performance of the proposed solver and we compare it to the performance of other benchmark solvers, such as the PGMRES with ILU preconditioner.

VI.6.1 Implementation Details

For the numerics of this section, we used the C++ framework PETSc [10, 11] and the domain specific language Utopia [141] for the parallel linear algebra and solvers, and the Cray-MPICH compiler. For the assembly of high order finite elements, we used the PetIGA package [34]. A parallel tensor-product routine was implemented to assemble space-time matrices. Numerical experiments have been performed on the Cray XC40 nodes of the Piz Daint supercomputer of the Swiss national supercomputing centre (CSCS).[1] The used partition features 1813 computation nodes, each of which holds two 18-core Intel Xeon E5-2695v4 (2.10GHz) processors. We stress that, when $\rho > 1$ processors are used, a block Jacobi iterative method as in (VI.18) is employed by default by PETSc before any other method in order to obtain scalable solution strategies. However, the PETSc default row-wise partition of the space-time matrix follows a load balancing principle and, except in the trivial case $\rho = N$, does not correspond to the row-wise partition described in Section VI.5; see Figure VI.2. Therefore, the partition must be adjusted by the user.

[1] https://www.cscs.ch/computers/piz-daint/

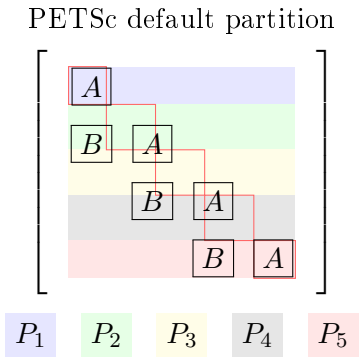

Figure VI.2: The PETSc default row-wise partition of the space-time matrix does not account for the structure of the space-time problem; compare with Figure VI.1.

VI.6.2 Experimental Setting

In the numerics of this section, we solve the linear system (VI.2) arising from the choices $k = 2$, $\psi(t, \mathbf{x}) = 1$, $\mathrm{T} = 1$, $\boldsymbol{n} = (n, n)$, $\boldsymbol{p} = (p, p)$, $\boldsymbol{r} = (p-1, p-1)$. The basis functions $\ell_{1,[q]}, \ldots, \ell_{q+1,[q]}$ are chosen as the Lagrange polynomials associated with the right Gauss-Radau nodes in $[-1, 1]$ (see, for instance, [72]). The values of $\mathcal{K}(\mathbf{x})$, N, n, q, p, are specified in each example. For each solver considered herein, we use $\varepsilon = 10^{-8}$ as a tolerance and the PETSc default stopping criterion. Whenever we report the run-time of a solver, the time spent in I/O operations and matrix assembly is ignored; run-times are always expressed in seconds. In all the tables below, the number of iterations needed by a given solver to converge within the tolerance $\varepsilon = 10^{-8}$ is reported in square brackets next to the corresponding run-time. Throughout this section, we use the following abbreviations for the solvers.

- $\boxed{\mathrm{ILU}(0) - \mathrm{PGMRES}}$

  PGMRES with preconditioner given by an ILU(0) factorization (ILU factorization with no fill-in) of the system matrix $C_{N,\boldsymbol{n}}^{[q,\boldsymbol{p}]}$.

- $\boxed{\mathrm{MG}_{\mu,\nu}^{L} - \mathrm{PGMRES}}$

  The proposed solver, as described in Section VI.4, with μ multigrid iterations applied to $K_{\boldsymbol{n},[\boldsymbol{p}]}(\mathcal{K})$. Each multigrid iterations involves ν Gauss-Seidel smoothing steps at the finest level (typically $\nu = 1$) and 1 Gauss-Seidel smoothing step at the coarse levels. The superscript L denotes the number of multigrid levels.

- $\boxed{\mathrm{TMG}_{\mu,\nu}^{L} - \mathrm{PGMRES}}$

  The same as "$\mathrm{MG}_{\mu,\nu}^{L}$-PGMRES", with the only difference that the multigrid iterations are performed with the telescopic option, thus giving rise to the telescopic multigrid (TMG) [45, 92]. This technique consists in halving the number of processors used across the grid hierarchy: if N_f processors are used on the fine grid ($l = 0$), then we use $N_f/2^l$ processors on level l. This strategy can be beneficial for the parallel multigrid performance, as shown in Section VI.6.4.

Table VI.4: PGMRES iterations and run-time (using 64 cores) to solve the linear system (VI.2) up to a precision of 10^{-8}, according to the experimental setting described in Section VI.6.2. We used $\mathcal{K}(\mathbf{x}) = I_2$, $q = 0$, $N = 32$ time steps and $n = 259 - p$. The total size of the space-time system (number of DoFs) is given by $32 \cdot 257^2$.

p	1	2	3	4	5	6	7	8
ILU(0)-GMRES	3.7 [579]	4.3 [367]	5.2 [269]	6.7 [226]	8.2 [193]	10.1 [174]	11.9 [156]	22.5 [234]
$MG^5_{3,2}$-GMRES	1.4 [33]	2.9 [33]	4.7 [33]	7.2 [33]	10.5 [35]	14.7 [36]	21.1 [41]	34.6 [53]
$MG^5_{1,2}$-GMRES	0.8 [33]	1.6 [33]	2.5 [33]	4.0 [35]	6.6 [42]	11.0 [52]	16.0 [60]	26.2 [77]
$MG^5_{3,1}$-GMRES	1.1 [33]	2.2 [33]	3.3 [33]	5.0 [34]	7.1 [36]	11.4 [43]	17.0 [51]	28.5 [67]
$MG^5_{1,1}$-GMRES	0.6 [33]	1.2 [33]	1.8 [34]	3.1 [39]	5.3 [50]	9.1 [63]	13.5 [75]	19.8 [87]

Table VI.5: PGMRES iterations and run-time (using 64 cores) to solve the linear system (VI.2) up to a precision of 10^{-8}, according to the experimental setting described in Section VI.6.2. We used $\mathcal{K}(x_1, x_2) = \begin{bmatrix} \cos(x_1) + x_2 & 0 \\ 0 & x_1 + \sin(x_2) \end{bmatrix}$, $q = 1$, $N = 20$ time steps and $n = 131 - p$. The total size of the space-time system (number of DoFs) is given by $40 \cdot 129^2$.

p	1	2	3	4	5	6	7	8
ILU(0)-GMRES	1.3 [449]	1.7 [283]	2.2 [219]	2.9 [183]	3.6 [158]	4.4 [141]	6.0 [148]	9.5 [186]
$MG^5_{2,3}$-GMRES	0.6 [55]	1.3 [55]	2.4 [55]	4.1 [58]	7.6 [64]	12.7 [90]	18.5 [101]	32.2 [139]
$MG^5_{1,3}$-GMRES	0.5 [57]	1.0 [56]	1.8 [56]	3.5 [68]	6.2 [85]	10.4 [103]	15.0 [116]	26.5 [161]
$MG^5_{2,1}$-GMRES	0.5 [57]	1.0 [57]	1.6 [58]	3.1 [77]	5.2 [91]	8.6 [112]	12.6 [128]	22.0 [179]
$MG^5_{1,1}$-GMRES	0.5 [67]	0.8 [65]	1.3 [68]	2.8 [90]	4.6 [110]	7.2 [125]	11.0 [150]	19.4 [210]

