UNIVERSITÀ DEGLI STUDI DELL'INSUBRIA

Dipartimento di Scienza ed Alta Tecnologia
Dottorato di Ricerca in Matematica del Calcolo: Modelli, Strutture, Algoritmi ed Applicazioni

Spectral features of matrix-sequences, GLT, symbol, and application in preconditioning Krylov methods, image deblurring, and multigrid algorithms

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

Ph.D. thesis of:
Mariarosa Mazza
Matr. N. 720737

Accademic year 2015-2016

Alla mia dolcissima

nonna Maria Rosa

Contents

Introduction

The final purpose and the very essence of any scientific discipline can be regarded as the solution of real-world problems. With this aim, a mathematical modeling of the considered phenomenon is often compulsory. Closed-form solutions of the arising functional (differential or integral) equations are usually not available and numerical discretization techniques are required. It is not uncommon that the used discretization leads to large linear or nonlinear systems, whose size depends on the number of discretization points and the greater the number of such points, the better the accuracy in the solution. In this setting, when approximating an infinite-dimensional linear equation via some linear approximation method, one finds a sequence of linear systems $\{A_n x_n = b_n\}$ of increasing dimension d_n. The coefficient matrices $\{A_n\}$ could inherit a *structure* from the continuous problem. Many applications such as Markov chains [43, 107], the reconstruction of signals with missing data [44], the inpainting problem [28], and of course the numerical approximation of constant coefficient $s \times s$ systems of Partial Differential Equations (PDEs) over k-dimensional domains [5], give rise to sequences $\{A_n\}$, where A_n is a *multilevel block Toeplitz matrix*. Nevertheless, there are situations, e.g. the approximation by local methods (finite differences, finite elements, isogeometric analysis, etc) of PDEs with nonconstant coefficients, general domains and nonuniform gridding, in which the class of Toeplitz matrices is no longer sufficient and we require the so-called *Generalized Locally Toeplitz* (GLT) algebra (see the pioneering work by Tilli [141], and the generalization in [131, 133]). In short, the latter is an algebra containing sequences of matrices including the Toeplitz sequences and virtually any sequence of matrices coming from 'reasonable' approximations by local discretization methods of PDEs.

For the resolution of structured large linear systems, the direct methods may require an high computation and, generally, they do not exploit the information on the matrix structure in order to accelerate the convergence. Conversely, iterative methods, and especially multigrid or preconditioned Krylov techniques, are more easily adaptable to problems with specific structural features.

It is well-known that the convergence properties of preconditioned Krylov and multigrid methods strongly depend on the spectral features of the resulting coefficient matrix. In the context of structured matrices, the spectral analysis is strictly related to the notion of *symbol*, a function whose role relies in describing the asymptotical distribution of the spectrum. More in detail, a sequence of matrices $\{A_n\}$ of size d_n is distributed in the eigenvalues sense as a measurable function $f : G \to \mathbb{C}^{s\times s}$, defined on a measurable set $G \subset \mathbb{R}^k$ with $0 < m_k(G) < \infty$, where m_k is the Lebesgue measure in $\mathbb{R}^k$, if for every continuous function F with bounded support on $\mathbb{C}$ we have

$$\lim_{n\to\infty} \frac{1}{d_n} \sum_{i=1}^{d_n} F(\lambda_i(A_n)) = \frac{1}{m_k(G)} \int_G \frac{\sum_{j=1}^{s} F(\lambda_j(f(t)))}{s} dt,$$

where $\lambda_i(A_n)$, $i = 1, \ldots, d_n$ are the eigenvalues of A_n, and $\lambda_j(f(t))$, $j = 1, \ldots, s$ are the eigenvalues of $f(t)$. In this situation, f is called the symbol of $\{A_n\}$, and it provides a 'compact' description of the asymptotic spectral distribution of A_n. In other words, an evaluation of $\lambda_j(f(t))$, $j = 1, \ldots, s$ over a uniform equispaced gridding in the domain leads to a reasonable approximation of the eigenvalues of A_n, when d_n is sufficiently large.

The purpose of this thesis is both of theoretical and practical nature. On the one hand, in Chapter 2 we enlarge some known spectral distribution tools presented in Chapter 1 by proving the eigenvalue distribution of matrix-sequences obtained as combination of some algebraic operations on multilevel block Toeplitz matrices. On the other hand, in the same chapter, we take advantage of the obtained results for designing efficient preconditioning techniques. Moreover, in the remaining Chapters 3-5, we focus on the numerical solution of structured linear systems coming from the following applications: image deblurring, Fractional Diffusion Equations (FDEs) and coupled PDEs. A spectral analysis of the arising structured sequences will allow us either to study the convergence and predict the behavior of iterative methods like preconditioned Krylov and multigrid methods applied to A_n, or to design effective

preconditioners and multigrid solvers for the associated linear systems.

The philosophy behind a *preconditioning* strategy for well-posed problems is to cluster as well as possible the spectrum of the coefficient matrix in order to speed up the convergence of iterative methods. Indeed, it is well-known that the more is clustered the spectrum, the higher is the convergence rate of a Conjugate Gradient like (CG-like) method. In the univariate case, circulant preconditioners as Strang or T. Chan preconditioners are superlinear (that is, ensure a cluster at 1 of the eigenvalues of the preconditioned matrix and the minimal eigenvalue is bounded from below by a constant independent of the size of the matrix) [38]. On the contrary, by negative results in [136, 137], the multilevel circulant preconditioning cannot ensure a superlinear convergence character of preconditioned Krylov methods. Moreover, classical circulant preconditioners fail in the case the symbol has a zero [144].

Taking into account these observations, instead of circulant preconditioners one can think to use a preconditioner which *preserves the structure* of the coefficient matrix. In this direction, an alternative technique is represented by the *band Toeplitz preconditioning* [32, 39]. Using band Toeplitz matrices, the symbol f of the coefficient matrix-sequence is approximated by a trigonometric polynomial g of fixed degree and the advantage here is that trigonometric polynomials can be chosen to match the zeros (with the same order) of f, so that the preconditioned method still works when f has zeros, that is in the ill-conditioned case. Moreover, at least in the univariate context, we recall that for solving banded systems we can apply specialized versions of the Gaussian Elimination maintaining an optimal linear cost. For optimal methods we mean methods such that the complexity of solving the given linear system is proportional to the cost of matrix-vector multiplication, see [7] for a precise notion in the context of iterative methods.

Concerning the band Toeplitz preconditioning, we emphasize that the technique has been explored for multilevel Toeplitz matrix with scalar-valued symbol in [48, 121, 123], even in the (asymptotically) ill-conditioned case, but with a specific focus on the Hermitian positive definite case. Specific attempts in the non-Hermitian case can be found in [34, 94]. Further results concerning genuine block Toeplitz structures with matrix-valued symbol are considered in [124, 126], but again for Hermitian positive definite matrix-valued symbols. In Chapter 2 (Section 2.1), analyzing the global asymptotic behavior of the spectrum of the preconditioned matrix, we enrich the literature focusing on band Toeplitz preconditioning for the non-Hermitian multilevel block case.

A different preconditioning strategy useful to deal with non-Hemitian multilevel block Toeplitz is related to the so-called Hermitian/Skew-Hermitian Splitting (HSS) scheme. When the real part of the coefficient matrix is positive definite, it is an unconditionally convergent iterative method, but when the matrix is ill-conditioned, it could show a poor convergence rate. A possibility to speed up its convergence is to use a preconditioning technique, giving rise to the Preconditioned Hermitian/Skew-Hermitian Splitting (PHSS). Choosing an Hermitian positive definite preconditioner, in Chapter 2 (Section 2.2) we perform a spectral analysis of the iteration matrix. Actually, we prove a more general spectral distribution result on sequences of matrices obtained as combination of some algebraic operations on multilevel block Toeplitz matrices. The knowledge of the eigenvalue distribution of iteration matrix will be used in practice either to provide a guess for its asymptotic spectral radius or to find efficient PHSS preconditioners.

For ill-conditioned linear systems arising from the discretization of ill-posed problems (discrete ill-posed problems [84]), such as linear systems arising in image deblurring problems, classical preconditioning may lead to wrong results. Indeed, if the preconditioner is a too close approximation of A_n, then it strongly "inherits" the ill-conditioning of A_n. In this case, the first iterations of a preconditioned CG-like method are already highly influenced by the noise of the input data, and the preconditioner gives rise to high instability, since the noise subspace has a nontrivial intersection with the subspace of ill-conditioning of the preconditioner. In order to avoid instability, the preconditioner should speed up the convergence only in the subspace where the components of the noise are negligible with respect to the components of the signal. In other words, a regularizing preconditioner has to be able to approximate A_n in the space of the signal and filters the noisy components simultaneously. Such a preconditioning is known as *regularization preconditioning*.

In Chapter 3, we propose two regularization preconditioning techniques for linear systems arising in the context of the image deblurring. On the one hand, since in image deblurring problems the structure of the coefficient matrix is a correction of a two-level Toeplitz depending on the imposed boundary conditions, with previous observations concerning the importance of preserving the structure in mind, we propose a preconditioner which has the same boundary conditions of the original problem and which shows better performances and higher stability (in relation to the choice of the regularization parameter) than the circulant preconditioner used in literature [45]. On the other hand, starting from the fact that images have sparse representation in the Fourier domain, we investigate a regularization preconditioning

technique which provides sparse solution in the Fourier domain. In detail, we use as preconditioner a diagonal matrix containing the Fourier coefficients of the observed image or of an approximation of the true image. Such a preconditioner can be interpreted as regularization matrix for Tikhonov method whose penalty term approximates the 1-norm. The resulting linear system is diagonal and hence the regularization parameter can be easily estimated, for instance by the generalized cross validation.

Aside from preconditioned Krylov methods, the *multigrid methods*, in the last decades, have gained a remarkable reputation as fast solvers for large and possibly ill-conditioned structured matrices associated to shift-invariant operators. Multigrid methods for Toeplitz matrices were firstly investigated in [71, 33, 95] and extended to multilevel Toeplitz matrices in [72, 140, 128]. The main contributions of these works were in the definition of proper grid transfer operators and in the convergence analysis of the corresponding two-grid method. The analysis of the V-cycle has been provided later in [4, 1]. As for preconditioned Krylov methods the convergence analysis of the two-grid and V-cycle can be handled by studying analytical properties of the symbol (so the study does not involve explicitly the entries of the matrix and, more importantly, the size of the system). Actually, the knowledge of the symbol is crucial to define both the symbol of the preconditioner and the grid transfer operator of a multigrid method. The advantage of multigrid methods is that for the grid transfer operator it is enough that the associated symbol possesses a proper zero with an order larger than the order of the zero of the symbol of the coefficient matrix. Conversely, the preconditioner symbol has to match exactly the order of such a zero.

In Chapter 4 both preconditioned Krylov and multigrid methods are used for solving linear systems whose coefficient matrix is a sum of two diagonal times Toeplitz and which arise from a discretization of FDE initial-boundary value problems (modeling anomalous diffusion phenomena as those encountered, e.g., in image processing [8] and in turbulent flow [31, 138]). Under appropriate conditions the coefficient matrix-sequence belongs to the GLT class. We compute the associated symbol and show that when the diffusion coefficients are equal (even if not necessarily constant), it also describes the eigenvalue distribution. Making use of such asymptotic spectral information, we study in more detail known methods of preconditioned Krylov and multigrid type for FDEs problems: for instance we prove that the circulant preconditioning described in [100] cannot be superlinear in the variable coefficient case, due to a lack of clustering at a single point, while the multigrid approach based on the symbol used in [113] can be optimal also in the variable coefficient setting. Moreover, we introduce two structure preserving tridiagonal preconditioners for Krylov methods, which preserve the computational cost per iteration of the used Krylov method.

A multigrid method for Toeplitz matrices with univariate block symbol has been proposed in [96], even though, up to our knowledge, when the block symbol is not diagonal it lacks of a convergence analysis yet. On the other hand, the block symbol is becoming a popular theoretical tool [87] and relevant applications are related to block (multilevel) Toeplitz matrices, e.g. [58]. In Chapter 5, we formally extend the idea presented in [96] also in the multidimensional setting and to nonconstant basis of eigenvectors. We derive a very efficient multigrid preconditioner for GLT sequences originating from Finite Element Method (FEM) discretization of a coupled system of PDEs. We illustrate the technique on a two dimensional linear elasticity problem, discretized by the stable FEM pair Q1isoQ1, and arising as a subproblem of the Glacial Isostatic Adjustment [101].

We now summarize the contents of Chapters 1–5.

- In Chapter 1 we introduce definitions and results useful throughout the thesis. Among others, the definition of distribution in the eigen/singular value sense, the clustering property of matrix-sequences, the notion of multilevel block Toeplitz matrices, the GLT class. Furthermore, an overview of preconditioning for Toeplitz linear systems and some introductory aspects of the HSS scheme and of multigrid methods as solvers for such linear systems are also given.

- Chapter 2 is devoted to preconditioning strategies for non-Hermitian multilevel block Toeplitz linear systems associated with a multivariate Lebesgue integrable matrix-valued function. On the one hand, we consider preconditioned matrices, where the preconditioner has a band multilevel block Toeplitz structure and we complement known results on the localization of the spectrum with global distribution results for the eigenvalues of the preconditioned matrices. On the other hand, we perform a spectral analysis of the PHSS method applied to multilevel block Toeplitz linear systems. When the preconditioner is chosen as a Hermitian positive definite multilevel block Toeplitz matrix, we are able to compute the symbol describing the asymptotic eigenvalue distribution of the iteration matrices, and, by minimizing the infinity norm of the spectral radius of the symbol, we are also able to identify effective PHSS preconditioners for the system matrix.

- In Chapter 3 we discuss two regularization preconditioning strategies for structured matrices arising in the context of the image deblurring problem. The first technique is a structure preserving preconditioning aimed to accelerate the convergence of iterative regularization methods, without spoiling the restoration. The second one consists in a diagonal regularization preconditioner containing the Fourier coefficients of the observed image or of an approximation of the true image, which is interpreted as a regularization matrix for the Tikhonov regularization in the Fourier domain.

- In Chapter 4 we focus on the sequence of linear systems arising in the discretization of FDEs, with particular attention to the nonconstant coefficient case. We show that the coefficient matrix-sequence belongs to the GLT class and then we compute the associated symbol. Making use of such asymptotic spectral information, we study in more detail recently developed techniques, by furnishing new positive and negative results. Moreover, we introduce two tridiagonal preconditioners for Krylov methods obtained approximating the matrix corresponding to the fractional derivative operator with the discretization matrix of the first derivative and with the Laplacian matrix. By construction, both preconditioners preserve the Toeplitz-like structure of the coefficient matrix and, due to their tridiagonal structure, both preserve the computational cost per iteration of the used Krylov method. A clustering analysis of the preconditioned matrix-sequences, even in case of nonconstant diffusion coefficients, is also provided.

- In Chapter 5 we consider the solution of linear systems arising from the finite element approximation of coupled PDEs. As an example we consider the linear elasticity problem in saddle point form. Discretizing by the stable FEM pair Q1isoQ1, we obtain a linear system with a two-by-two block matrix. We are interested in the efficient iterative solution of the involved linear systems, aiming at constructing optimal preconditioning methods that are robust with respect to the relevant parameters of the problem. We consider the case when the originating systems are solved by a preconditioned Krylov method, as inner solver, and performing a spectral analysis of two-level block Toeplitz structure of the arising block matrices, we design an ad hoc multigrid method to be used as preconditioner.

All our principal findings are summarized in the conclusion chapter.

The results of our research have been published or are in the process of publication in [53, 54, 56, 57, 46, 51]

Chapter 1

Preliminary definitions and results

The aim of this introductory chapter is to fix the notation used throughout the thesis, to illustrate some known tools necessary for dealing with sequences of (Toeplitz) matrices, and to recall some classical solving strategies for (Toeplitz) linear systems. In particular, we define the spectral distribution and the clustering property of the spectrum of a sequence of matrices. Moreover, we introduce the definition of multilevel block Toeplitz matrices and briefly recall some known localization and spectral distribution results, focusing both in the Hermitian and non-Hermitian matrix-valued symbol case, with particular attention to preconditioned sequences of multilevel block Toeplitz matrices. Furthermore, due to extensively use that we will make of, a section of this chapter is devoted to the Generalized Locally Toeplitz algebra which extends the multilevel block Toeplitz one. An overview of preconditioning for Toeplitz linear systems and some introductory aspects of the Hermitian/Skew-Hermitian splitting scheme and of multigrid methods as solvers for such linear systems are also given. We conclude the chapter recalling some properties of trigonometric polynomials.

1.1 Notation and norms

- $\mathcal{M}_n$ is the linear space of the complex $n \times n$ matrices.
- $\mathcal{C}_0(\mathbb{C})$ and $\mathcal{C}_0(\mathbb{R}_0^+)$ are the set of continuous functions with bounded support defined over $\mathbb{C}$ and $\mathbb{R}_0^+ = [0, \infty)$, respectively.
- Given a matrix $A \in \mathcal{M}_n$, we denote by
  - $\lambda_j(A),\ j = 1, \ldots, n$ the eigenvalues of A and by $\sigma_j(A),\ j = 1, \ldots, n$ its singular values;
  - $\Lambda(A)$ the diagonal matrix containing the eigenvalues of A;
  - $\rho(A) = \max_{i=1,\ldots,n} |\lambda_i(A)|$ the spectral radius of A;
  - $\mathrm{tr}(A)$ the trace of A;
  - A^T the transpose of A;
  - A^* the conjugate transpose of A;
  - $\mathrm{rank}(A)$ the rank of A;
  - $\mathcal{N}(A)$ the null space of the matrix A.
- Whenever A is a Hermitian positive semidefinite matrix, $A^{1/2}$ is the nonnegative square root of A.
- Given two matrices $A, B \in \mathcal{M}_n$, $A \sim B$ means that A is similar to B.
- If $r \in \mathbb{C}$ and $\epsilon > 0$, then $D(r, \epsilon)$ is the open disk in the complex plane centered at r with radius ϵ.
- We denote by $D(S, \epsilon)$ the ϵ-expansion of S, defined as $D(S, \epsilon) = \bigcup_{r \in S} D(r, \epsilon)$.
- For any $X \subseteq \mathbb{C}$, $\mathrm{Coh}[X]$ is the convex hull of X and $d(X, z)$ is the (Euclidean) distance of X from the point $z \in \mathbb{C}$.
- We say that a function $f : G \to \mathcal{M}_s$, defined on some measurable set $G \subseteq \mathbb{R}^k$, is in $L^p(G)$/measurable/continuous, if its components $f_{ij} : G \to \mathbb{C},\ i, j = 1, \ldots, s$, are in $L^p(G)$/measurable/continuous.

- I_k is the k-cube $(-\pi,\pi)^k$.
- m_k is the Lebesgue measure in $\mathbb{R}^k$.
- For $1 \le p \le \infty$, $L^p(k,s)$ is the linear space of k-variate functions $f : I_k \to \mathcal{M}_s$ belonging to $L^p(I_k)$.
- $\mathbf{i}$ is the imaginary unit ($\mathbf{i}^2 = -1$).
- $\otimes$ denotes the Kronecker tensor product.
- A sequence of matrices parameterized by an index $n \in \mathbb{N}$ will be denoted by $\{A_n\}_n$ and will be called 'matrix-sequence'.
- 'HPD' is an abbreviation for Hermitian Positive Definite, while 'HPSD' is an abbreviation for Hermitian Positive SemiDefinite.
- 'BCCB' is an abbreviation for Block Circulant with Circulant Blocks, while 'BTTB' is an abbreviation for Block Toeplitz with Toeplitz Blocks.
- 'FFT' is an abbreviation for Fast Fourier Transform, 'DFT' is an abbreviation for Discrete Fourier Transform, and 'DCT' is an abbreviation for Discrete Cosine Transform.
- 'CG' is an abbreviation for Conjugate Gradient, 'CGLS' is an abbreviation for Conjugate Gradient for Least Squares, and 'GMRES' is an abbreviation for Generalized Minimal Residual.

1.1.1 Multi-index notation

Let us fix the multi-index notation that will be extensively used throughout this thesis. A multi-index $m \in \mathbb{Z}^k$, also called a k-index, is simply a vector in $\mathbb{Z}^k$ and its components are denoted by $m_1,\dots,m_k$. Standard operations defined for vectors in $\mathbb{C}^k$, such as addition, subtraction and scalar multiplication, are also defined for k-indices. We will use the letter e for the vector of all ones, whose size will be clear from the context. If $m \in \mathbb{N}^k$, we set $\widehat{m} := \prod_{i=1}^k m_i$ and we write $m \to \infty$ to indicate that $\min_i m_i \to \infty$. Inequalities involving multi-indices are always understood in the componentwise sense. For instance, given $h, m \in \mathbb{Z}^k$, the inequality $h \le m$ means that $h_l \le m_l$ for all $l = 1,\dots,k$. If $h, m \in \mathbb{Z}^k$ and $h \le m$, the multi-index range $h,\dots,m$ is the set $\{j \in \mathbb{Z}^k : h \le j \le m\}$. We assume for the multi-index range $h,\dots,m$ the standard lexicographic ordering:

$$\Big[\dots \Big[\big[(j_1,\dots,j_k) \big]_{j_k=h_k,\dots,m_k} \Big]_{j_{k-1}=h_{k-1},\dots,m_{k-1}} \dots \Big]_{j_1=h_1,\dots,m_1}. \tag{1.1}$$

For instance, if $k = 2$ then the ordering is

$$(h_1,h_2),(h_1,h_2{+}1),\dots,(h_1,m_2),(h_1{+}1,h_2),(h_1{+}1,h_2{+}1),\dots,(h_1{+}1,m_2),\dots\dots,(m_1,h_2),(m_1,h_2{+}1),\dots,(m_1,m_2).$$

The notation $\sum_{j=h}^m$ indicates the summation over all j in the multi-index range $h,\dots,m$. When a multi-index j varies over a multi-index range $h,\dots,m$ (this may be written as $j = h,\dots,m$), it is always understood that j varies from h to m according to the lexicographic ordering (1.1). For instance, if $m \in \mathbb{N}^k$ and if $y = [y_i]_{i=e}^m$, then y is a vector of size $m_1 \cdots m_k$ whose components y_i, $i = e,\dots,m$, are ordered in accordance with the ordering (1.1) for the multi-index range $e,\dots,m$. Similarly, if $Y = [y_{ij}]_{i,j=e}^m$, then Y is a matrix of size $m_1 \cdots m_k$ whose components are indexed by two multi-indices i, j, both varying over the multi-index range $e,\dots,m$ in accordance with (1.1).

In order to highlight the multi-index notation, throughout the thesis, a sequence of matrices parameterized by a multi-index $n \in \mathbb{N}^k$ will be denoted by $\{A_n\}_{n\in\mathbb{N}^k}$. Moreover, we will refer to it as 'matrix-family'.

1.1.2 Vector, matrix and functional norms

Throughout this thesis we give to $\|\cdot\|_p$ two meanings depending on whether it is applied to a vector or to a matrix. Given a vector $x \in \mathbb{C}^n$, we denote by $\|x\|_p$, $p \in [1,\infty]$ the p-norm of x defined as

$$\|x\|_p = \left(\sum_{i=1}^n |x_i|^p\right)^{\frac{1}{p}}, \qquad p \in [1,+\infty),$$
$$\|x\|_\infty = \max_{i=1,\dots,n} |x_i| = \|x\|, \qquad p = \infty.$$

Given a matrix $A \in \mathcal{M}_n$, we denote by $\|A\|_p$, $p \in [1,\infty]$ the Schatten p-norm of A, defined as the p-norm of the vector formed by the singular values of A. In symbols,

$$\|A\|_p = \left(\sum_{j=1}^{n} \sigma_j^p(A)\right)^{1/p}, \qquad p \in [1,+\infty),$$
$$\|A\|_\infty = \max_{j=1,\dots,n} \sigma_j(A) = \|A\|, \qquad p = \infty.$$

The Schatten 1-norm is also called the trace-norm, while $\|\cdot\|_\infty = \|\cdot\|$ is known as spectral norm. We refer the reader to [19] for the properties of the Schatten p-norms. We only recall from [19, Problem III.6.2 and Corollary IV.2.6] the Hölder inequality $\|AB\|_1 \le \|A\|_p \|B\|_q$, which is true for all square matrices A, B of the same size and whenever $p, q \in [1,\infty]$ are conjugate exponents (i.e. $\frac{1}{p} + \frac{1}{q} = 1$). In particular, we will need the Hölder inequality with $p = 1$ and $q = \infty$, which involves the spectral norm and the trace-norm:

$$\|AB\|_1 \le \|A\| \|B\|_1. \tag{1.2}$$

Other well-known inequalities involving the Schatten 1-norm are the following

$$|\mathrm{tr}\,(A)| \le \|A\|_1, \tag{1.3}$$
$$\|A\|_1 \le \mathrm{rank}(A)\|A\|. \tag{1.4}$$

The p-norm of a square matrix $A \in \mathcal{M}_n$, that is the p-norm of the vector of length n^2 obtained by putting all the columns of A one below the other will be denoted by $\|A\|_{(p)}$, $p \in [1,\infty]$

$$\|A\|_{(p)} = \left(\sum_{i=1}^{n}\sum_{j=1}^{n} |a_{ij}|^p\right)^{\frac{1}{p}}, \qquad p \in [1,+\infty),$$
$$\|A\|_{(\infty)} = \max_{i,j=1,\dots,n} |a_{ij}|, \qquad p = \infty.$$

Actually, $\|\cdot\|_{(p)}$ is not submultiplicative and so, according to some books, it is not a matrix norm. However, we will keep on calling it a norm, both for simplicity and for the fact that it possesses all the other properties of matrix norms. Note that the famous Frobenius norm is just $\|\cdot\|_{(p)}$ for $p = 2$. For historical reasons we will denote it by $\|\cdot\|_F$.

Now, in order to define a functional norm on $L^p(k,s)$ by means of the Schatten p-norm, let us show that, for any $1 \le p \le \infty$, $L^p(k,s) = L^p(I_k, dx, \mathcal{M}_s)$, where

$$L^p(I_k, dx, \mathcal{M}_s) := \left\{ f : I_k \to \mathcal{M}_s \;\middle|\; f \text{ is measurable, } \int_{I_k} \|f(x)\|_p^p dx < \infty \right\}, \qquad \text{if } 1 \le p < \infty,$$
$$L^\infty(I_k, dx, \mathcal{M}_s) := \left\{ f : I_k \to \mathcal{M}_s \;\middle|\; f \text{ is measurable, } \operatorname*{ess\,sup}_{x\in I_k} \|f(x)\|_\infty < \infty \right\}.$$

Since $\mathcal{M}_s$ is a finite-dimensional vector space, all the norms on $\mathcal{M}_s$ are equivalent. In particular, $\|\cdot\|_{(p)}$ and $\|\cdot\|_p$ are equivalent, and so there are two positive constants α, β such that

$$\alpha\|f(x)\|_p \le \|f(x)\|_{(p)} \le \beta\|f(x)\|_p, \quad \forall x \in I_k.$$

It follows that

$$\alpha^p \int_{I_k} \|f(x)\|_p^p dx \le \int_{I_k} \|f(x)\|_{(p)}^p dx \le \beta^p \int_{I_k} \|f(x)\|_p^p dx, \qquad \text{if } 1 \le p < \infty, \tag{1.5}$$
$$\alpha \operatorname*{ess\,sup}_{x\in I_k} \|f(x)\|_\infty \le \operatorname*{ess\,sup}_{x\in I_k} \|f(x)\|_{(\infty)} \le \beta \operatorname*{ess\,sup}_{x\in I_k} \|f(x)\|_\infty. \tag{1.6}$$

Therefore, if $f \in L^p(k,s)$ then each component $f_{ij} : I_k \to \mathbb{C}$, $i,j = 1,\dots,s$, belongs to $L^p(I_k)$ and the first inequalities in (1.5)–(1.6) say that $f \in L^p(I_k, dx, \mathcal{M}_s)$. Conversely, if $f \in L^p(I_k, dx, \mathcal{M}_s)$, the second inequalities in (1.5)–(1.6) says that $f \in L^p(k,s)$. This concludes the proof of the identity $L^p(k,s) = L^p(I_k, dx, \mathcal{M}_s)$ and allows us to define the following functional norm on $L^p(k,s)$:

$$\|f\|_{L^p} := \| \|f(x)\|_p \|_{L^p(I_k)} = \begin{cases} \left(\int_{I_k} \|f(x)\|_p^p dx\right)^{1/p}, & \text{if } 1 \le p < \infty, \\ \operatorname{ess\,sup}_{x\in I_k} \|f(x)\|_\infty, & \text{if } p = \infty. \end{cases} \tag{1.7}$$

If $p, q \in [1, \infty]$ are conjugate exponents and $f \in L^p(k,s),\ g \in L^q(k,s)$, then a computation involving the Hölder inequalities for both Schatten p-norms and $L^p(I_k)$-norms shows that $fg \in L^1(k,s)$ and, in fact, $\|fg\|_{L^1}, \|gf\|_{L^1} \leq \|f\|_{L^p}\|g\|_{L^q}$. In particular, we will need the inequality with $p = 1$ and $q = \infty$, i.e.

$$\|fg\|_{L^1}, \|gf\|_{L^1} \leq \|f\|_{L^1}\|g\|_{L^\infty}. \tag{1.8}$$

1.2 Essential range and sectorial functions

In this section we introduce the notion of essential range, essential numerical range and sectoriality of a matrix-valued function f.