VI.6.3 Iteration Count and Timing

Tables VI.4–VI.6 illustrate the performance of the proposed solver in terms of number of iterations and run-times. It is clear from the table that the solver is superior to the classical PGMRES with preconditioner given by the ILU factorization of the system matrix $C^{[q,\boldsymbol{p}]}_{N,\boldsymbol{n}}(\mathcal{K})$. Moreover, the best performance of the solver is obtained when applying to $K_{\boldsymbol{n},[\boldsymbol{p}]}(\mathcal{K})$ a single multigrid iteration ($\mu = 1$) with one smoothing step at the finest level ($\nu = 1$). It should also be noted that the solver is considerably robust with respect to the spline degree p as both number of iterations and run-time do not grow significantly with p.

Table VI.6: PGMRES iterations and run-time (using 64 cores) to solve the linear system (VI.2) up to a precision of 10^{-8}, according to the experimental setting described in Section VI.6.2. We used $\mathcal{K}(x_1, x_2) = \begin{bmatrix} (2 + \cos x_1)(1 + x_2) & \cos(x_1 + x_2)\sin(x_1 + x_2) \\ \cos(x_1 + x_2)\sin(x_1 + x_2) & (2 + \sin x_2)(1 + x_1) \end{bmatrix}$, $q = 0$, $N = 20$ time steps and $n = 259 - p$. The total size of the space-time system (number of DoFs) is given by $20 \cdot 257^2$.

p	1	2	3	4	5	6	7	8
ILU(0)-GMRES	1.9 [450]	2.2 [284]	2.6 [205]	3.4 [170]	4.4 [154]	5.2 [135]	6.4 [125]	12.6 [195]
$MG^5_{2,2}$-GMRES	0.2 [11]	0.5 [11]	0.8 [11]	1.5 [13]	2.6 [17]	4.1 [20]	5.9 [23]	8.8 [27]
$MG^5_{1,2}$-GMRES	0.2 [12]	0.4 [11]	0.6 [12]	1.2 [15]	2.1 [20]	3.3 [23]	4.6 [26]	7.2 [31]
$MG^5_{2,1}$-GMRES	0.2 [11]	0.4 [11]	0.6 [12]	1.1 [15]	2.0 [20]	3.1 [23]	4.6 [27]	6.2 [31]
$MG^5_{1,1}$-GMRES	0.2 [12]	0.3 [11]	0.5 [14]	1.0 [19]	1.7 [23]	2.5 [26]	3.6 [30]	5.5 [36]

Table VI.7: Strong scaling: PGMRES iterations and run-time to solve the linear system (VI.2) up to a precision of 10^{-8}, according to the experimental setting described in Section VI.6.2. We used $\mathcal{K}(\mathbf{x}) = I_2$, $q = 0$, $p = 3$, $N = 64$ time steps and $n = 384$. The total size of the space-time system (number of DoFs) is given by $64 \cdot 385^2$.

Cores	1	2	4	8	16	32
ILU(0)-GMRES	1319.0 [414]	671.1 [415]	328.7 [415]	178.9 [415]	105.6 [415]	84.7 [416]
$\text{MG}^7_{1,1}$-GMRES	339.1 [64]	173.7 [64]	87.5 [64]	48.8 [64]	30.1 [64]	26.5 [64]
$\text{TMG}^7_{1,1}$-GMRES	339.1 [64]	173.7 [64]	87.6 [64]	48.8 [64]	30.0 [64]	26.3 [64]
Cores	64	128	256	512	1024	2048
ILU(0)-GMRES	38.1 [417]	22.0 [500]	10.3 [519]	6.7 [550]	4.0 [619]	2.7 [753]
$\text{MG}^7_{1,1}$-GMRES	12.9 [64]	7.0 [64]	3.4 [65]	2.4 [65]	2.3 [65]	5.5 [65]
$\text{TMG}^7_{1,1}$-GMRES	12.8 [64]	6.3 [64]	3.1 [64]	1.8 [63]	1.0 [64]	0.6 [64]

Table VI.8: Space-time weak scaling: PGMRES iterations and run-time to solve the linear system (VI.2) up to a precision of 10^{-8}, according to the experimental setting described in Section VI.6.2. We used $\mathcal{K}(\mathbf{x}) = I_2$, $q = 0$, $p = 2$, and $(N, n) = (8, 65)$, $(16, 129)$, $(32, 256)$, $(64, 512)$. The ratio DoFs/Cores is constant in the table.

[Cores, n, N, L]	[1, 65, 8, 4]	[8, 129, 16, 5]	[64, 257, 32, 6]	[512, 513, 64, 7]
ILU(0)-GMRES	0.22 [50]	0.69 [121]	4.3 [367]	13.8 [989]
$\text{TMG}^L_{1,1}$-GMRES	0.08 [10]	0.17 [17]	0.89 [33]	2.1 [64]

VI.6.4 Scaling

In the scaling experiments, besides the multigrid already considered above, we also use a TMG for performance reasons (see Section VI.6.2 for some details/references about the TMG). From Table VI.7 we see that the proposed solver, especially when using the TMG option, shows a nearly optimal strong scaling with respect to the number of cores. Table VI.8 illustrates the weak scaling properties of the proposed solver, which possesses a remarkably superior parallel efficiency with respect to the standard ILU approach in terms of iteration count and run-time. In fact, the efficiency of the proposed solver can be estimated to be about three times the one of the standard ILU approach.

Conclusions

In the present thesis we dealt with the spectral analysis and the development of fast solvers for matrices with a Toeplitz-related structure by using a symbol approach. In this conclusive chapter we summarize the presented results and we suggest some possible future lines of research.