If $f : G \to \mathbb{C}$ is a complex-valued measurable function, defined on some measurable set $G \subseteq \mathbb{R}^k$, the essential range of f, $\mathcal{ER}(f)$, is defined as the set of points $r \in \mathbb{C}$ such that, for every $\epsilon > 0$, the measure of $f^{-1}(D(r,\epsilon)) := \{t \in G : f(t) \in D(r,\epsilon)\}$ is positive. In symbols,

$$\mathcal{ER}(f) := \{r \in \mathbb{C} : \forall \epsilon > 0,\ m_k\{t \in G : f(t) \in D(r,\epsilon)\} > 0\}.$$

Note that $\mathcal{ER}(f)$ is always closed (the complement is open). Moreover, it can be shown that $f(t) \in \mathcal{ER}(f)$ for almost every $t \in G$, i.e., $f \in \mathcal{ER}(f)$ a.e.

Definition 1. Given a measurable matrix-valued function $f : G \to \mathcal{M}_s$, defined on some measurable set $G \subseteq \mathbb{R}^k$,

- the *essential range* of f, $\mathcal{ER}(f)$, is the union of the essential ranges of the eigenvalue functions $\lambda_j(f) : G \to \mathbb{C}$, $j = 1, \ldots, s$, that is $\mathcal{ER}(f) := \bigcup_{j=1}^{s} \mathcal{ER}(\lambda_j(f))$;
- the *essential numerical range* of f, $\mathcal{ENR}(f)$, is the set of points $r \in \mathbb{C}$ such that, for every $\epsilon > 0$, the measure of $\{t \in G : \exists v \in \mathbb{C}^s$ with $\|v\|_2 = 1$ such that $v^* f(t) v \in D(r,\epsilon)\}$ is positive. In symbols,

  $$\mathcal{ENR}(f) := \{r \in \mathbb{C} : \forall \epsilon > 0,\ m_k\{t \in G : \exists v \in \mathbb{C}^s \text{ with } \|v\|_2 = 1 \text{ such that } v^* f(t) v \in D(r,\epsilon)\} > 0\}.$$

Note that $\mathcal{ER}(f)$ is closed, being the union of a finite number of closed sets. $\mathcal{ENR}(f)$ is also closed, because its complement is open. Moreover, it can be proved that, for a.e. $t \in G$, the following property holds: $v^* f(t) v \in \mathcal{ENR}(f)$ for all $v \in \mathbb{C}^s$ with $\|v\|_2 = 1$. In other words, $v^* f v \in \mathcal{ENR}(f)$ for all $v \in \mathbb{C}^s$ with $\|v\|_2 = 1$, a.e. In addition, it can be shown that $\mathcal{ENR}(f) \supseteq \mathcal{ER}(f)$. In the case $s = 1$, we have $\mathcal{ENR}(f) = \mathcal{ER}(f)$.

Now we turn to the definition of sectorial function. Given a straight line z in the complex plane, let H_1 and H_2 be the two open half-planes such that $\mathbb{C}$ is the disjoint union $H_1 \bigcup z \bigcup H_2$; we call H_1 and H_2 the open half-planes determined by z. Moreover, we denote by $\omega(z) \in \mathbb{C}$ the rotation number (of modulus 1) such that $\omega(z) \cdot z = \{w \in \mathbb{C} : \mathrm{Re}(w) = d(z,0)\}$. Note that $\omega(z)$ is uniquely determined if $d(z,0) > 0$. If $d(z,0) = 0$, there are two possible candidates for $\omega(z)$, one the opposite of the other. This is not really a problem for what follows, but, for definiteness, let us say that, in the case where $d(z,0) = 0$, $\omega(z)$ is the candidate lying in the half-plane $\{w \in \mathbb{C} :\ \mathrm{Re}(w) > 0\} \bigcup \{w \in \mathbb{C} :\ \mathrm{Re}(w) = 0,\ \mathrm{Im}(w) > 0\}$.

Definition 2. A function $f \in L^1(k,s)$ is *weakly sectorial* if there exists a straight line z in the complex plane with the following property: one of the two open half-planes determined by z, say H_1, is such that $\mathcal{ENR}(f) \bigcap H_1 = \emptyset$ and $0 \in H_1 \bigcup z$. Whenever $f \in L^1(k,s)$ is weakly sectorial, every straight line z with the previous property is called a *separating line* for $\mathcal{ENR}(f)$. A function $f \in L^1(k,s)$ is *sectorial* if it is weakly sectorial and there exists a separating line z such that the eigenvalue of minimum modulus of $\frac{1}{2}(\omega(z) f(x) + \overline{\omega(z)} f^*(x))$ is not a.e. equal to $d(z,0)$.

Remark 1. Given a weakly sectorial function $f \in L^1(k,s)$ and a separating line z for $\mathcal{ENR}(f)$, the following relation holds:

$$\mathcal{ENR}(\omega_z f) = \omega_z \cdot \mathcal{ENR}(f) \subseteq \{w \in \mathbb{C} :\ \mathrm{Re}(w) \geq d(z,0)\}, \tag{1.9}$$

where ω_z is either $\omega(z)$ or $-\omega(z)$ (for sure, $\omega_z = \omega(z)$ if $d(z,0) > 0$). Recalling from the discussion after Definition 1 that, a.e., $v^*[\omega_z f] v \in \mathcal{ENR}(\omega_z f)$ for all $v \in \mathbb{C}^s$ with $\|v\|_2 = 1$, from (1.9) it follows that, a.e.,

$$v^* \mathrm{Re}(\omega_z f) v = \mathrm{Re}(v^*[\omega_z f] v) \geq d(z,0) \qquad \forall v \in \mathbb{C}^s \text{ with } \|v\|_2 = 1.$$

This implies, by the minimax principle [19], that

$$\lambda_{\min}(\mathrm{Re}(\omega_z f)) \geq d(z,0) \quad \text{a.e.} \tag{1.10}$$

The sectoriality requirement can then be expressed in an equivalent way by saying that there exists a certain separating line z for which the minimal eigenvalue of the real part $\mathrm{Re}(\omega_z f) = \frac{1}{2}(\omega_z f + \overline{\omega_z} f^*)$, which is a.e. greater or equal to $d(z,0)$, is not a.e. equal to $d(z,0)$.

1.3 Spectral distribution and clustering of matrix-sequences

In this section we begin with the definition of spectral distribution and clustering, in the sense of eigenvalues and singular values, of a matrix-sequence, and we define the area of K, in the case where K is a compact subset of $\mathbb{C}$. Then, we present the main tool, taken from [60], for proving our distribution results presented in Chapter 2, i.e. Theorems 13,15, which provide the asymptotic spectral distribution of preconditioned multilevel block Toeplitz matrices and of some algebraic combinations of multilevel block Toeplitz matrices.

Given a function F and given a matrix A of order M, we set

$$\Sigma_\lambda(F,A) := \frac{1}{M}\sum_{j=1}^{M} F(\lambda_j(A)), \qquad \Sigma_\sigma(F,A) := \frac{1}{M}\sum_{j=1}^{M} F(\sigma_j(A)).$$

Definition 3. Let $f : G \to \mathcal{M}_s$ be a measurable function, defined on a measurable set $G \subset \mathbb{R}^k$ with $0 < m_k(G) < \infty$. Let $\{A_n\}_n$ be a matrix-sequence, with A_n of size d_n tending to infinity.

- $\{A_n\}_n$ is *distributed as the pair* (f,G) *in the sense of the eigenvalues,* in symbols $\{A_n\}_n \sim_\lambda (f,G)$, if for all $F \in \mathcal{C}_0(\mathbb{C})$ we have

$$\lim_{n\to\infty} \Sigma_\lambda(F,A_n) = \frac{1}{m_k(G)} \int_G \frac{\sum_{i=1}^{s} F(\lambda_i(f(t)))}{s} dt = \frac{1}{m_k(G)} \int_G \frac{\mathrm{tr}(F(f(t)))}{s} dt. \tag{1.11}$$

  In this case, we say that f is the *symbol* of the matrix-sequence $\{A_n\}_n$.

- $\{A_n\}_n$ is *distributed as the pair* (f,G) *in the sense of the singular values,* in symbols $\{A_n\}_n \sim_\sigma (f,G)$, if for all $F \in \mathcal{C}_0(\mathbb{R}_0^+)$ we have

$$\lim_{n\to\infty} \Sigma_\sigma(F,A_n) = \frac{1}{m_k(G)} \int_G \frac{\sum_{i=1}^{s} F(\sigma_i(f(t)))}{s} dt = \frac{1}{m_k(G)} \int_G \frac{\mathrm{tr}(F(|f(t)|))}{s} dt, \tag{1.12}$$

  where $|f(t)| := (f^*(t)f(t))^{1/2}$.

Remark 2. If f is smooth enough, an informal interpretation of the limit relation (1.11) (resp. (1.12)) is that when the matrix-size of A_n is sufficiently large, then d_n/s eigenvalues (resp. singular values) of A_n can be approximated by a sampling of $\lambda_1(f(t))$ (resp. $\sigma_1(f(t))$) on a uniform equispaced grid of the domain G, and so on until the last d_n/s eigenvalues can be approximated by an equispaced sampling of $\lambda_s(f(t))$ (resp. $\sigma_s(f(t))$) in the domain. For instance, if f is continuous, $k = 1$, $d_n = ns$, and $G = [a,b]$, the eigenvalues of A_n are approximately equal to $\lambda_i(f(a + j\frac{b-a}{n}))$, $j = 1,\ldots,n$, $i = 1,\ldots,s$. Analogously, if $k = 2$, $d_n = n^2 s$, and $G = [a,b]\times[c,d]$, the eigenvalues of A_n are approximately equal to $\lambda_i(f(a + j\frac{b-a}{n}, c + l\frac{d-c}{n}))$, $j,l = 1,\ldots,n$, $i = 1,\ldots,s$ (and so on in a k-variate setting).

If $\{A_n\}_{n\in\mathbb{N}^h}$ is a matrix-family (parameterized by a multi-index), with A_n of size d_n tending to infinity when $n \to \infty$ (i.e. when $\min_j n_j \to \infty$), we still write $\{A_n\}_{n\in\mathbb{N}^h} \sim_\lambda (f,G)$ to indicate that (1.11) is satisfied for all $F \in \mathcal{C}_0(\mathbb{C})$, but we point out that now '$n \to \infty$' in (1.11) means '$\min_j n_j \to \infty$', in accordance with the multi-index notation introduced before. Similarly, we write $\{A_n\}_{n\in\mathbb{N}^h} \sim_\sigma (f,G)$ if (1.12) is satisfied for all $F \in \mathcal{C}_0(\mathbb{R}_0^+)$, where again $n \to \infty$ means $\min_j n_j \to \infty$. We note that $\{A_n\}_{n\in\mathbb{N}^h} \sim_\lambda (f,G)$ (resp. $\{A_n\}_{n\in\mathbb{N}^h} \sim_\sigma (f,G)$) is equivalent to saying that $\{A_{n(m)}\}_m \sim_\lambda (f,G)$ (resp. $\{A_{n(m)}\}_m \sim_\sigma (f,G)$) for every matrix-sequence $\{A_{n(m)}\}_m$ extracted from $\{A_n\}_{n\in\mathbb{N}^h}$ and such that $\min_j n_j(m) \to \infty$ as $m \to \infty$.

Definition 4. Let $\{A_n\}_n$ be a matrix-sequence, with A_n of size d_n tending to infinity, and let $S \subseteq \mathbb{C}$ be a closed subset of $\mathbb{C}$. We say that $\{A_n\}_n$ *is strongly clustered at S in the sense of the eigenvalues* if, for every $\epsilon > 0$, the number of eigenvalues of A_n outside $D(S, \epsilon)$ (the ϵ-expansion of S) is bounded by a constant q_ϵ independent of n. In other words

$$q_\epsilon(n, S) := \#\{j \in \{1, \ldots, d_n\} : \lambda_j(A_n) \notin D(S, \epsilon)\} = O(1), \quad \text{as } n \to \infty. \tag{1.13}$$

We say that $\{A_n\}_n$ *is weakly clustered at S in the sense of the eigenvalues* if, for every $\epsilon > 0$,

$$q_\epsilon(n, S) = o(n), \quad \text{as } n \to \infty.$$

If $\{A_n\}_n$ is strongly or weakly clustered at S and S is not connected, then its disjoint parts are called sub-clusters.

By replacing 'eigenvalues' with 'singular values' and $\lambda_j(A_n)$ with $\sigma_j(A_n)$ in (1.13), we obtain the definitions of a matrix-sequence strongly or weakly clustered at a closed subset of $\mathbb{C}$, in the sense of the singular values. When speaking of strong/weak clustering, matrix-sequence strongly/weakly clustered, etc., without further specifications, it is always understood 'in the sense of the eigenvalues' (when the clustering is intended in the sense of the singular values, this is always specified). It is worth noting that, since the singular values are always nonnegative, any matrix-sequence is strongly clustered in the sense of the singular values at a certain $S \subseteq [0, \infty)$. Similarly, any matrix-sequence formed by matrices with only real eigenvalues (e.g. by Hermitian matrices) is strongly clustered at some $S \subseteq \mathbb{R}$ in the sense of the eigenvalues.

Remark 3. If $\{A_n\}_n \sim_\lambda (f, G)$, with $\{A_n\}_n, f, G$ as in Definition 3, then $\{A_n\}_n$ is weakly clustered at $\mathcal{ER}(f)$ in the sense of the eigenvalues. This result is proved in [77, Theorem 4.2]. It is clear that $\{A_n\}_n \sim_\lambda (f, G)$, with $f \equiv r$ equal to a constant function, is equivalent to saying that $\{A_n\}_n$ is weakly clustered at $r \in \mathbb{C}$ in the sense of the eigenvalues. The reader is referred to [127, Section 4] for several relationships which link the concepts of equal distribution, equal localization, spectral distribution, spectral clustering, etc.

Definition 5. Let K be a compact subset of $\mathbb{C}$. We define

$$Area(K) := \mathbb{C} \backslash U,$$

where U is the (unique) unbounded connected component of $\mathbb{C} \backslash K$.

Now we are ready for stating the main tool we shall use for the proof of our distribution results in Chapter 2 (Theorems 13, 15).

Theorem 1. [60] *Let $\{A_n\}_n$ be a matrix-sequence, with A_n of size d_n tending to infinity as $n \to \infty$. If:*

(c_1) *the spectrum of A_n, $\Lambda(A_n)$, is uniformly bounded, i.e., $|\lambda| < C$ for all $\lambda \in \Lambda(A_n)$, for all n, and for some constant C independent of n;*

(c_2) *there exists a measurable function $f : G \to \mathcal{M}_s$ in $L^\infty(k, s)$, defined over a certain domain $G \subset \mathbb{R}^k$ of finite and positive Lebesgue measure, such that, for every positive integer N, we have*

$$\lim_{n \to \infty} \frac{\mathrm{tr}(A_n^N)}{d_n} = \frac{1}{m_k(G)} \int_G \frac{\mathrm{tr}(f^N(x))}{s} \, dx;$$

(c_3) *$\{P(A_n)\}_n \sim_\sigma (P(f), G)$ for every polynomial P;*

then the matrix-sequence $\{A_n\}_n$ is weakly clustered at $Area(\mathcal{ER}(f))$ in the sense of the eigenvalues and relation (1.11) *is true for every $F \in \mathcal{C}_0(\mathbb{C})$ which is holomorphic in the interior of $Area(\mathcal{ER}(f))$. If moreover:*

(c_4) *$\mathbb{C} \backslash \mathcal{ER}(f)$ is connected and the interior of $\mathcal{ER}(f)$ is empty;*

then $\{A_n\}_n \sim_\lambda (f, G)$.

In the previous theorem, the assumption **(c_4)** on the function f only concerns the topological structure of its essential range, and is completely independent of its smoothness properties. We call the set of functions f satisfying **(c_4)** the 'Tilli class'. Here is the precise definition.

Definition 6. [143] We say that a measurable function $f : G \to \mathcal{M}_s$, defined on some measurable set $G \subseteq \mathbb{R}^k$, is in the *Tilli class*, if $\mathbb{C}\backslash\mathcal{ER}(f)$ is connected and the interior of $\mathcal{ER}(f)$ is empty.

We end this section by providing another useful theorem for proving the asymptotic spectral distribution of a matrix-sequence.

Theorem 2 (Theorem 3.4 in [77])**.** *Let $\{A_n\}_n$ be a matrix-sequence with $A_n = B_n + C_n$ and B_n Hermitian. Assume that A_n, B_n and C_n are of size d_n tending to infinity. If*

- *$\{B_n\}_n \sim_\lambda (f, G)$, where $f : G \to \mathbb{C}$ is defined on a measurable set $G \subset \mathbb{R}^k$ with $0 < m_k(G) < \infty$;*
- *$\|B_n\|$, $\|C_n\|$ are bounded by a constant independent of n;*
- *$\|C_n\|_1 = o(d_n)$.*

Then $\{A_n\}_n \sim_\lambda (f, G)$.

1.4 Multilevel block Toeplitz matrices

We begin this section by introducing the definition of multilevel block Toeplitz matrix. Then, we recall some localization and spectral distribution results focusing both in the Hermitian and non-Hermitian matrix-valued symbol case, with particular attention to preconditioned sequences of multilevel block Toeplitz matrices. Recall that if $n := (n_1, \ldots, n_k)$ is a multi-index in $\mathbb{N}^k$, we define $\widehat{n} := \prod_{i=1}^k n_i$.

Definition 7. Fixed $f \in L^1(k, s)$, the Fourier coefficients of f are defined as

$$\widehat{f}_j := \frac{1}{(2\pi)^k} \int_{I_k} f(x) \mathrm{e}^{-\mathbf{i}\langle j,x\rangle} dx \in \mathcal{M}_s, \quad j \in \mathbb{Z}^k, \tag{1.14}$$

where $\langle j, x\rangle = \sum_{t=1}^k j_t x_t$, and the integrals in (1.14) are done componentwise. Then the n-th Toeplitz matrix associated with f is the matrix of order $s\widehat{n}$ given by

$$T_n(f) = \sum_{|j_1|<n_1} \cdots \sum_{|j_k|<n_k} \left[J_{n_1}^{(j_1)} \otimes \cdots \otimes J_{n_k}^{(j_k)} \right] \otimes \widehat{f}_{(j_1,\ldots,j_k)}, \tag{1.15}$$

where $J_t^{(l)}$ is the matrix of order t whose (i, j) entry equals 1 if $i - j = l$ and equals zero otherwise. $\{T_n(f)\}_{n\in\mathbb{N}^k}$ is called the *family of Toeplitz matrices generated by f*, which in turn is called the *symbol (or the generating function)* of $\{T_n(f)\}_{n\in\mathbb{N}^k}$.

Multilevel block Toeplitz matrices arise in important applications such as Markov chains [43, 107] (with $k = 1$ and $s > 1$), in the reconstruction of signals with missing data [44] (with $k = 1$ and $s = 2$), in the inpainting problem [28] (with $k = 2$ and $s = 2$), and of course in the numerical approximation of constant coefficient $r \times r$ systems of Partial Differential Equations (PDEs) over d-dimensional domains [5] (with $k = d$ and $s = r$).

We point out that the multilevel block Toeplitz matrix $T_n(f)$ displayed in (1.15) can be expressed in multi-index notation as

$$T_n(f) = [\widehat{f}_{i-j}]_{i,j=e}^n. \tag{1.16}$$

Moreover, we recall some known facts concerning the spectral norm and the Schatten 1-norm of Toeplitz matrices, see [130, Corollary 3.5]:

$$f \in L^1(k, s) \quad \Rightarrow \quad \|T_n(f)\|_1 \le \frac{\widehat{n}}{(2\pi)^k} \|f\|_{L^1}, \quad \forall n \in \mathbb{N}^k; \tag{1.17}$$

$$f \in L^\infty(k, s) \quad \Rightarrow \quad \|T_n(f)\| \le \|f\|_{L^\infty}, \quad \forall n \in \mathbb{N}^k. \tag{1.18}$$

1.4.1 Localization results

We start this subsection with a localization result for the spectrum of $T_n(f)$ in the case where $f : I_k \to \mathcal{M}_s$ is a Hermitian matrix-valued function, i.e., $f(x)$ is Hermitian for a.e. $x \in I_k$ (Proposition 1). Next, we state Proposition 2 which provides a localization result for the spectrum of the preconditioned matrix $T_n^{-1}(g)T_n(f)$ in the case where f, g are Hermitian matrix-valued functions with g HPD a.e. The results in Propositions 1 and 2 can be found in [126, Theorems 2.5, 2.6, 3.1] for the unilevel case $k = 1$, and the extension to the multilevel setting is plain, so we decided not to report the proofs in this thesis.

Proposition 1. *Let $f : I_k \to \mathcal{M}_s$ be a Hermitian matrix-valued function in $L^1(k,s)$ and let*

$$m_f := \operatorname*{ess\,inf}_{x\in I_k} \lambda_{\min}(f(x)), \qquad M_f := \operatorname*{ess\,sup}_{x\in I_k} \lambda_{\max}(f(x)).$$

Then, the following properties hold.

1. *$\Lambda(T_n(f)) \subseteq [m_f, M_f]$ for every $n \in \mathbb{N}^k$.*
2. *If $\lambda_{\min}(f)$ is not a.e. constant, then $\Lambda(T_n(f)) \subset (m_f, M_f]$.*
   If $\lambda_{\max}(f)$ is not a.e. constant, then $\Lambda(T_n(f)) \subset [m_f, M_f)$.

In particular, if f is HPSD a.e. and $\lambda_{\min}(f)$ is not a.e. equal to 0, then $T_n(f)$ is HPD for all $n \in \mathbb{N}^k$.

Note that under the hypothesis that,f is a Hermitian matrix-valued function, every matrix $T_n(f)$ is Hermitian. This follows from the relation $T_n^*(f) = T_n(f^*)$, which holds for every $f \in L^1(k,s)$ and $n \in \mathbb{N}^k$, and from the fact that, if f is Hermitian a.e., then $f^* = f$ a.e.

Proposition 2. *Let f, g be Hermitian matrix-valued functions in $L^1(k,s)$ with g HPD a.e. Let $h = g^{-1}f$ and let*

$$m_h := \operatorname*{ess\,inf}_{x\in I_k} \lambda_{\min}(h(x)), \qquad M_h := \operatorname*{ess\,sup}_{x\in I_k} \lambda_{\max}(h(x)).$$

Then, $\Lambda(T_n^{-1}(g)T_n(f)) \subseteq [m_h, M_h]$ for all $n \in \mathbb{N}^k$.

In the following localization result the symbol is an integrable weakly sectorial/sectorial function, not necessarily Hermitian matrix-valued. It is taken from [135, Theorem 2.4]. We note that the statement of Theorem 2.4 in [135] contains a typo, which is present again in [94, Theorem 1] and which is corrected in the statement below.

Theorem 3. [135] *Let $f \in L^1(k,s)$ and let $d := d(\mathrm{Coh}[\mathcal{ENR}(f)], 0)$.*

- *Suppose f is weakly sectorial. Then $\sup_{z\in\mathcal{S}} d(z,0) = \max_{z\in\mathcal{S}} d(z,0) = d$, where $\mathcal{S}$ is the set of all separating lines for $\mathcal{ENR}(f)$. Moreover, $\sigma \geq d$ for all singular values σ of $T_n(f)$ and for all $n \in \mathbb{N}^k$.*
- *Suppose f is sectorial and let z be a separating line for $\mathcal{ENR}(f)$ such that the eigenvalue of minimum modulus of $\frac{1}{2}(\omega(z)f(x) + \overline{\omega(z)}f^*(x))$ is not a.e. equal to $d(z,0)$. Then $\sigma > d(z,0)$ for all singular values σ of $T_n(f)$ and for all $n \in \mathbb{N}^k$.*

In particular, if f is sectorial then all the matrices $T_n(f)$, $n \in \mathbb{N}^k$, are invertible.

Theorem 4 below (straightforward block extension of a theorem taken from [135]) is, to our knowledge, the first tool for devising spectrally equivalent preconditioners in the non-Hermitian multilevel block case.

We remark that, if $f \in L^1(k,s)$ and if $\widetilde{f}(x)$ is similar to $f(x)$ via a constant transformation C (independent of x), that is $f(x) = C\widetilde{f}(x)C^{-1}$ a.e., then $\widetilde{f} \in L^1(k,s)$ and $T_n(\widetilde{f}) = (I_{\widehat{n}} \otimes C)^{-1} T_n(f)(I_{\widehat{n}} \otimes C)$ for all $n \in \mathbb{N}^k$ ($I_{\widehat{n}}$ is the identity matrix of order $\widehat{n}$). This result follows from the definitions of $T_n(\widetilde{f}), T_n(f)$, see (1.15), and from the properties of the tensor Kronecker product of matrices.

Theorem 4. [94] *Suppose $f, g \in L^1(k,s)$ with g sectorial, and let $R(f,g) := \{\lambda \in \mathbb{C} : f - \lambda g$ is sectorial$\}$. Then, for any n, the eigenvalues of $T_n^{-1}(g)T_n(f)$ belong to $[R(f,g)]^c$, i.e. to the complementary set of $R(f,g)$. In addition, if $\widetilde{f}(x)$ is similar to $f(x)$ via a constant transformation and if $\widetilde{g}$ is similar to g via the same constant transformation, then $T_n^{-1}(g)T_n(f)$ is similar to $T_n^{-1}(\widetilde{g})T_n(\widetilde{f})$ by the above discussion and therefore, for any n, the eigenvalues of $T_n^{-1}(g)T_n(f)$ belong to $[R(\widetilde{f},\widetilde{g})]^c$ as well. As a consequence, if $\mathcal{F}$ denotes the set of all pairs $(\widetilde{f},\widetilde{g})$ satisfying the previous assumptions, then, for any n, the eigenvalues of $T_n^{-1}(g)T_n(f)$ belong to $\bigcap_{(\widetilde{f},\widetilde{g})\in\mathcal{F}} [R(\widetilde{f},\widetilde{g})]^c$.*

1.4.2 Distribution results

In this subsection we recall some distribution results both for unilevel and multilevel block Toeplitz sequences. The pioneering result in the unilevel context is due to Szegö and deals with bounded real-valued symbols.

Theorem 5. *Let $f \in L^\infty(I_1)$ be a real-valued function. Then,*

$$\{T_n(f)\}_n \sim_\lambda (f, I_1).$$

As regards the multilevel setting, in [143] Tilli showed that the sentence 'real-valued function' must be replaced by 'a function in the Tilli class' according to Definition 6.

Theorem 6. *Let $f \in L^\infty(I_k)$. If f is in the Tilli class, then $\{T_n(f)\}_{n\in\mathbb{N}^k} \sim_\lambda (f, I_k)$.*

Using Theorem 1, in [60] the authors generalized Theorem 6 to the matrix-valued symbol case.

Theorem 7. *Let $f \in L^\infty(k, s)$. If f is in the Tilli class, then $\{T_n(f)\}_{n\in\mathbb{N}^k} \sim_\lambda (f, I_k)$.*

In the case where $f \in L^1(k, s)$ is a Hermitian matrix-valued function, another version of Theorem 7, due to Tilli, is available (see [142]).

Theorem 8. *Let $f \in L^1(k, s)$ be a Hermitian matrix-valued function. Then, $\{T_n(f)\}_{n\in\mathbb{N}^k} \sim_\lambda (f, I_k)$.*

The cases $k = 1$, $s = 1$ and $k > 1$, $s = 1$, were contemporary studied by Tyrtyshnikov and Zamarashkin in [145, 146].

We conclude this subsection recalling a distribution result for preconditioned sequences of Hermitian multilvel block Toeplitz matrices (see [126, Theorem 3.10 and the discussion at the end of Subsection 3.3]).

Theorem 9. *Let $f, g \in L^1(k, s)$ be Hermitian matrix-valued functions with g HPD a.e., and let $h = g^{-1}f$. Then $\{T_n^{-1}(g)T_n(f)\}_{n\in\mathbb{N}^k} \sim_\lambda (h, I_k)$.*

In the next chapter we will enrich the previous scenario adding a distribution result on preconditioned sequences of non-Hermitian multilvel block Toeplitz matrices. Moreover, we will focus on some algebraic operations on multilevel block Toeplitz matrices (Theorems 13, 15).

1.5 Generalized Locally Toeplitz

As already argued, multilevel block Toeplitz matrices arise in many applications. Although, there are situations, e.g. the approximation by local methods (finite differences, finite elements, isogeometric analysis, etc.) of PDEs with nonconstant coefficients, general domains and nonuniform gridding, in which the class of Toeplitz matrices is no longer sufficient and a further structure of matrices is needed. With this objective, the Generalized Locally Toeplitz (GLT) algebra has been introduced in the pioneering work by Tilli [141], and widely generalized in [131, 133]. As we will point out in a moment, the sequences of multilevel block Toeplitz matrices generated by a $L^1(k, s)$ function, as well as their corresponding algebra, form a subset of the GLT class (see Section 3.3.1 in [133]). Unfortunately, the formal definitions are rather technical, difficult, and involve a heavy notation: therefore we just give and briefly discuss the notion of GLT class in one dimension. Moreover, we report few properties of the GLT class [74] in the general multidimensional setting.