In **Chapter II** we described the singular and spectral distribution of special 2-by-2 block matrix-sequences. In particular, focusing on the symmetrization of the matrix-sequence $\{T_n[f]\}_n$ generated by f, we proved that $\{Y_nT_n[f]\}_n$ is essentially distributed as $\pm|f|$ in the eigenvalue sense, which informally means that roughly half of the eigenvalues of $Y_nT_n[f]$ are positive and they are approximated by a uniform sampling of $|f|$ and roughly half of the eigenvalues are negative and they are approximated by a uniform sampling of $-|f|$. As a consequence, with the choice of a suitable circulant preconditioner C_n, we proved that the preconditioned matrix-sequence $\{|C_n|^{-1}Y_nT_n[f]\}_n$ is distributed as ±1 in the eigenvalue sense. Moreover, we showed that the extension of the results to the block-Toeplitz case is possible with no particular difficulties. Conversely, extending the latter results to the multilevel case would require more work. On one hand, proving the spectral distribution of the symmetrized multilevel Toeplitz matrix-sequence $\{Y_{\mathbf{n}}T_{\mathbf{n}}[f]\}_{\mathbf{n}}$ is not as straightforward[2]. On the other hand, the performances of multilevel circulant preconditioners deteriorate as the dimensionality increases, as it has been proven in [101, 120, 124]. However, in future works we intend to derive and exploit the spectral features of such symmetrized multilevel matrix-sequences in order to mimic the unilevel construction of efficient preconditioners, which in the multilevel setting will possibly be of non-circulant type.

The encouraging results given in **Chapter II** suggested us to investigate the spectral and singular value distributions of other matrix-sequences of interest in practical applications. Hence, in **Chapter III** we have described the singular value distribution of a sequence of the form $\{h(T_n[f])\}_n$ and the eigenvalue distribution of the symmetrized sequence $\{Y_nh(T_n[f])\}_n$ in the case where $f \in L^\infty([-\pi,\pi])$ and h has convergence radius r such that $\|f\|_\infty < r$. In particular, making use of the properties of GLT sequences and under the aforementioned hypotheses on f and on the convergence radius of h, we proved that the matrix-sequence $\{h(T_n[f])\}_n$ is distributed in the singular value sense as $h \circ f$. In addition, we exploited this property to study the spectral distribution of the symmetrized sequence $\{Y_nh(T_n[f])\}_n$ and we discovered that its spectral symbol is given by

$$\phi_{|h\circ f|}(\vartheta) = \begin{cases} |h \circ f(\vartheta)|, & \vartheta \in [0, 2\pi], \\ -|h \circ f(-\vartheta)|, & \vartheta \in [-2\pi, 0), \end{cases}$$

[2]The spectral distribution of the symmetrized multilevel Toeplitz matrix-sequence $\{Y_{\mathbf{n}}T_{\mathbf{n}}[f]\}_{\mathbf{n}}$ was derived in [54] during the thesis revision time.

Finally, we numerically confirmed the derived distribution results with several experiments in different settings, also stemming from computational finance problems.

A desirable future development is the investigation on the possibility of relaxing the condition $f \in L^\infty([-\pi,\pi])$ taking a generic $f \in L^1([-\pi,\pi])$. In the latter case, a further step of analysis is required. Some preliminary considerations suggest that we could use the cut-off argument as in [136, 140] and the versatility of the a.c.s. notion. An alternative to the cut-off idea is the use of Cesàro sums to obtain sequences of polynomials that converge to f with the techniques and results derived in [112].

In **Chapter IV**, we studied multigrid strategies for linear systems having Toeplitz coefficient matrices with block entries. Our main aim was to start filling the existing theoretical gap in the convergence analysis of such methods. The resulting study indicates that the generalization is not trivial, since the commutativity property of multiplication played an essential role in the scalar case and it cannot be used in the block setting. Indeed, we proposed a general two-grid convergence analysis for positive definite block-circulant matrices, proving an optimal convergence rate independent of the matrix size under specific assumptions on the block symbol of the grid-transfer operator. In particular, we analysed a first case where the trigonometric polynomial that generates the block-circulant matrix used in the construction of the grid transfer operator is unitarily diagonalizable at all points and fulfils an appropriate commutativity condition. Moreover, we proved the approximation property for a grid transfer operator with a block symbol that might be non-diagonalizable, paying particular attention to the role of eigenvectors. Furthermore, we provided the generalization of the convergence results to multilevel block-circulant matrices, where the multilevel grid transfer operator possesses a tensor structure, and we explained how all the theory developed for block-circulants can be transferred to block-Toeplitz matrices.

A full convergence analysis for the V-cycle in our block-Toeplitz setting is still not present, but it is currently under investigation in [20], following the strategies devised in [97]. However, in the subsequent chapter we proposed a measuring instrument for the ill-conditioning of the symbol at the coarser levels that serves as a guideline to empirically choose a suitable prolongation operator for achieving fast multigrid convergence for more than two grids.

In **Chapter V**, we developed and analysed multigrid procedures for the solution of linear systems stemming from the $\mathbb{Q}_s$ Finite Elements approximation of elliptic partial differential equations with Dirichlet boundary conditions and where the operator is $\operatorname{div}(-a(\mathbf{x})\nabla\cdot)$, with a continuous and positive over $[0,1]^k$. Firstly, we proposed a classical multigrid strategy following a functional approach and we analysed the prolongation matrix as a cut block-Toeplitz matrix. Indeed, we demonstrated the convergence and optimality of such two-grid method for polynomial degree $s = 1, 2, 3$ exploiting the results of **Chapter IV**. Moreover, we performed an analogous analysis for a linear interpolation prolongation operator and in this case the convergence was proven for all even polynomial degrees. The extension of the convergence results to all polynomial degrees for both prolongation operators is currently under investigation [20]. Finally, we tested a third class of grid transfer operators, constructed according to the analysis of **Chapter IV**, that is, focusing only on algebraic considerations on the symbol of the linear system block-Toeplitz matrix-sequence. Results of numerical experiments that test all the considered methods were presented, both in one dimension and in higher dimension, showing an optimal behaviour in

terms of the dependency on the matrix size and a substantial robustness with respect to the dimensionality. We highlight that the choice of the optimal smoother from a computational point of view will be object of a further analysis aimed at devising a more competitive method, especially in the case where the matrices possess a tensor structure.

Even though we focused on the $\mathbb{Q}_s$ stiffness matrices, the presented procedures have a wider interest. Firstly, our procedure can be applied with slight changes in the case of a variation of Equation (V.1) obtained imposing different boundary conditions. In fact, the resulting stiffness matrices differ from the ones that we analysed only of a small rank correction matrix. Therefore, they share the same asymptotic spectral properties, which means that we only have to take care of possible outliers, affecting the choice of the proper smoother. Furthermore, both the geometric and the algebraic strategies could be mimicked for other discretizations and problems, given that the system matrix-sequences fulfil the required hypotheses. Among them, we cite the case of staggered discontinuous Galerkin methods for the incompressible Navier–Stokes equations [47], for whose linear systems both a two-grid and a V-cycle method have been studied in [39]. Moreover, it is of future interest the development of a multigrid method for the block-Toeplitz linear systems stemming from an IgA discretization of the Poisson problem with splines of non-maximal regularity, which would also be useful for an extension of the work that we did in **Chapter VI**.