Since a GLT sequence is a sequence of matrices obtained from a combination of some algebraic operations on multilevel block Toeplitz matrices and diagonal sampling matrices, we need the following definition.

Definition 8. Given a Riemann-integrable function a defined over $[0, 1]$, by *diagonal sampling matrix* of order n we mean $D_n(a) = \mathrm{diag}_{j=1,\dots,n}\, a\left(\frac{j}{n}\right)$.

A part from previous definition, to define the GLT class we need also the notion of approximating class of sequences (a.c.s.), which generalizes the concept of perturbation by small-norm plus low-rank terms, widely used in the preconditioning literature (see [136] and references therein). The guiding idea is that any reasonable approximation by local methods of PDEs leads to matrix-sequences that can be approximated in the a.c.s. sense by a finite sum of products of Toeplitz and diagonal sampling matrices; see [141, 131, 133, 74].

Definition 9 (a.c.s.). Let $\{A_n\}_n$ be a matrix-sequence. An *approximating class of sequences (a.c.s.)* for $\{A_n\}_n$ is a sequence of matrix-sequences $\{\{B_{n,m}\}_n\}_m$ with the following property: for every m there exists n_m such that, for $n \geq n_m$,

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

$$\operatorname{rank}(R_{n,m}) \le c(m)n, \qquad \|E_{n,m}\| \le \omega(m),$$

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

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

Roughly speaking, $\{\{B_{n,m}\}_n\}_m$ is an a.c.s. for $\{A_n\}_n$ if A_n is equal to $B_{n,m}$ plus a low-rank matrix (with respect to the matrix size), plus a small-norm matrix.

Definition 10. Let $m, n \in \mathbb{N}$, let $a : [0,1] \to \mathbb{C}$, and let $f : I_1 \to \mathbb{C}$ in $L^1(I_1)$. Then, we define the $n \times n$ matrix

$$\begin{aligned} \mathrm{LT}_n^m(a,f) &= D_m(a) \otimes T_{\lfloor n/m \rfloor}(f) \ \oplus \ O_{n \bmod m} \\ &= \operatorname*{diag}_{j=1,\dots,m} a\Big(\frac{j}{m}\Big) \otimes T_{\lfloor n/m \rfloor}(f) \ \oplus \ O_{n \bmod m}, \end{aligned}$$

where the tensor Kronecker product operation $\otimes$ is applied before the direct sum $\oplus$. It is understood that $\mathrm{LT}_n^m(a,f) = O_n$ when $n < m$ and that the term $O_{n \bmod m}$ is not present when n is a multiple of m.

Definition 11 (LT sequence). Let $\{A_n\}_n$ be a matrix-sequence. We say that $\{A_n\}_n$ is a *separable Locally Toeplitz (sLT)* sequence if there exist

- a Riemann-integrable function $a : [0,1] \to \mathbb{C}$,
- a function $f \in L^1(I_1)$,

such that $\{\{\mathrm{LT}_n^m(a,f)\}_n\}_m$ is an a.c.s. for $\{A_n\}_n$. In this case, we write $\{A_n\}_n \sim_{\mathrm{sLT}} a(x)f(\theta)$. The function $a(x)f(\theta)$ is referred to as the *symbol* of the sequence $\{A_n\}_n$, a is the *weight function* and f is the *generating function*.

Definition 12 (GLT sequence). Let $\{A_n\}_n$ be a matrix-sequence and let $\kappa : [0,1] \times I_1 \to \mathbb{C}$ be a measurable function. We say that $\{A_n\}_n$ is a *GLT sequence* with symbol $\kappa(x,\theta)$, and we write

$$\{A_n\}_n \sim_{\mathrm{GLT}} \kappa(x,\theta),$$

if:

- for any $\epsilon > 0$ there exist matrix-sequences $\{A_n^{(i,\epsilon)}\}_n \sim_{\mathrm{sLT}} a_{i,\epsilon}(x)f_{i,\epsilon}(\theta),\ i = 1,\dots,\eta_\epsilon$;
- $\sum_{i=1}^{\eta_\epsilon} a_{i,\epsilon}(x)f_{i,\epsilon}(\theta) \to \kappa(x,\theta)$ in measure over $[0,1] \times I_1$ when $\epsilon \to 0$;
- $\Big\{\big\{\sum_{i=1}^{\eta_\epsilon} A_n^{(i,\epsilon)}\big\}_n\Big\}_m$, with $\epsilon = (m+1)^{-1}$, is an a.c.s. for $\{A_n\}_n$.

Now we shortly mention four main features of the GLT class in the general multidimensional setting in which $\{A_n\}_{n\in\mathbb{N}^k} \sim_{\mathrm{GLT}} \kappa(x,\theta)$ with $\kappa : G \to \mathbb{C}$, $G = [0,1]^k \times I_k$.

GLT1 Let $\{A_n\}_{n\in\mathbb{N}^k} \sim_{\mathrm{GLT}} \kappa(x,\theta)$ with $\kappa : G \to \mathbb{C}$, $G = [0,1]^k \times I_k$, then $\{A_n\}_{n\in\mathbb{N}^k} \sim_\sigma (\kappa, G)$. If the matrices A_n are Hermitian, then it holds also $\{A_n\}_{n\in\mathbb{N}^k} \sim_\lambda (\kappa, G)$.

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

GLT3 Every Toeplitz family $\{T_n(f)\}_{n\in\mathbb{N}^k}$ generated by a $L^1(k,s)$ function $f = f(\theta)$ is $\{T_n(f)\}_{n\in\mathbb{N}^k} \sim_{\mathrm{GLT}} f(\theta)$, with the specifications reported in item **[GLT1]**: we notice that the function f does not depend on the spacial variable $x \in [0,1]^k$. Every diagonal sampling sequence $\{D_n(a)\}_{n\in\mathbb{N}^k}$, where $a : [0,1]^k \to \mathbb{C}$ is a Riemann integrable function, is $\{D_n(a)\}_{n\in\mathbb{N}^k} \sim_{\mathrm{GLT}} a(x)$.

GLT4 Let $\{A_n\}_{n\in\mathbb{N}^k} \sim_\sigma (0, G)$, $G = [0,1]^k \times I_k$, then $\{A_n\}_{n\in\mathbb{N}^k} \sim_{\mathrm{GLT}} 0$.

Remark 4. The approximation by local methods (finite differences, finite elements, isogeometric analysis, etc.) of PDEs with nonconstant coefficients, general domains, nonuniform gridding leads to GLT sequences, under very mild assumptions (see [141, 131, 133] for the case of finite differences, [15, 75] for the finite element setting, and [52, 73] for the case of isogeometric analysis approximations): in Chapter 4, as a byproduct, we show that the approximation of fractional diffusion equations leads to GLT sequences as well.

1.6 Preconditioning for Toeplitz matrices

In order to solve linear systems of type

$$A_n x = b \tag{1.19}$$

where $A_n = T_n(f)$ is the Toeplitz matrix associated to the symbol f, iterative methods like Krylov methods, e.g., Conjugate Gradient (CG) (when A_n is an HPD matrix) or Conjugate Gradient for Least Squares (CGLS), Generalized Minimal Residual (GMRES) (when A_n is a non-Hermitian matrix), can be applied. A motivation for using iterative methods when solving linear systems with Toeplitz structure is that the matrix-vector products $A_n y$ can be computed efficiently. The convergence rate of the CG-like methods depends on the condition number of the matrix A_n and on how the spectrum of A_n is clustered (see [5]). For example, if the condition number of A_n is large, the calculation of the solution by means of iterative methods may become very slow. One way to speed up their convergence rate is to precondition the linear system. Thus, instead of solving (1.19), we solve the preconditioned system

$$P_n^{-1} A_n x = P_n^{-1} b.$$

The matrix P_n, called the *preconditioner*, should be chosen according to the following criteria

1. the solution of a linear system with coefficient matrix P_n should cost as a matrix-vector product;
2. P_n must have similar spectral properties as the original matrix A_n, in such a way that the spectrum of $P_n^{-1} A_n$ is clustered, according to Definition 4, around 1.

An HPD matrix P_n is an *optimal* preconditioner for A_n if and only if there exists an $\bar{n}$ such that for any $n \geq \bar{n}$ all the eigenvalues of $P_n^{-1} A_n$ belong to a positive bounded universal interval independent of n. Under this assumption the optimal convergence of a CG-like method is guaranteed. For optimal methods we mean methods such that the complexity of solving the given linear system is proportional to the cost of matrix-vector multiplication (see [7] for a precise notion). For an iterative method this implies a convergence, within a preassigned accuracy, in a number of iterations independent of n and that the cost of every iteration is of the same order as that of the matrix-vector product.

If A_n and P_n are HPD matrices, then the superlinear convergence for preconditioned CG-like methods is guaranteed whenever the eigenvalues of $P_n^{-1} A_n$ have a strong cluster at 1. In the non-Hermitian case, a preconditioned CG-like method can be applied to the symmetrized preconditioned system, and the superlinear convergence can be seen as long as the singular values of $P_n^{-1} A_n$ have a strong cluster at 1. P_n is called *superlinear* preconditioner if $P_n^{-1} A_n - I$ has a strong cluster around zero (see [136]). Superlinear preconditioners provide the superlinear convergence of CG-like methods.

The literature of preconditioners for structured matrices is really vast, in this section we just recall the results for both well-conditioned and ill-conditioned matrices needed throughout the thesis (see the review [37]).

Circulant preconditioners An $n \times n$ matrix is said to be circulant if

$$C_n = \begin{bmatrix} c_0 & c_{-1} & \cdots & c_{2-n} & c_{1-n} \\ c_1 & c_0 & c_{-1} & \vdots & c_{2-n} \\ \vdots & c_1 & c_0 & \ddots & \vdots \\ c_{n-2} & & \ddots & \ddots & c_{-1} \\ c_{n-1} & c_{n-2} & \cdots & c_1 & c_0 \end{bmatrix}_{n \times n},$$

where $c_{-k} = c_{n-k}$, $k = 1, \ldots, n-1$. Circulant matrices are diagonalized by the Discrete Fourier Transform (DFT) F_n, i.e.

$$C_n = F_n^* \Lambda(C_n) F_n, \tag{1.20}$$

where the entries of F_n are given by

$$[F_n]_{j,k} = \frac{1}{\sqrt{n}} \mathrm{e}^{2\pi \mathbf{i} jk/n}, \quad 0 \leq j, k \leq n-1. \tag{1.21}$$

The matrix $\Lambda(C_n)$ can be obtained in $O(n \log n)$ operation by taking the Fast Fourier Transform (FFT) of the first column of C_n. In fact, the diagonal entries $\lambda_k(C_n)$ of $\Lambda(C_n)$ are given by

$$\lambda_k(C_n) = \sum_{j=0}^{n-1} c_j \mathrm{e}^{2\pi \mathbf{i} jk/n}, \quad k = 0, \ldots, n-1.$$

Once $\Lambda(C_n)$ is obtained, the product $C_n y$ and $C_n^{-1} y$ for any vector y can be computed by FFTs in $O(n \log n)$ operations, using (1.20).

The matrix-vector multiplications $A_n y$ can also be computed by FFts by embedding A_n into a $2n \times 2n$ circulant matrix, i.e.

$$\begin{bmatrix} A_n & * \\ * & A_n \end{bmatrix}$$

and then carrying out the multiplication by using the decomposition of circulant matrices. The matrix-vector multiplication thus requires $O(2n \log 2n)$ operations.

Strang ([139]) and T. Chan ([40]) proposed the use of circulant matrices to precondition Toeplitz matrices in CG-like iterations.

For an $n \times n$ Toeplitz matrix $A_n = [t_{k-\ell}]_{k,\ell=1}^n$

- the *Strang circulant* preconditioner is defined to be the matrix that copies the central diagonals of A_n and reflects them around to complete the circulant requirement. The diagonals s_j of the Strang preconditioner $s(A_n) = [s_{k-\ell}]_{k,\ell=1}^n$ are given by

$$s_j = \begin{cases} t_j, & 0 < j \le \lfloor n/2 \rfloor, \\ t_{j-n}, & \lfloor n/2 \rfloor < j < n, \\ s_{n+j}, & 0 < -j < n. \end{cases}$$

- the *Chan circulant* preconditioner $c(A_n)$ is defined to be the minimizer of

$$\|C_n - A_n\|_F,$$

over all $n \times n$ circulant matrices C_n. The j-th diagonals of $c(A_n)$ are

$$c_j = \begin{cases} \frac{(n-j)t_j + j t_{j-n}}{n}, & 0 \le j < n, \\ c_{n+j}, & 0 < -j < n. \end{cases}$$

which are just the average of the diagonals of A_n, with the diagonals being extended to length n by a wrap-around.

For the convergence of these preconditioners we need the definition of Wiener class.

Definition 13. The *Wiener class* is the set of functions $f(\theta) = \sum_{k=-\infty}^{\infty} f_k \mathrm{e}^{\mathbf{i}k\theta}$ such that $\sum_{k=-\infty}^{\infty} |f_k| < \infty$.

Note that the Wiener class forms a subalgebra of the continuous and 2π–periodic functions.
If A_n is associated to a positive symbol in the Wiener class, the circulant preconditioners proposed by Chan and Strang are superlinear [38]. More precisely, this is true for the Strang preconditioner when f belongs to the Dini–Lipschitz class, while for the Chan preconditioner this is true when f is merely continuous.

We point out that the above preconditioning techniques have not been designed for ill-conditioned Toeplitz matrices whose generating function vanishes at some points. Indeed, a preconditioned CG-like method with those circulant preconditioners fails in the case where f has zeros [144].

In the case $A_n = T_n(f)$ is a 2-level matrix ($n = (n_1, n_2)$), that is when A_n is a Block Toeplitz with Toeplitz Blocks (BTTB), a circulant preconditioner is intended to be a Block Circulant with Circulant Blocks (BCCB). Such kind of matrices can be spectrally decomposed using the two-dimensional DFT

$$F = F_{n_2} \otimes F_{n_1} \in \mathcal{M}_{\widehat{n}}, \tag{1.22}$$

with F_{n_1}, F_{n_2} as in (1.21).

BCCB preconditioners for BTTB matrices (cf. T. Chan and Olkin [41]) and low-rank perturbations thereof have been investigated by Ku and Kuo [99], Tyrtyshnikov [145], and R. Chan and Jin [35].

Unfortunately, it is well-known by seminal results in [136], that a BCCB preconditioner for a BTTB matrix cannot be superlinear even for an optimal (i.e., "almost exact") approximation of its generating function. More in general, if A_n is a k-level Toeplitz matrix, general results in [136, 137] tell us that, even in the well-conditioned case, the performances of multilevel circulant preconditioners deteriorate when k increases and the superlinear behaviour of any preconditioned Krylov method is lost. This behaviour is not limited to circulant preconditioners, but extends to preconditioners in appropriate algebras of matrices with fast transforms like as trigonometric and Hartley algebras, wavelet algebras, etc. ([129, 111]).

Toeplitz preconditioners Instead of circulant preconditioners one can think to use a preconditioner which preserves the structure of the coefficient matrix. In this direction, an alternative technique is represented by the *band Toeplitz preconditioning*. The motivation behind using band Toeplitz matrices is to approximate the generating function f by a trigonometric polynomial g of fixed degree. The advantage here is that trigonometric polynomials can be chosen to match the zeros (with the same order) of f, so that the preconditioned method still works when f has zeros, that is in the ill-conditioned case. Moreover, at least in the univariate context, we recall that for solving banded systems we can apply specialized versions of the Gaussian Elimination maintaining an optimal linear cost.

Concerning the band Toeplitz preconditioning, we emphasize that the technique has been explored for k-level Toeplitz matrix with scalar-valued symbol in [32, 48, 121, 39, 123], even in the (asymptotically) ill-conditioned case, but with a specific focus on the HPD case. In detail, in [121], it has been proved that a BTTB preconditioner of a BTTB matrix is optimal even when the spectrum of the preconditioner is a poor approximation of the spectrum of the system matrix (in particular, it is enough that all the zeros of the generating function are exactly located with the same multiplicity, regardless the other values of the generating function).

Specific attempts in the non-Hermitian case can be found in [34, 94]. Further results concerning genuine block Toeplitz structures with matrix-valued symbol are considered in [124, 126], but again for HPD matrix-valued symbols. In the next chapter we enrich the literature focusing on band Toeplitz preconditioning for the non-Hermitian multilevel block case.

Clearly, instead of polynomials any function g that matches the zeros of f and gives rise to Toeplitz matrices can also be considered. For example, in [36], R. Chan and Ng used the Toeplitz matrix generated by $1/f$ to approximate the inverse of the Toeplitz matrix A_n generated by f. In that paper, it has been proved that the spectrum of the preconditioned matrix $P_n^{-1}A_n$ is clustered around 1. However, in general it may be difficult to compute the Fourier coefficients of $1/f$ explicitly, and hence P_n cannot be formed efficiently. R. Chan and Ng thus have derived families of Toeplitz preconditioners by using different kernel functions and different levels of approximation for the Fourier coefficients of $1/f$.

1.7 The HSS and PHSS methods

Given a square matrix A_n, the corresponding *Hermitian/Skew-Hermitian Splitting (HSS)* is given by

$$A_n = \mathrm{Re}(A_n) + \mathbf{i}\,\mathrm{Im}(A_n),$$

where $\mathrm{Re}(A_n)$ and $\mathrm{Im}(A_n)$ are, respectively, the real and imaginary part of A_n. They are defined as

$$\mathrm{Re}(A_n) := \frac{A_n + A_n^*}{2}, \qquad \mathrm{Im}(A_n) := \frac{A_n - A_n^*}{2\mathbf{i}}.$$

By construction, $\mathrm{Re}(A_n)$ and $\mathrm{Im}(A_n)$ are Hermitian.

The HSS method

Let us consider the linear system

$$A_n x = b, \qquad A_n \in \mathcal{M}_{d_n}, \qquad x, b \in \mathbb{C}^{d_n}, \qquad d_n \to \infty \ \text{ as } \ n \to \infty, \tag{1.23}$$

where the real part $\mathrm{Re}(A_n)$ is positive definite.[1] This assumption, which, by the Fan-Hoffman theorem [19, Proposition III.5.1], ensures the invertibility of A_n, is satisfied in important applications; see e.g. [17] and references therein.

Now, given a positive parameter α and assuming that $\mathrm{Re}(A_n)$ is positive definite, the HSS of A_n is related to the following convergent two-step iteration for the solution of the linear system (1.23):

$$\begin{cases} (\alpha I + \mathrm{Re}(A_n))\, x^{(k+\frac{1}{2})} = (\alpha I - \mathbf{i}\,\mathrm{Im}(A_n))\, x^{(k)} + b, \\ (\alpha I + \mathbf{i}\,\mathrm{Im}(A_n))\, x^{(k+1)} = (\alpha I - \mathrm{Re}(A_n))\, x^{(k+\frac{1}{2})} + b, \end{cases} \tag{1.24}$$

where I is the identity matrix, while $x^{(0)}$ is a initial guess. It is not difficult to see that the above iteration, called *HSS iteration*, gives rise to a stationary iterative method, the *HSS method*, whose

[1] In the literature, a matrix A with positive definite real part is often referred to as a *positive definite matrix*. Note that, according to this general definition, a positive definite matrix need not to be Hermitian; moreover, a *real* matrix A is positive definite if and only if $x^*Ax > 0$ for all nonzero real vectors x.

iteration matrix is

$$M_n(\alpha) := (\alpha I + \mathbf{i}\,\mathrm{Im}(A_n))^{-1}\,(\alpha I - \mathrm{Re}(A_n))\,(\alpha I + \mathrm{Re}(A_n))^{-1}\,(\alpha I - \mathbf{i}\,\mathrm{Im}(A_n))\,.$$

Note that $M_n(\alpha)$ is well-defined. Indeed, $\alpha I + \mathbf{i}\,\mathrm{Im}(A_n)$ is invertible, because α is nonzero and $\mathbf{i}\,\mathrm{Im}(A_n)$ is skew-Hermitian, and $\alpha I + \mathrm{Re}(A_n)$ is also invertible, because α is positive and $\mathrm{Re}(A_n)$ is positive definite by hypothesis. As proved in [11], the convergence rate of the HSS method, that is the spectral radius of $M_n(\alpha)$, is bounded from above by the spectral radius of the Hermitian matrix

$$(\alpha I - \mathrm{Re}(A_n))\,(\alpha I + \mathrm{Re}(A_n))^{-1}\,. \tag{1.25}$$

In formulas,

$$\rho(M_n(\alpha)) \le \rho((\alpha I - \mathrm{Re}(A_n))\,(\alpha I + \mathrm{Re}(A_n))^{-1}) = \max_{\lambda\in\Lambda(\mathrm{Re}(A_n))}\left|\frac{\alpha-\lambda}{\alpha+\lambda}\right|. \tag{1.26}$$

Note that the matrix (1.25) forms the 'central part' of $M_n(\alpha)$ or, in other words, the part of $M_n(\alpha)$ associated with $\mathrm{Re}(A_n)$. The upper bound in (1.26) is unconditionally less than 1 under the only assumptions that α is positive and $\mathrm{Re}(A_n)$ is positive definite. This means that the HSS method is unconditionally convergent. However, if A_n is ill-conditioned (this may happen for instance when n is large, especially if A_n has small eigenvalues tending to 0 as $n \to \infty$), then it can be checked that, independently of α, the upper bound in (1.26) is not satisfactory and, in general, also the convergence rate $\rho(M_n(\alpha))$ is poor. Such a situation typically occurs in the approximation of convection diffusion PDEs, as reported in [17].

Several generalizations of the HSS method have been proposed in the literature for accelerating the convergence rate (see [10, 12, 16] and references therein). A possibility in this direction is to use a preconditioning technique, giving rise to the *Preconditioned HSS (PHSS) method*.

The PHSS method

Let P_n be a HPD matrix and let us define the PHSS method as follows [17]: given a positive parameter α and a initial guess $x^{(0)}$, for $k = 0, 1, 2, \ldots$ compute

$$\begin{cases} \left(\alpha I + P_n^{-1}\mathrm{Re}(A_n)\right) x^{(k+\frac{1}{2})} = \left(\alpha I - P_n^{-1}\mathbf{i}\,\mathrm{Im}(A_n)\right) x^{(k)} + P_n^{-1}b, \\ \left(\alpha I + P_n^{-1}\mathbf{i}\,\mathrm{Im}(A_n)\right) x^{(k+1)} = \left(\alpha I - P_n^{-1}\mathrm{Re}(A_n)\right) x^{(k+\frac{1}{2})} + P_n^{-1}b, \end{cases} \tag{1.27}$$

until convergence. A simple check shows that the iteration matrix of the PHSS method (1.27) is

$$\begin{aligned} M_n(\alpha) &:= (\alpha I + \mathbf{i}\,P_n^{-1}\mathrm{Im}(A_n))^{-1}(\alpha I - P_n^{-1}\mathrm{Re}(A_n))(\alpha I + P_n^{-1}\mathrm{Re}(A_n))^{-1}(\alpha I - \mathbf{i}\,P_n^{-1}\mathrm{Im}(A_n)) \\ &= (\alpha P_n + \mathbf{i}\,\mathrm{Im}(A_n))^{-1}(\alpha P_n - \mathrm{Re}(A_n))(\alpha P_n + \mathrm{Re}(A_n))^{-1}(\alpha P_n - \mathbf{i}\,\mathrm{Im}(A_n)). \end{aligned} \tag{1.28}$$

From this, it is clear that the PHSS cannot be interpreted as the HSS applied to the matrix $P_n^{-1}A_n$, because $P_n^{-1}\mathrm{Re}(A_n) + \mathbf{i}\,P_n^{-1}\mathrm{Im}(A_n)$ is not the HSS of $P_n^{-1}A_n$. However, we know that any HPD matrix P_n can be factorized as $P_n = L_nL_n^*$ for some nonsingular matrix L_n; for instance, we can take $L_n = P_n^{1/2}$ or L_n equal to the lower triangular matrix in the Cholesky factorization of P_n. The PHSS (1.27) can then be interpreted as the HSS (1.24) applied to the matrix $L_n^{-1}A_nL_n^{-*}$ (which is similar to $P_n^{-1}A_n$ via the similarity transformation $X \mapsto L_n^{-*}XL_n^*$); note in fact that these two methods have the same convergence rate, because their respective iteration matrices are similar. Indeed, denoting by

$$\begin{aligned} \hat{M}_n(\alpha) := &\left(\alpha I + \mathbf{i}\,\mathrm{Im}(L_n^{-1}A_nL_n^{-*})\right)^{-1}\left(\alpha I - \mathrm{Re}(L_n^{-1}A_nL_n^{-*})\right) \\ &\left(\alpha I + \mathrm{Re}(L_n^{-1}A_nL_n^{-*})\right)^{-1}\left(\alpha I - \mathbf{i}\,\mathrm{Im}(L_n^{-1}A_nL_n^{-*})\right) \end{aligned}$$

the iteration matrix of the HSS applied to $L_n^{-1}A_nL_n^{-*}$, and taking into account the relations

$$\mathrm{Re}(L_n^{-1}A_nL_n^{-*}) = L_n^{-1}\mathrm{Re}(A_n)L_n^{-*}, \qquad \mathrm{Im}(L_n^{-1}A_nL_n^{-*}) = L_n^{-1}\mathrm{Im}(A_n)L_n^{-*},$$

we obtain that the PHSS iteration matrix (1.28) can be expressed as

$$M_n(\alpha) = L_n^*\hat{M}_n(\alpha)L_n^{-*} \sim \hat{M}_n(\alpha).$$

Another viewpoint is as follows: the PHSS (1.27) coincides with the following method, obtained from the original HSS (1.24) by replacing the identity matrix I with the preconditioner P_n:

$$\begin{cases} (\alpha P_n + \mathrm{Re}(A_n))\, x^{(k+\frac{1}{2})} = (\alpha P_n - \mathbf{i}\,\mathrm{Im}(A_n))\, x^{(k)} + b, \\ (\alpha P_n + \mathbf{i}\,\mathrm{Im}(A_n))\, x^{(k+1)} = (\alpha P_n - \mathrm{Re}(A_n))\, x^{(k+\frac{1}{2})} + b. \end{cases}$$

From the first interpretation of the PHSS given above and from the discussion concerning the unconditional convergence of the HSS, it follows that the PHSS (1.27), like the HSS (1.24), converges unconditionally to the solution of the linear system in (1.23) under the only assumptions that $\alpha > 0$ and $\mathrm{Re}(A_n)$ is positive definite. More specifically, the PHSS convergence rate, that is the spectral radius of the PHSS iteration matrix $M_n(\alpha)$ in (1.28), is bounded from above by the spectral radius of the Hermitian matrix

$$(\alpha I - P_n^{-1/2}\mathrm{Re}(A_n)P_n^{-1/2})(\alpha I + P_n^{-1/2}\mathrm{Re}(A_n)P_n^{-1/2})^{-1}$$

which is similar to

$$(\alpha I - P_n^{-1}\mathrm{Re}(A_n))(\alpha I + P_n^{-1}\mathrm{Re}(A_n))^{-1}. \tag{1.29}$$

Therefore, we have

$$\rho(M_n(\alpha)) \le \rho((\alpha I - P_n^{-1}\mathrm{Re}(A_n))(\alpha I + P_n^{-1}\mathrm{Re}(A_n))^{-1}) = \max_{\lambda\in\Lambda(P_n^{-1}\mathrm{Re}(A_n))} \left|\frac{\alpha-\lambda}{\alpha+\lambda}\right|. \tag{1.30}$$

Note that the matrix (1.29) forms the 'central part' of $M_n(\alpha)$, which is the one related to $\mathrm{Re}(A_n)$; cf. (1.28). The best parameter α that minimizes the upper bound (1.30) is the geometric mean of the extreme eigenvalues of $P_n^{-1}\mathrm{Re}(A_n)$. However, we should say that there exist situations in which the skew-Hermitian contributions in the PHSS iteration matrix have a role in accelerating the convergence rate. The explanation of this phenomenon falls in the theory of multi-iterative methods [120]. In these situations, the upper bound (1.30) may be quite inaccurate, since it is only based on the knowledge of the real part $\mathrm{Re}(A_n)$ and completely ignores the effects due to the imaginary part $\mathrm{Im}(A_n)$.