Indeed, in **Chapter VI**, we have proposed a new solver for the space-time IgA-DG discretization of the anisotropic diffusion problem (VI.1), where the spline functions used for the spacial component have maximum regularity. The method combined a suitable preconditioned GMRES algorithm with a few iterations of an appropriate multigrid method, both devised taking inspiration from the spectral analysis in [16]. Through numerical experiments, we have illustrated the competitiveness of our proposal with respect to other benchmark solvers in terms of iteration count, run-times and scaling. In particular, the solver is suited for parallel computation as it shows remarkable scalability properties with respect to the number of cores. In addition, we highlight that the proposed solver is highly flexible as it does not depend on the domain or the space-time discretization, as long as a tensor-product structure is maintained between space and time.

However, many significant steps could still be performed. Firstly, a future item of research is the theoretical convergence analysis of the proposed solver. Moreover, it would be interesting to investigate the performance of the solver for the anisotropic diffusion problem (VI.1) in the case of a space domain Ω more complex than the hypersquare $(0,1)^k$ introducing a geometry parametrization. Finally, a computational improvement could be obtained by considering an inner/outer multilevel hierarchy in time to improve the overall scalability of the proposed solver, for example, using it as a smoother in a multigrid-in-time algorithm.

In conclusion, we think that the present thesis inserts some missing pieces in the beautiful and intricate puzzle of Toeplitz-related structures, which still needs to be completed with the efforts of future theoretical and applicative research studies.

Bibliography

[1] R. Abedi, B. Petracovici, and R. B. Haber. A space-time discontinuous Galerkin method for linearized elastodynamics with element-wise momentum balance. *Comput. Methods Appl. Mech. Engrg.*, 195(25-28):3247–3273, 2006. p. 101

[2] P. Arbenz, D. Hupp, and D. Obrist. A parallel solver for the time-periodic Navier-Stokes equations. In *Parallel processing and applied mathematics. Part II*, volume 8385 of *Lecture Notes in Comput. Sci.*, pages 291–300. Springer, Heidelberg, 2014. p. 101

[3] A. Aricò and M. Donatelli. A V-cycle multigrid for multilevel matrix algebras: proof of optimality. *Numer. Math.*, 105(4):511–547, 2007. p. ix

[4] A. Aricò, M. Donatelli, and S. Serra-Capizzano. V-cycle optimal convergence for certain (multilevel) structured linear systems. *SIAM J. Matrix Anal. Appl.*, 26(1):186–214, 2004. p. ix, 20, 55, 56, 78

[5] F. Auricchio, L. Beirão da Veiga, T. J. R. Hughes, A. Reali, and G. Sangalli. Isogeometric collocation methods. *Math. Models Methods Appl. Sci.*, 20(11):2075–2107, 2010. p. v, 102

[6] F. Avram. On bilinear forms in Gaussian random variables and Toeplitz matrices. *Probab. Theory Related Fields*, 79(1):37–45, 1988. p. vi, 12

[7] O. Axelsson and G. Lindskog. On the rate of convergence of the preconditioned conjugate gradient method. *Numer. Math.*, 48(5):499–523, 1986. p. 18

[8] O. Axelsson and M. Neytcheva. The algebraic multilevel iteration methods—theory and applications. In *Proceedings of the Second International Colloquium on Numerical Analysis (Plovdiv, 1993)*, pages 13–23. VSP, Utrecht, 1994. p. 16

[9] A. K. Aziz and P. Monk. Continuous finite elements in space and time for the heat equation. *Math. Comp.*, 52(186):255–274, 1989. p. 101

[10] S. Balay, S. Abhyankar, M. F. Adams, J. Brown, P. Brune, K. Buschelman, L. Dalcin, A. Dener, V. Eijkhout, W.D. Gropp, D. Karpeyev, D. Kaushik, M. G. Knepley, D.A. May, L. C. McInnes, R. T. Mills, T. Munson, K. Rupp, P. Sanan, B.F. Smith, S. Zampini, H. Zhang, and H. Zhang. PETSc Web page. https://www.mcs.anl.gov/petsc, 2019. p. 110

[11] S. Balay, S. Abhyankar, M. F. Adams, J. Brown, P. Brune, K. Buschelman, L. Dalcin, A. Dener, V. Eijkhout, W.D. Gropp, D. Karpeyev, D. Kaushik, M. G. Knepley, D.A. May,

L. C. McInnes, R. T. Mills, T. Munson, K. Rupp, P. Sanan, B.F. Smith, S. Zampini, H. Zhang, and H. Zhang. PETSc users manual. Technical Report ANL-95/11 - Revision 3.13, Argonne National Laboratory, 2020. p. 110

[12] G. Barbarino, C. Garoni, and S. Serra-Capizzano. Block generalized locally Toeplitz sequences: theory and applications in the multidimensional case. *Electron. Trans. Numer. Anal.*, 53:113–216, 2020. p. vii, 14

[13] G. Barbarino, C. Garoni, and S. Serra-Capizzano. Block generalized locally Toeplitz sequences: theory and applications in the unidimensional case. *Electron. Trans. Numer. Anal.*, 53:28–112, 2020. p. vii, 105

[14] L. Beirão da Veiga, A. Buffa, G. Sangalli, and R. Vázquez. Mathematical analysis of variational isogeometric methods. *Acta Numer.*, 23:157–287, 2014. p. 102

[15] P. Benedusi, C. Garoni, R. Krause, P. Ferrari, and S. Serra-Capizzano. Fast parallel solver for the space-time iga-dg discretization of the anisotropic diffusion equation. Technical Report 2019-011, Department of Information Technology, Uppsala University, 2019. p. x, 102

[16] P. Benedusi, C. Garoni, R. Krause, X. Li, and S. Serra-Capizzano. Space-time FE-DG discretization of the anisotropic diffusion equation in any dimension: the spectral symbol. *SIAM J. Matrix Anal. Appl.*, 39(3):1383–1420, 2018. p. vii, x, 101, 102, 104, 117

[17] P. Benedusi, D. Hupp, P. Arbenz, and R. Krause. A parallel multigrid solver for time-periodic incompressible Navier-Stokes equations in 3D. In *Numerical mathematics and advanced applications—ENUMATH 2015*, volume 112 of *Lect. Notes Comput. Sci. Eng.*, pages 265–273. Springer, [Cham], 2016. p. 101

[18] P. Betsch and P. Steinmann. Conservation properties of a time FE method. II. Time-stepping schemes for non-linear elastodynamics. *Internat. J. Numer. Methods Engrg.*, 50(8):1931–1955, 2001. p. 101

[19] R. Bhatia. *Matrix analysis*, volume 169 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, 1997. p. 5

[20] M. Bolten, M. Donatelli, P. Ferrari, and I. Furci. A symbol based analysis for multigrid methods for block-circulant and block-toeplitz systems. In preparation. p. ix, 56, 74, 89, 91, 116

[21] M. Bolten, M. Donatelli, T. Huckle, and C. Kravvaritis. Generalized grid transfer operators for multigrid methods applied on Toeplitz matrices. *BIT*, 55(2):341–366, 2015. p. ix, 91

[22] A. Böttcher and S. M. Grudsky. *Toeplitz matrices, asymptotic linear algebra, and functional analysis*. Birkhäuser Verlag, Basel, 2000. p. 8

[23] A. Böttcher, S. M. Grudsky, and E. A. Maximenko. Inside the eigenvalues of certain Hermitian Toeplitz band matrices. *J. Comput. Appl. Math.*, 233:2245–2264, 2010. p. v

[24] A. Böttcher and B. Silbermann. *Introduction to Large Truncated Toeplitz Matrices.* Springer, 1999. p. 8

[25] A. Böttcher and B. Silbermann. *Analysis of Toeplitz operators.* Springer Monographs in Mathematics. Springer-Verlag, Berlin, second edition, 2006. Prepared jointly with Alexei Karlovich. p. 8

[26] A. Brandt. Multi-level adaptive solutions to boundary-value problems. *Math. Comp.*, 31(138):333–390, 1977. p. 79, 84

[27] R. H. Chan, Q.-S. Chang, and H.-W. Sun. Multigrid method for ill-conditioned symmetric Toeplitz systems. *SIAM J. Sci. Comput.*, 19(2):516–529, 1998. p. ix, 55

[28] R. H. Chan and M. Ng. Conjugate gradient methods for Toeplitz systems. *SIAM Review*, 38(3):427–482, 1996. p. 19

[29] R. H. Chan and G. Strang. Toeplitz equations by Conjugate Gradients with circulant preconditioner. *SIAM J. Sci. Statist. Comput.*, 10(1):104–119, 1989. p. v, viii

[30] R. H. Chan and M. C. Yeung. Circulant preconditioners for Toeplitz matrices with positive continuous generating functions. *Math. Comp.*, 58:233–240, 1992. p. viii, 20

[31] T. Chan. An optimal circulant preconditioner for Toeplitz systems. *SIAM J. Sci. Comput.*, 9(4):766–771, 1988. p. v, viii

[32] J. W. Cooley and J. W. Tukey. An algorithm for the machine calculation of complex Fourier series. *Math. Comp.*, 19:297–301, 1965. p. v, 14

[33] J. A. Cottrell, T. J. R. Hughes, and Y. Bazilevs. *Isogeometric analysis.* John Wiley & Sons, Ltd., Chichester, 2009. Toward integration of CAD and FEA. p. v, 102

[34] L. Dalcin, N. Collier, P. Vignal, A. M. A. Côrtes, and V. M. Calo. PetIGA: a framework for high-performance isogeometric analysis. *Comput. Methods Appl. Mech. Engrg.*, 308:151–181, 2016. p. 110

[35] P. Davis. *Circulant Matrices.* J. Wiley and Sons, New York, 1979. p. v

[36] V. Del Prete, F. Di Benedetto, M. Donatelli, and S. Serra-Capizzano. Symbol approach in a signal-restoration problem involving block Toeplitz matrices. *J. Comput. Appl. Math.*, 272:399–416, 2014. p. vi, 55

[37] M. Donatelli. An algebraic generalization of local Fourier analysis for grid transfer operators in multigrid based on Toeplitz matrices. *Numer. Linear Algebra Appl.*, 17(2-3):179–197, 2010. p. 56, 58

[38] M. Donatelli, A. Dorostkar, M. Mazza, M. Neytcheva, and S. Serra-Capizzano. Function-based block multigrid strategy for a two-dimensional linear elasticity-type problem. *Comput. Math. Appl.*, 74(5):1015–1028, 2017. p. 55, 60

[39] M. Donatelli, P. Ferrari, I. Furci, D. Sesana, and S. Serra-Capizzano. Multigrid methods for block-circulant and block-Toeplitz large linear systems: Algorithmic proposals and two-grid optimality analysis. *Numer. Linear Algebra Appl.* Accepted. p. ix, 56, 74, 117

[40] M. Donatelli, C. Garoni, C. Manni, S. Serra-Capizzano, and H. Speleers. Robust and optimal multi-iterative techniques for IgA Galerkin linear systems. *Comput. Methods Appl. Mech. Engrg.*, 284:230–264, 2015. p. viii

[41] M. Donatelli, C. Garoni, C. Manni, S. Serra-Capizzano, and H. Speleers. Spectral analysis and spectral symbol of matrices in isogeometric collocation methods. *Math. Comp.*, 85(300):1639–1680, 2016. p. vi

[42] M. Donatelli, C. Garoni, C. Manni, S. Serra-Capizzano, and H. Speleers. Symbol-based multigrid methods for Galerkin B-spline isogeometric analysis. *SIAM J. Numer. Anal.*, 55(1):31–62, 2017. p. vi

[43] M. Donatelli, M. Molteni, V. Pennati, and S. Serra-Capizzano. Multigrid methods for cubic spline solution of two point (and 2D) boundary value problems. *Appl. Numer. Math.*, 104:15–29, 2016. p. 55

[44] M. Donatelli, S. Serra-Capizzano, and D. Sesana. Multigrid methods for Toeplitz linear systems with different size reduction. *BIT*, 52(2):305–327, 2012. p. 58

[45] C.C. Douglas. *A Review of Numerous Parallel Multigrid Methods*, chapter 17, pages 187–202. Society for Industrial and Applied Mathematics, 1996. p. 111

[46] D. J. Duffy. *Finite difference methods in financial engineering.* Wiley Finance Series. John Wiley & Sons, Ltd., Chichester, 2006. p. 43

[47] M. Dumbser, F. Fambri, I. Furci, M. Mazza, S. Serra-Capizzano, and M. Tavelli. Staggered discontinuous Galerkin methods for the incompressible Navier-Stokes equations: spectral analysis and computational results. *Numer. Linear Algebra Appl.*, 25(5):e2151, 31, 2018. p. 117

[48] K. Eriksson, C. Johnson, and A. Logg. *Adaptive Computational Methods for Parabolic Problems*, chapter 24. American Cancer Society, 2004. p. 101

[49] C. Estatico and S. Serra-Capizzano. Superoptimal approximation for unbounded symbols. *Linear Algebra Appl.*, 428(2-3):564–585, 2008. p. 19, 30

[50] R. D. Falgout, S. Friedhoff, Tz. V. Kolev, S. P. MacLachlan, J. B. Schroder, and S. Vandewalle. Multigrid methods with space-time concurrency. *Comput. Vis. Sci.*, 18(4-5):123–143, 2017. p. 101

[51] D. Fasino and P. Tilli. Spectral clustering properties of block multilevel Hankel matrices. *Linear Algebra Appl.*, 306(1-3):155–163, 2000. p. 27