All these remarks on the PHSS convergence properties are collected in Theorem 10. In the following, if $S \in \mathcal{M}_{d_n}$ is any invertible matrix, we denote by $\|\cdot\|_S$ both the vector norm and the matrix norm induced by S. They are defined as

$$\begin{aligned} \|y\|_S &= \|Sy\|_2, \qquad y \in \mathbb{C}^{d_n}, \\ \|X\|_S &= \max_{\|y\|_S=1} \|Xy\|_S = \|SXS^{-1}\|, \qquad X \in \mathcal{M}_{d_n}. \end{aligned}$$

Theorem 10. [17] *Let A_n, $P_n \in \mathcal{M}_{d_n}$ be matrices such that $\mathrm{Re}(A_n)$, P_n are HPD, and let $\alpha > 0$. Then, the following results hold.*

1. *The PHSS method* (1.27) *has iteration matrix $M_n(\alpha)$, as reported in (1.28), and*

$$\rho(M_n(\alpha)) \le \|M_n(\alpha)\|_{S_n(\alpha)} \le \sigma_n(\alpha), \tag{1.31}$$

where $S_n(\alpha) := (\alpha I + P_n^{-1/2}\mathbf{i}\,\mathrm{Im}(A_n)P_n^{-1/2})P_n^{1/2}$ and

$$\sigma_n(\alpha) := \max_{\lambda\in\Lambda(P_n^{-1}\mathrm{Re}(A_n))} \left|\frac{\alpha-\lambda}{\alpha+\lambda}\right| < 1.$$

In particular, the PHSS iteration (1.27) *converges to the unique solution of the linear system* (1.23).

2. *The best parameter α that minimizes the upper bound $\sigma_n(\alpha)$ is*

$$\alpha_n^* := \sqrt{\lambda_{\min}(P_n^{-1}\mathrm{Re}(A_n))\lambda_{\max}(P_n^{-1}\mathrm{Re}(A_n))} \tag{1.32}$$

and, consequently, the best upper bound is

$$\sigma_n(\alpha_n^*) = \frac{\sqrt{\kappa_n}-1}{\sqrt{\kappa_n}+1}, \tag{1.33}$$

where

$$\kappa_n := \frac{\lambda_{\max}(P_n^{-1}\mathrm{Re}(A_n))}{\lambda_{\min}(P_n^{-1}\mathrm{Re}(A_n))} \tag{1.34}$$

is the spectral condition number of the Hermitian matrix $P_n^{-1/2}\mathrm{Re}(A_n)P_n^{-1/2} \sim P_n^{-1}\mathrm{Re}(A_n)$.

1.8 Multigrid methods for Toeplitz matrices

In this section, we recall the basic idea of a multigrid method, then we focus on algebraic multigrid methods for multilevel Toeplitz matrices with scalar-valued symbol. The reported information will be used in Chapters 4 and 5.

1.8.1 Multigrid methods: basic idea

When a classical stationary iterative method is used to solve a linear system, the error components corresponding to large eigenvalues are damped efficiently, while the error components corresponding to the small eigenvalues are reduced slowly. Since in the discretized PDE the former correspond to rough error modes, while the latter to smooth error modes, methods like Jacobi are known as *smoothers*. The main aim of a multigrid (MG) method is to combine a smoother with some strategy able to damp the error components corresponding to the small eigenvalues, using some geometrically or algebraically constructed hierarchy of linear systems.

Let $Ax = b$ be the linear system we want to solve, with $x, b \in \mathbb{C}^n$ and $A \in \mathcal{M}_n$ HPD matrix. Fix $m+1$ integers $n = N_0 > N_1 > \cdots > N_m > 0$, where $0 < m < n$ denotes the maximum number of levels we decided to use. To define a multigrid method the following ingredients are needed for every level $i = 0, \ldots, m-1$:

1. appropriate smoothers $\mathcal{S}_i$, $\tilde{\mathcal{S}}_i$, and the corresponding smoothing steps ν_i, $\tilde{\nu}_i$;
2. restriction operators $R_i : \mathbb{C}^{N_i} \to \mathbb{C}^{N_{i+1}}$ and prolongation operators $P_i : \mathbb{C}^{N_{i+1}} \to \mathbb{C}^{N_i}$ to transfer a quantity between levels i and $i+1$;
3. the matrix at the coarser level $A_{i+1} \in \mathbb{C}^{N_{i+1} \times N_{i+1}}$ ($A_0 = A$, $b_0 = b$).

One iteration of MG in the *V-cycle version* consists of the following steps:

- ν_i pre-smoothing steps are performed using $\mathcal{S}_i$;
- the current iteration is corrected using the coarser level, process which is known as *coarse grid correction*. More precisely, the residual $r_i \in \mathbb{C}^{N_i}$ is computed and restricted to the coarse grid obtaining r_{i+1}, which is used to solve the error equation on the coarse grid

  $$A_{i+1} e_{i+1} = r_{i+1},$$

  by a recursive application of MG. The error e_{i+1} is interpolated back to obtain the finer level error e_i which is used to update the current iteration. The iteration matrix of the coarse grid correction is

  $$CGC_i = I_{N_i} - P_i A_{i+1}^{-1} R_i A_i;$$

- the iterate is improved by $\tilde{\nu}_i$ steps performed using $\tilde{\mathcal{S}}_i$.

The following algorithm summarize one iteration of a V-cycle MG.

Algorithm 1 $\mathcal{MG}(i, x_i, b_i)$

if $i = m$ **then**

$x_m \leftarrow A_m^{-1} r_m$

else

$x_i \leftarrow \mathcal{S}_i^{\nu_i}(x_i, b_i)$

$r_i \leftarrow b_i - A_i x_i$

$r_{i+1} \leftarrow R_i r_i$

$e_{i+1} \leftarrow \mathcal{MG}(i+1, 0_{N_{i+1}}, r_{i+i})$

$e_i \leftarrow P_i e_{i+1}$

$x_i \leftarrow x_i + e_i$

$x_i \leftarrow \tilde{\mathcal{S}}_i^{\tilde{\nu}_i}(x_i, b_i)$

end

For a given initial guess $x^{(0)}$, one MG iteration can be described as $x^{(l+1)} = \mathcal{MG}(0, x^{(l)}, b_0)$, for $l = 0, 1, \ldots$

A simple choice for both pre- and post-smoothing is relaxed Richardson

$$\begin{aligned} \mathcal{S}_i(x_i, b_i) &= S_i x_i + \omega_i b_i, \quad S_i = (I_{N_i} - \omega_i A_i), \quad \omega > 0, \\ \tilde{\mathcal{S}}_i(x_i, b_i) &= \tilde{S}_i x_i + \tilde{\omega}_i b_i, \quad \tilde{S}_i = (I_{N_i} - \tilde{\omega}_i A_i), \quad \tilde{\omega} > 0. \end{aligned} \tag{1.35}$$

When deriving convergence estimates for MG, usually, R_i is chosen to be the adjoint of P_i and the coarse grid matrix A_{i+1} is chosen as $P_i^* A_i P_i$. These conditions are known in the related literature as *Galerkin conditions* and the resulting method is the so-called *algebraic multigrid* (AMG). Note that, if the projectors have full rank, the matrix at the next coarse level is nonsingular and still HPD.

Let us define $\|\cdot\|_X = \|X^{1/2} \cdot\|_2$, where $\|\cdot\|_2$ is the usual Euclidean norm on $\mathbb{C}^n$ and X is an HPD matrix. The following theorem presents a convergence result for AMG ([1, 117]).

Theorem 11. *Let m, n be integers satisfying $0 < m < n$ and suppose that $A \in \mathcal{M}_n$ is a HPD matrix; given a sequence of $m+1$ positive integers $n = N_0 > N_1 > \cdots > N_m > 0$, let $P_i : \mathbb{C}^{N_{i+1}} \to \mathbb{C}^{N_i}$, $i = 0, \ldots, m-1$ be a full-rank matrix and $R_i = P_i^*$. Define $A_0 = A$, $b_0 = b$, $A_{i+1} = P_i^* A_i P_i$, $i = 0, \ldots, m-1$ and choose two classes of iterative methods $\mathcal{S}_i$, $\tilde{\mathcal{S}}_i$ whose iteration matrices are S_i, $\tilde{S}_i$, respectively. If there exists three real positive numbers α_i, β_i, γ_i such that*

$$\|S_i^{\nu_i} x\|_{A_i}^2 \le \|x\|_{A_i}^2 - \alpha_i \|S_i^{\nu_i} x\|_{A_i^2}^2 \quad \forall x \in \mathbb{C}^{N_i} \qquad \text{(pre-smoothing property)} \tag{1.36}$$

$$\|\tilde{S}_i^{\tilde{\nu}_i} x\|_{A_i}^2 \le \|x\|_{A_i}^2 - \beta_i \|x\|_{A_i^2}^2 \quad \forall x \in \mathbb{C}^{N_i} \qquad \text{(post-smoothing property)} \tag{1.37}$$

$$\|CGC_i x\|_{A_i}^2 \le \gamma_i \|x\|_{A_i^2}^2 \quad \forall x \in \mathbb{C}^{N_i} \qquad \text{(approximation property)} \tag{1.38}$$

for every $i = 0, \ldots, m-1$, defined

$$\delta_{\text{pre}} := \min_{0 \le i < m} \frac{\alpha_i}{\gamma_i}, \quad \delta_{\text{post}} := \min_{0 \le i < m} \frac{\beta_i}{\gamma_i},$$

it holds that $\delta_{\text{post}} \le 1$ and

$$\|AMG_0\|_A \le \sqrt{\frac{1 - \delta_{\text{post}}}{1 + \delta_{\text{pre}}}} < 1,$$

where AMG_0 is the V-cycle iteration matrix.

Remark 5. From Theorem 11, the sequence $\{x^{(l)}\}_{l \in \mathbb{N}}$ generated by AMG converges to the solution of $Ax = b$. Moreover, when at least one of the following conditions holds

$$\inf_t \min_{0 \le i < m(t)} \frac{\alpha_i}{\gamma_i} > 0, \quad \inf_t \min_{0 \le i < m(t)} \frac{\beta_i}{\gamma_i} > 0,$$

the convergence is optimal, that is, the method converges with a constant error reduction not depending on n and m.

Remark 6. Observe that the smoothing properties (1.36)-(1.37) are related to the smoothers, but not to the projectors, while the approximation property (1.38) depends only on the choice of the projectors.

When we consider only two levels in Algorithm 1, the V-cycle MG is known as Two-Grid Method (TGM). As regards the convergence under Galerkin conditions, the following theorem holds ([117]).

Theorem 12. *Suppose that $A \in \mathcal{M}_n$ is a HPD matrix; given a positive integer $\tilde{n} < n$, let $P : \mathbb{C}^{\tilde{n}} \to \mathbb{C}^n$, be a full-rank matrix and $R = P^*$. Define, $\tilde{A} = P^* A P$ and choose two classes of iterative methods $\mathcal{S}$, $\tilde{\mathcal{S}}$ whose iteration matrices are S, $\tilde{S}$, respectively. If there exists three real positive numbers α, β, γ such that*

$$\|S^{\nu} x\|_A^2 \le \|x\|_A^2 - \alpha \|S^{\nu} x\|_{A^2}^2 \quad \forall x \in \mathbb{C}^n \qquad \text{(pre-smoothing property)} \tag{1.39}$$

$$\|\tilde{S}^{\tilde{\nu}} x\|_A^2 \le \|x\|_A^2 - \beta_i \|x\|_{A^2}^2 \quad \forall x \in \mathbb{C}^n \qquad \text{(post-smoothing property)} \tag{1.40}$$

$$\min_{y \in \mathbb{C}^{\tilde{n}}} \|x - Py\| \le \gamma \|x\|_A \quad \forall x \in \mathbb{C}^n \qquad \text{(approximation property)} \tag{1.41}$$

defined

$$\delta_{\text{pre}} := \frac{\alpha}{\gamma}, \quad \delta_{\text{post}} := \frac{\beta}{\gamma},$$

it holds that $\delta_{\text{post}} \leq 1$ *and*

$$\|TGM\|_A \leq \sqrt{\frac{1-\delta_{\text{post}}}{1+\delta_{\text{pre}}}} < 1,$$

where $TGM = S^\nu CGC\tilde{S}^{\tilde{\nu}}$, *with* $CGC = I_n - P\tilde{A}^{-1}P^*A$ *is the two-grid iteration matrix.*

From Theorem 12, the convergence of the TGM is optimal, that is, the method converges with a constant error reduction not depending on n.

Remark 7. Note that, as is often the case in literature, a multigrid process can be definite by introducing first the two-grid algorithm and then by recursively solving the coarse-grid equation with this two-grid process.

1.8.2 Algebraic multigrid for Toeplitz matrices

AMG for multilevel Toeplitz matrices with scalar-valued symbols has been investigated in detail in [72, 140, 128, 4, 1]. To guarantee the optimal convergence of the AMG method, certain smoothing and approximation properties have to be fulfilled. As observed in Remark 6, they can be treated separately. We start with the conditions that ensure the validity of smoothing properties. First of all, note that since the diagonal of the coefficient matrix is a multiple of the identity, for Toeplitz matrices the relaxed Richardson in (1.35) are equivalent to the weighted Jacobi iterations. As the following proposition shows, using appropriate weights, these smoothers fulfill the smoothing properties.

Proposition 3. [1] *Let* $A = T_n(f)$ *with* $f : I_k \to \mathbb{C}$ *nonnegative and not identically zero. Defined* $S = I - \omega A$ *and* $\tilde{S} = I - \tilde{\omega}A$, *if*

$$0 \leq \omega, \tilde{\omega} \leq \frac{2}{\|f\|_{L^\infty}}, \tag{1.42}$$

then there exist $\alpha, \beta > 0$ *such that the two smoothing properties*

$$\|S^\nu x\|_A^2 \leq \|x\|_A^2 - \alpha\|S^\nu x\|_{A^2}^2 \quad \forall x \in \mathbb{C}^{\widehat{n}}, \tag{1.43}$$

$$\|\tilde{S}^\theta x\|_A^2 \leq \|x\|_A^2 - \beta\|x\|_{A^2}^2 \quad \forall x \in \mathbb{C}^{\widehat{n}}, \tag{1.44}$$

hold with $\nu, \theta \in \mathbb{N}$.