[52] P. Ferrari, N. Barakitis, and S. Serra-Capizzano. Asymptotic spectra of large matrices coming from the symmetrization of Toeplitz structure functions and applications to preconditioning. *Numer. Linear Algebra Appl.*, e2332, 2020. p. v, vi, viii, 44

[53] P. Ferrari, I. Furci, S. Hon, M. A. Mursaleen, and S. Serra-Capizzano. The eigenvalue distribution of special 2-by-2 block matrix-sequences with applications to the case of symmetrized Toeplitz structures. *SIAM J. Matrix Anal. Appl.*, 40(3):1066–1086, 2019. p. vi, viii, 24

[54] P. Ferrari, I. Furci, and S. Serra-Capizzano. Multilevel symmetrized toeplitz structures and spectral distribution results for the related matrix-sequences. *ArXiv*:2011.10835, 2020. p. 115

[55] P. Ferrari, R. I. Rahla, C. Tablino-Possio, S. Belhaj, and S. Serra-Capizzano. Multigrid for $\mathbb{Q}_k$ finite element matrices using a (block) Toeplitz symbol approach. *Mathematics*, 8(5), 2020. p. x, 74

[56] G. Fiorentino and S. Serra. Multigrid methods for Toeplitz matrices. *Calcolo*, 28(3-4):283–305 (1992), 1991. p. ix, 55, 56, 57, 78

[57] G. Fiorentino and S. Serra. Multigrid methods for symmetric positive definite block Toeplitz matrices with nonnegative generating functions. *SIAM J. Sci. Comput.*, 17(5):1068–1081 (1996), 1996. p. ix, 55

[58] D. A. French. A space-time finite element method for the wave equation. *Comput. Methods Appl. Mech. Engrg.*, 107(1-2):145–157, 1993. p. vii

[59] M. J. Gander. 50 years of time parallel time integration. In *Multiple shooting and time domain decomposition methods*, volume 9 of *Contrib. Math. Comput. Sci.*, pages 69–113. Springer, Cham, 2015. p. 101

[60] M. J. Gander and L. Halpern. Techniques for Locally Adaptive Time Stepping Developed over the Last Two Decades. In *Domain Decomposition Methods in Science and Engineering XX*, pages 377–385, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg. p. 101

[61] M. J. Gander and M. Neumüller. Analysis of a new space-time parallel multigrid algorithm for parabolic problems. *SIAM J. Sci. Comput.*, 38(4):A2173–A2208, 2016. p. 102

[62] C. Garoni and S. Serra-Capizzano. *Generalized locally Toeplitz sequences: theory and applications. Vol. I.* Springer, Cham, 2017. p. vii, 6, 7, 8, 9, 10, 12, 14, 25, 28, 30, 45

[63] C. Garoni and S. Serra-Capizzano. *Generalized locally Toeplitz sequences: theory and applications. Vol. II.* Springer, Cham, 2018. p. vii, 14

[64] C. Garoni, S. Serra-Capizzano, and D. Sesana. Spectral analysis and spectral symbol of d-variate $\mathbb{Q}_p$ Lagrangian FEM stiffness matrices. *SIAM J. Matrix Anal. Appl.*, 36(3):1100–1128, 2015. p. v, ix, 73, 75, 76, 98

[65] G. H. Golub and C. F. Van Loan. *Matrix computations*, volume 3 of *Johns Hopkins Series in the Mathematical Sciences*. Johns Hopkins University Press, Baltimore, MD, 1983. p. 4, 63

[66] A. Greenbaum. *Iterative methods for solving linear systems*, volume 17 of *Frontiers in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1997. p. vii, 18

[67] U. Grenander and G. Szegő. *Toeplitz forms and their applications*. Chelsea Publishing Co., New York, second edition, 1984. p. vi, 8, 12

[68] M. Griebel and D. Oeltz. A sparse grid space-time discretization scheme for parabolic problems. *Computing*, 81(1):1–34, 2007. p. 101

[69] W. Hackbusch. *Multigrid methods and applications*, volume 4 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, 1985. p. 78, 79, 84

[70] W. Hackbusch. *Iterative solution of large sparse systems of equations*, volume 95 of *Applied Mathematical Sciences*. Springer-Verlag, New York, 1994. Translated and revised from the 1991 German original. p. vii, 79, 84

[71] M. R. Hestenes and E. Stiefel. Methods of conjugate gradients for solving linear systems. *J. Research Nat. Bur. Standards*, 49:409–436 (1953), 1952. p. vii, 18

[72] J. S. Hesthaven and T. Warburton. *Nodal discontinuous Galerkin methods*, volume 54 of *Texts in Applied Mathematics*. Springer, New York, 2008. Algorithms, analysis, and applications. p. 104, 111

[73] N. J. Higham. *Accuracy and stability of numerical algorithms*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, second edition, 2002. p. vii

[74] N. J. Higham. *Functions of matrices*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2008. Theory and computation. p. 15, 44

[75] N. J. Higham and P. Kandolf. Computing the action of trigonometric and hyperbolic matrix functions. *SIAM J. Sci. Comput.*, 39(2):A613–A627, 2017. p. 43

[76] C. Hofer, U. Langer, M. Neumüller, and R. Schneckenleitner. Parallel and robust preconditioning for space-time isogeometric analysis of parabolic evolution problems. *SIAM J. Sci. Comput.*, 41(3):A1793–A1821, 2019. p. 101, 102

[77] S. Hon, M. A. Mursaleen, and S. Serra-Capizzano. A note on the spectral distribution of symmetrized Toeplitz sequences. *Linear Algebra Appl.*, 579:32–50, 2019. p. 24, 30

[78] S. Hon and A. Wathen. Circulant preconditioners for analytic functions of Toeplitz matrices. *Numer. Algorithms*, 79(4):1211–1230, 2018. p. 15, 43, 47, 50, 51, 52

[79] S. Hon and A. Wathen. *Numerical Investigation of the Spectral Distribution of Toeplitz-Function Sequences*, volume 36. Springer, Cham, 2019. Computational Methods for Inverse Problems in Imaging. Springer INdAM Series. p. 43

[80] Roger A. Horn and Charles R. Johnson. *Matrix analysis*. Cambridge University Press, Cambridge, 1985. p. 23

[81] G. Horton and S. Vandewalle. A space-time multigrid method for parabolic partial differential equations. *SIAM J. Sci. Comput.*, 16(4):848–864, 1995. p. 101

[82] T. Huckle and J. Staudacher. Multigrid preconditioning and Toeplitz matrices. *Electron. Trans. Numer. Anal.*, 13:81–105, 2002. p. ix

[83] T. Huckle and J. Staudacher. Multigrid methods for block Toeplitz matrices with small size blocks. *BIT*, 46(1):61–83, 2006. p. 55

[84] T. J. R. Hughes, J. A. Cottrell, and Y. Bazilevs. Isogeometric analysis: CAD, finite elements, NURBS, exact geometry and mesh refinement. *Comput. Methods Appl. Mech. Engrg.*, 194(39-41):4135–4195, 2005. p. 102

[85] T. J. R. Hughes and G. M. Hulbert. Space-time finite element methods for elastodynamics: formulations and error estimates. *Comput. Methods Appl. Mech. Engrg.*, 66(3):339–363, 1988. p. 101

[86] C. M. Klaij, J. J. W. van der Vegt, and H. van der Ven. Space-time discontinuous Galerkin method for the compressible Navier-Stokes equations. *J. Comput. Phys.*, 217(2):589–611, 2006. p. 101, 102

[87] D. Krause and R. Krause. Enabling local time stepping in the parallel implicit solution of reaction-diffusion equations via space-time finite elements on shallow tree meshes. *Appl. Math. Comput.*, 277:164–179, 2016. p. 101

[88] O. A. Ladyženskaja, V. A. Solonnikov, and N. N. Ural'ceva. *Linear and quasilinear equations of parabolic type*. Translated from the Russian by S. Smith. Translations of Mathematical Monographs, Vol. 23. American Mathematical Society, Providence, R.I., 1968. p. 101

[89] U. Langer, S. E. Moore, and M. Neumüller. Space-time isogeometric analysis of parabolic evolution problems. *Comput. Methods Appl. Mech. Engrg.*, 306:342–363, 2016. p. 101

[90] S. T. Lee, X. Liu, and H.-W. Sun. Fast exponential time integration scheme for option pricing with jumps. *Numer. Linear Algebra Appl.*, 19(1):87–101, 2012. p. 47

[91] S. T. Lee, H.-K. Pang, and H.-W. Sun. Shift-invert Arnoldi approximation to the Toeplitz matrix exponential. *SIAM J. Sci. Comput.*, 32(2):774–792, 2010. p. 47

[92] D. A. May, P. Sanan, K. Rupp, M. G. Knepley, and B. F. Smith. Extreme-scale multigrid components within PETSc. In *Proceedings of the Platform for Advanced Scientific Computing Conference*, PASC '16, New York, NY, USA, 2016. Association for Computing Machinery. p. 111

[93] M. Mazza and J. Pestana. Spectral properties of flipped Toeplitz matrices and related preconditioning. *BIT*, 59(2):463–482, 2019. p. 24

[94] M. Mazza, A. Ratnani, and S. Serra-Capizzano. Spectral analysis and spectral symbol for the 2D curl-curl (stabilized) operator with applications to the related iterative solutions. *Math. Comp.*, 88(317):1155–1188, 2019. p. 55

[95] D. Meidner and B. Vexler. Adaptive space-time finite element methods for parabolic optimization problems. *SIAM J. Control Optim.*, 46(1):116–142, 2007. p. 101

[96] S. T. Miller and R. B. Haber. A spacetime discontinuous Galerkin method for hyperbolic heat conduction. *Comput. Methods Appl. Mech. Engrg.*, 198(2):194–209, 2008. p. 101

[97] A. Napov and Y. Notay. When does two-grid optimality carry over to the V-cycle? *Numer. Linear Algebra Appl.*, 17(2-3):273–290, 2010. p. 91, 116

[98] M. Neumüller and O. Steinbach. Refinement of flexible space-time finite element meshes and discontinuous Galerkin methods. *Comput. Vis. Sci.*, 14(5):189–205, 2011. p. 101

[99] M. Ng. *Iterative methods for Toeplitz systems (Numerical Mathematics and Scientific Computation)*. Oxford University Press, New York., 2004. p. vii, 9, 19, 31

[100] E. Ngondiep, S. Serra-Capizzano, and D. Sesana. Spectral features and asymptotic properties for g-circulants and g-Toeplitz sequences. *SIAM J. Matrix Anal. Appl.*, 31(4):1663–1687, 2009/10. p. v, 58

[101] D. Noutsos, S. Serra-Capizzano, and P. Vassalos. Matrix algebra preconditioners for multilevel Toeplitz systems do not insure optimal convergence rate. *Theoret. Comput. Sci.*, 315(2):557–579, 2004. p. viii, 115

[102] C. C. Paige and M. A. Saunders. Solutions of sparse indefinite systems of linear equations. *SIAM J. Numer. Anal.*, 12(4):617–629, 1975. p. vii, 18

[103] S. V. Parter. On the distribution of the singular values of Toeplitz matrices. *Linear Algebra Appl.*, 80:115–130, 1986. p. vi, 12

[104] J. Pestana and A. J. Wathen. A preconditioned MINRES method for nonsymmetric Toeplitz matrices. *SIAM J. Matrix Anal. Appl.*, 36(1):273–288, 2015. p. vi, viii, xi, 24, 30, 31

[105] A. Quarteroni. *Numerical models for differential problems*, volume 2 of *MS&A. Modeling, Simulation and Applications*. Springer-Verlag Italia, Milan, 2009. Translated from the 4th (2008) Italian edition by Silvia Quarteroni. p. 101

[106] R. I. Rahla, S. Serra-Capizzano, and C. Tablino-Possio. Spectral analysis of $\mathbb{P}_k$ finite element matrices in the case of friedrichs–keller triangulations via generalized locally Toeplitz technology. *Numer. Linear Algebra Appl.*, 27(4):e2302, 2020. p. 76

[107] J. W. Ruge and K. Stüben. Algebraic multigrid. In *Multigrid methods*, volume 3 of *Frontiers Appl. Math.*, pages 73–130. SIAM, Philadelphia, PA, 1987. p. ix, 20, 55

[108] Y. Saad. *Iterative methods for sparse linear systems*. Society for Industrial and Applied Mathematics, Philadelphia, PA, second edition, 2003. p. vii, 17, 102, 105

[109] Y. Saad and M. H. Schultz. GMRES: a generalized minimal residual algorithm for solving nonsymmetric linear systems. *SIAM J. Sci. Statist. Comput.*, 7:856–869, 1986. p. vii, 18

[110] D. Schötzau and C. Schwab. An *hp* a priori error analysis of the DG time-stepping method for initial value problems. *Calcolo*, 37(4):207–232, 2000. p. 102

[111] S. Serra. On the conditioning and the solution, by means of multigrid methods, of symmetric (block) Toeplitz linear systems. In *Proceedings of the Sixth International Colloquium on Differential Equations (Plovdiv, 1995)*, pages 249–256. VSP, Utrecht, 1996. p. 16

[112] S. Serra and P. Tilli. On unitarily invariant norms of matrix-valued linear positive operators. *J. Inequal. Appl.*, 7(3):309–330, 2002. p. 116

[113] S. Serra-Capizzano. On the extreme spectral properties of Toeplitz matrices generated by L^1 functions with several minima/maxima. *Linear Algebra Appl.*, 36:135–142, 1996. p. vi

[114] S. Serra-Capizzano. A Korovkin-type theory for finite Toeplitz operators via matrix algebras. *Numer. Math.*, 82:117–142, 1999. p. v, viii, 19, 20

[115] S. Serra-Capizzano. Spectral and computational analysis of block Toeplitz matrices having nonnegative definite matrix-valued generating functions. *BIT*, 39(1):152–175, 1999. p. vi

[116] S. Serra-Capizzano. Superlinear PCG methods for symmetric Toeplitz systems. *Math. Comp.*, 88:793–803, 1999. p. viii, 19, 20

[117] S. Serra-Capizzano. Korovkin tests, approximation, and ergodic theory. *Math. Comp.*, 69(232):1533–1558, 2000. p. 30

[118] S. Serra-Capizzano. Distribution results on the algebra generated by Toeplitz sequences: a finite-dimensional approach. *Linear Algebra Appl.*, 328(1-3):121–130, 2001. p. 8

[119] S. Serra-Capizzano. Convergence analysis of two-grid methods for elliptic Toeplitz and PDEs matrix–sequences. *Numer. Math.*, 92(3):433–465, 2002. p. ix, 68

[120] S. Serra-Capizzano. Matrix algebra preconditioners for multilevel Toeplitz matrices are not superlinear. *Linear Algebra Appl.*, 343:303–319, 2002. p. viii, 115

[121] S. Serra-Capizzano. Generalized locally Toeplitz sequences: spectral analysis and applications to discretized partial differential equations. *Linear Algebra Appl.*, 366:371–402, 2003. Special issue on structured matrices: analysis, algorithms and applications (Cortona, 2000). p. 14

[122] S. Serra-Capizzano. The GLT class as a generalized Fourier analysis and applications. *Linear Algebra Appl.*, 419(1):180–233, 2006. p. 14

[123] S. Serra-Capizzano and C. Tablino-Possio. Multigrid methods for multilevel circulant matrices. *SIAM J. Sci. Comput.*, 26(1):55–85, 2004. p. v

[124] S. Serra-Capizzano and E. Tyrtyshnikov. Any circulant-like preconditioner for multilevel matrices is not superlinear. *SIAM J. Matrix Anal. Appl.*, 21(2):431–439, 2000. p. viii, 115

[125] F. Shakib, T. J. R. Hughes, and Z. Johan. A new finite element formulation for computational fluid dynamics. X. The compressible Euler and Navier-Stokes equations. *Comput. Methods Appl. Mech. Engrg.*, 89:141–219, 1991. Second World Congress on Computational Mechanics, Part I (Stuttgart, 1990). p. 101

[126] J. J. Sudirham, J. J. W. van der Vegt, and R. M. J. van Damme. Space-time discontinuous Galerkin method for advection-diffusion problems on time-dependent domains. *Appl. Numer. Math.*, 56(12):1491–1518, 2006. p. 101, 102

[127] H.-W. Sun, X.-Q. Jin, and Q.-S. Chang. Convergence of the multigrid method for ill-conditioned block Toeplitz systems. *BIT*, 41(1):179–190, 2001. p. ix

[128] T. E. Tezduyar, M. Behr, and J. Liou. A new strategy for finite element computations involving moving boundaries and interfaces—the deforming-spatial-domain/space-time procedure. I. The concept and the preliminary numerical tests. *Comput. Methods Appl. Mech. Engrg.*, 94(3):339–351, 1992. p. 101

[129] T. E. Tezduyar, M. Behr, S. Mittal, and J. Liou. A new strategy for finite element computations involving moving boundaries and interfaces—the deforming-spatial-domain/space-time procedure. II. Computation of free-surface flows, two-liquid flows, and flows with drifting cylinders. *Comput. Methods Appl. Mech. Engrg.*, 94(3):353–371, 1992. p. 101

[130] T. E. Tezduyar, S. Sathe, R. Keedy, and K. Stein. Space-time finite element techniques for computation of fluid-structure interactions. *Comput. Methods Appl. Mech. Engrg.*, 195(17-18):2002–2027, 2006. p. 101

[131] S. Thite. Adaptive spacetime meshing for discontinuous Galerkin methods. *Comput. Geom.*, 42(1):20–44, 2009. p. 101

[132] V. Thomée. *Galerkin finite element methods for parabolic problems*, volume 25 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, second edition, 2006. p. 102

[133] P. Tilli. A note on the spectral distribution of Toeplitz matrices. *Linear and Multilinear Algebra*, 45(2-3):147–159, 1998. p. v, 12

[134] E. E. Tyrtyshnikov. New theorems on the distribution of eigenvalues and singular values of multilevel Toeplitz matrices. *Dokl. Akad. Nauk*, 333(3):300–303, 1993. p. vi, 12

[135] E. E. Tyrtyshnikov. A unifying approach to some old and new theorems on distribution and clustering. *Linear Algebra Appl.*, 232:1–43, 1996. p. vi, 12

[136] E. E. Tyrtyshnikov and N. L. Zamarashkin. Spectra of multilevel Toeplitz matrices: advanced theory via simple matrix relationships. *Linear Algebra Appl.*, 270:15–27, 1998. p. 116

[137] J. J. W. van der Vegt and H. van der Ven. Space-time discontinuous Galerkin finite element method with dynamic grid motion for inviscid compressible flows. I. General formulation. *J. Comput. Phys.*, 182(2):546–585, 2002. p. 101

[138] C. Van Loan. *Computational Frameworks for the Fast Fourier Transform.* SIAM, Philadelphia, 1992. p. v, 14

[139] J. Česenek and M. Feistauer. Theory of the space-time discontinuous Galerkin method for nonstationary parabolic problems with nonlinear convection and diffusion. *SIAM J. Numer. Anal.*, 50(3):1181–1206, 2012. p. 101

[140] N. Zamarashkin and E. Tyrtyshnikov. Distribution of the eigenvalues and singular numbers of Toeplitz matrices under weakened requirements on the generating function. *Mat. Sb.*, 188(3):83–92, 1997. p. v, vi, 12, 116

[141] P. Zulian, Kopaničáková A., Nestola M. C. G., Fink A., Fadel N., Magri V., Schneider T., Botter E., and Mankau J. Utopia: a C++ embedded domain specific language for scientific computing. git repository. https://bitbucket.org/zulianp/utopia, 2016. p. 110

Acknowledgements

Firstly, I would like to thank my supervisors Marco Donatelli and Stefano Serra–Capizzano for their continuous and precious support during these three years.

Many thanks to all my collaborators, without whom this work would have not been possible, and to the referees for their remarks, which have improved the quality of my thesis.

Many thanks to Matthias Bolten and Isabella Furci for welcoming me in Wuppertal during such a difficult time for travelling. A very special "thank you" to Isabella for making me feel at home.

Finally, I would like to thank everybody who gave me the strength to continue during all the difficult times, with a kind word, a hug or a song.