Apart from the smoother, to define an optimal AMG method, the choice of the projector is crucial. Assume that f has only one zero of order at most 2. Under this hypothesis, in the 1-level case, the different strategies proposed in [71, 33, 95, 128, 4] are equivalent. For a fixed $N_i = 2^{t-i} - 1$, with t, i integer numbers such that $i < t$, chose as projectors the product between $T_{N_i}(p_i)$, with a nonnegative trigonometric polynomial p_i and the transpose of the following cutting matrix $K_{N_i} \in \mathbb{R}^{N_{i+1} \times N_i}$

$$K_{N_i} = \begin{bmatrix} 0 & 1 & 0 & & & & & \\ & & 0 & 1 & 0 & & & \\ & & & \ddots & \ddots & \ddots & & \\ & & & & & 0 & 1 & 0 \end{bmatrix}. \tag{1.45}$$

When $N_i = 2^{t-i}$, the cutting matrix K_{N_i} is given by

$$K_{N_i} = \begin{bmatrix} 1 & 0 & & & & & & \\ & & 1 & 0 & & & & \\ & & & & \ddots & \ddots & & \\ & & & & & & 1 & 0 \end{bmatrix}. \tag{1.46}$$

Then the matrix at the so-defined coarse level is still a Toeplitz matrix $A_{i+1} = P_i^* A_i P_i = T_{N_{i+1}}(f_{i+1})$, where

$$f_{i+1}(\theta) = \frac{1}{2}\left[p_i^2 f_i\left(\frac{\theta}{2}\right) + p_i^2 f_i\left(\frac{\theta}{2} + \pi\right)\right].$$

The k-level case has been discussed in [72, 140, 128, 1]. There, the authors consider k-level Toeplitz matrices $T_n(f)$ of order $n = (2^t - 1)e \in \mathbb{N}^k$ and of order $n = 2^t e \in \mathbb{N}^k$, where $e = (1, \ldots, 1) \in \mathbb{N}^k$ and t is a positive integer. Starting with $N_0 = n$, the order on the coarse levels is defined as $N_i = (2^{t-i} - 1)e$ or $N_i = (2^{t-i})e$, respectively. The projector P_i is defined as the product between a matrix $T_{N_i}(p_i)$

and the transpose of a k-level cutting matrix K_{N_i}, where p_i is a nonnegative k-variate trigonometric polynomial. More precisely, the cutting matrix K_{N_i} is obtained as the Kronecker product of k 1-level cutting matrices $K_{(N_i)_1}, \ldots, K_{(N_i)_k}$ (with $K_{(N_i)_\ell}$ as in (1.45) for $(N_i)_\ell$ odd and as is (1.46) for $(N_i)_\ell$ even). If f has only one zero of order at most 2 in every direction, all these choices preserve the k-level Toeplitz structure on the coarse levels.

Multigrid methods are widely used for solving PDEs, independently of the boundary conditions. In that context, the convergence analysis is performed by Local Fourier analysis (LFA) which does not consider the boundary effects, i.e. it assumes periodic boundary conditions or an infinite domain [25]. Analogously, since Toeplitz matrices are difficult to manipulate, multigrid convergence results are usually investigated using matrix algebra approximations such as τ or circulant matrices having the same symbol, i.e. spectral distribution, as the original Toeplitz matrix (see [49] for more details).

In the following, we recall the conditions that ensure the validity of the approximation property in the circulant algebra. For a fixed $\theta \in \mathbb{R}^k$, define the set of all *corners points* $\Omega(\theta)$ and the set of all *mirror points* $\mathcal{M}(\theta)$ as

$$\Omega(\theta) = \{\eta | \eta_j \in \{\theta_j, \theta_j + \pi\}\} \text{ and } \mathcal{M}(\theta) = \Omega(\theta) \backslash \{\theta\}.$$

Let $\{f_i\}_{i=0}^m$ be the sequence of symbols on the coarse levels, where

$$f_{i+1}(\theta) = \frac{1}{2^k} \sum_{\eta \in \Omega(\theta)} p_i^2 f_i(\eta). \tag{1.47}$$

Proposition 4. [1] *Let $A_i = C_{N_i}(f_i)$ with $N_i = 2^{t-i}e$, and f_i a k-variate nonnegative trigonometric polynomial. Let θ_i^0 be the unique zero of f_i in $[0,\pi]^k$ and let $CGC_i = I_{\widehat{N}_i} - P_i(P_i^* A_i P_i)^{-1} P_i^* A_i$ with $P_i = C_{N_i}(p_i) K_{N_i}^T$ and p_i a nonnegative k-variate trigonometric polynomial. Then the approximation property* (1.38) *holds if for all $\theta \in [0,\pi]^k$ p_i is such that*

$$\limsup_{\theta \to \theta_i^0} \left| \frac{p_i(\eta)}{f_i(\theta)} \right| < +\infty, \quad \eta \in \mathcal{M}(\theta), \tag{1.48}$$

$$\sum_{\eta \in \Omega(\theta)} p_i^2(\eta) > 0. \tag{1.49}$$

Remark 8. If θ_i^0 is a zero of order q for f_i, by condition (1.48), $\eta \in \mathcal{M}(\theta_i^0)$ is such that p_i has a zero at η at least of the same order. From conditions (1.48)–(1.49) holds that $p_i(\theta_i^0) \neq 0$, which means that the projectors are full-rank and that the ill-conditioned subspace of A_i is in the image of the projector P_i.

Proposition 5. *Let f_{i+1} be defined as in* (1.47)*, and suppose that p_i satisfies* (1.48)-(1.49)*. Then, if θ_i^0 is a zero of order q for f_i, $\theta_{i+1}^0 = 2\theta_i^0$ is a zero of order q for f_{i+1}.*

If θ_i^0 has order (at most) $2q$, a natural choice for p_i is

$$p_i(\theta) = c \cdot \prod_{j=1}^{k} [1 + \cos(\theta_j - (\theta_i^0)_j)]^q. \tag{1.50}$$

with c constant. Indeed, the polynomial p_i has a zero of order $2q$ at $\eta \in \mathcal{M}(\theta_i^0)$ and does not vanish at θ_i^0.

For the TGM, the smoothing properties are satisfied under the same condition (1.42), while for the approximation property the following proposition holds ([33, 128]).

Proposition 6. *Let $A = T_n(f)$ with $n = (2^t - 1)e$, f a k-variate nonnegative trigonometric polynomial, and let $\tilde{n} = (\tilde{n}_1, \ldots, \tilde{n}_k) < (n_1, \ldots, n_k)$. Let θ^0 be the unique zero of f in $[0,\pi]^k$ of order at most 2, and let $CGC = I_{\widehat{n}} - P(P^* A P)^{-1} P^* A$ with $P = T_n(p) K_n^T$, $K_n^T = K_{n_1} \otimes \cdots \otimes K_{n_k}$, $K_{n_j} \in \mathbb{R}^{\tilde{n}_j \times n_j}$ and p as in* (1.50)*. Then the approximation property* (1.41) *holds if for all $\theta \in [0,\pi]^k$ p is such that*

$$\limsup_{\theta \to \theta^0} \frac{|p(\eta)|^2}{f(\theta)} < +\infty, \quad \eta \in \mathcal{M}(\theta), \tag{1.51}$$

$$\sum_{\eta \in \Omega(\theta)} p^2(\eta) > 0.$$

Remark 9. If θ^0 is a zero of order q for f, by condition (1.51), $\eta \in \mathcal{M}(\theta^0)$ is such that p has a zero at η at least of order $\lfloor q/2 \rfloor$.

1.9 Trigonometric polynomials

We conclude this chapter by recalling the definition of k-variate complex-/matrix-valued trigonometric polynomial and by proving a proposition needed in Chapter 2.

Definition 14. We say that $g : \mathbb{C}^k \to \mathbb{C}$ is a k-variate trigonometric polynomial if g is a finite linear combination of the k-variate functions (Fourier frequencies) $\{\mathrm{e}^{\mathbf{i}\langle j,x\rangle} : j \in \mathbb{Z}^k\}$.

Let us observe that, if g is a k-variate trigonometric polynomial, then g has only a finite number of nonzero Fourier coefficients $\widehat{g}_j$.
The degree $r = (r_1, \ldots, r_k)$ of a k-variate trigonometric polynomial g is defined as follows: for each $i = 1, \ldots, k$, r_i is the maximum of $|j_i|$, where $j = (j_1, \ldots, j_k)$ varies among all multi-indices in $\mathbb{Z}^k$ such that $\widehat{g}_j \neq 0$ (r_i is called the degree of $g(x)$ with respect to the i-th variable x_i).
Observe that a k-variate trigonometric polynomial g of degree $r = (r_1, \ldots, r_k)$ can be written in the form $g(x) = \sum_{j=-r}^{r} \widehat{g}_j \mathrm{e}^{\mathbf{i}\langle j,x\rangle}$, that is, defined $\Pi_r = \mathrm{span}\{\mathrm{e}^{\mathbf{i}\langle j,x\rangle}, j = -r, \ldots, r\}$, $g \in \Pi_r$.

Definition 15. We say that $g : \mathbb{C}^k \to \mathcal{M}_s$ is a *k-variate trigonometric polynomial* if, equivalently:

- all the components $g_{l,t} : \mathbb{C}^k \to \mathbb{C}$, $l, t = 1, \ldots, s$, are k-variate trigonometric polynomials.
- g is a finite linear combination (with coefficients in $\mathcal{M}_s$) of the k-variate functions $\{\mathrm{e}^{\mathbf{i}\langle j,x\rangle} : j \in \mathbb{Z}^k\}$.

If g is a k-variate trigonometric polynomial, then g has only a finite number of nonzero Fourier coefficients $\widehat{g}_j \in \mathcal{M}_s$ and the degree $r = (r_1, \ldots, r_k)$ of g is defined in two equivalent ways:

- for each $i = 1, \ldots, k$, r_i is the maximum degree among all the polynomials $g_{l,t}(x)$ with respect to the i-th variable x_i;
- for each $i = 1, \ldots, k$, r_i is the maximum of $|j_i|$, where $j = (j_1, \ldots, j_k)$ varies among all multi-indices in $\mathbb{Z}^k$ such that $\widehat{g}_j$ is nonzero.

We note that a k-variate trigonometric polynomial g of degree $r = (r_1, \ldots, r_k)$ can be written in the form $g(x) = \sum_{j=-r}^{r} \widehat{g}_j \mathrm{e}^{\mathbf{i}\langle j,x\rangle}$, where the Fourier coefficients $\widehat{g}_j$ belong to $\mathcal{M}_s$.

Proposition 7 provides an estimate of the rank of $T_n(g)T_n(f) - T_n(gf)$, in the case where $f \in L^1(k, s)$ and g is a k-variate trigonometric polynomial of degree $r = (r_1, \ldots, r_k)$ taking values in $\mathcal{M}_s$. For $s = 1$, we can find the proof of this result (full for $k = 1$ and sketched for $k > 1$) in [134]. For completeness, we report the full proof for $k > 1$, also considering the generalization to $s > 1$.

Proposition 7. *Let $f, g \in L^1(k, s)$, with g a k-variate trigonometric polynomial of degree $r = (r_1, \ldots, r_k)$, and let n be a k-index such that $n \geq 2r + e$. Then*

$$\mathrm{rank}(T_n(g)T_n(f) - T_n(gf)) \leq s \left[\widehat{n} - \prod_{i=1}^{k} (n_i - 2r_i) \right]. \tag{1.52}$$

Proof. Since $g : \mathbb{C}^k \to \mathcal{M}_s$ is a k-variate trigonometric polynomial of degree r, we can write g in the form

$$g(x) = \sum_{j=-r}^{r} \widehat{g}_j \mathrm{e}^{\mathbf{i}\langle j,x\rangle}.$$

The Fourier coefficients of $(gf)(x) = g(x)f(x)$ are given by

$$(\widehat{gf})_\ell = \frac{1}{(2\pi)^k} \int_{I_k} g(x)f(x)\mathrm{e}^{-\mathbf{i}\langle \ell,x\rangle} dx = \sum_{j=-r}^{r} \widehat{g}_j \frac{1}{(2\pi)^k} \int_{I_k} f(x)\mathrm{e}^{-\mathbf{i}\langle \ell-j,x\rangle} dx = \sum_{j=-r}^{r} \widehat{g}_j \widehat{f}_{\ell-j}.$$

Now, using the definition of multilevel block Toeplitz matrices, see (1.16), for all $l, t = e, \ldots, n$ we have

$$T_n(gf)_{l,t} = (\widehat{gf})_{l-t} = \sum_{j=-r}^{r} \widehat{g}_j \widehat{f}_{l-t-j}, \tag{1.53}$$

and

$$(T_n(g)T_n(f))_{l,t} = \sum_{v=e}^{n} T_n(g)_{l,v} T_n(f)_{v,t} = \sum_{v=e}^{n} \widehat{g}_{l-v} \widehat{f}_{v-t} = \sum_{j=l-n}^{l-e} \widehat{g}_j \widehat{f}_{l-j-t} = \sum_{j=\max(l-n,-r)}^{\min(l-e,r)} \widehat{g}_j \widehat{f}_{l-t-j}, \tag{1.54}$$

where the last equality is motivated by the fact that $\widehat{g}_j$ is zero if $j < -r$ or $j > r$. Therefore, (1.53) and (1.54) coincide when $r + e \leq l \leq n - r$. Observe that the multi-index range $r + e, \ldots, n - r$ is nonempty because of the assumption $n \geq 2r + e$. We conclude that the only possible nonzero rows of $T_n(g)T_n(f) - T_n(gf)$ are those corresponding to multi-indices l in the set $\{e, \ldots, n\} \backslash \{r + e, \ldots, n - r\}$. This set has cardinality $\widehat{n} - \prod_{i=1}^{k}(n_i - 2r_i)$ and so $T_n(g)T_n(f) - T_n(gf)$ has at most $\widehat{n} - \prod_{i=1}^{k}(n_i - 2r_i)$ nonzero rows. Now we should notice that each row of $T_n(g)T_n(f) - T_n(gf)$ is actually a block-row of size s, i.e., a $s \times s\widehat{n}$ submatrix of $T_n(g)T_n(f) - T_n(gf)$. Indeed, each component of $T_n(f)$, $T_n(g)$, $T_n(g)T_n(f) - T_n(gf)$ is actually a $s \times s$ matrix, see (1.16). Therefore, the actual nonzero rows of $T_n(g)T_n(f) - T_n(gf)$ are at most $s[\widehat{n} - \prod_{i=1}^{k}(n_i - 2r_i)]$ and (1.52) is proved. □

Chapter 2

Spectral analysis and preconditioning for non-Hermitian multilevel block Toeplitz matrices

The problem considered throughout this chapter is the numerical solution of a linear system with coefficient matrix $T_n(f)$, where $f \in L^1(k,s)$. To do that, we enlarge the known distribution results presented in Chapter 1 and take advantage of the resulting tools for designing appropriate preconditioners for multilevel block Toeplitz matrices. More in detail, in Section 2.1, we extend Theorem 7 to the case of preconditioned matrix-families of the form $\{T_n^{-1}(g)T_n(f)\}_{n\in\mathbb{N}^k}$ (Theorem 13) and exploit the knowledge of the obtained spectral information as a guide for defining a satisfactorily band Toeplitz preconditioner $T_n(g)$ for $T_n(f)$ to be used in connection with Krylov methods. In Section 2.2 we provide a further extension of Theorem 13 to sequences of matrices obtained from a combination of some algebraic operations on multilevel block Toeplitz matrices (Theorems 15,17). These new results will be used to perform a spectral analysis of the PHSS method applied to Toeplitz matrices and then to find efficient PHSS preconditioners.

2.1 Band Toeplitz preconditioning for multilevel block Toeplitz sequences

We are interested in preconditioning a non-Hermitian multilevel block Toplitz $T_n(f)$ by $T_n(g)$, where g is a trigonometric polynomial, that is by a band Toeplitz preconditioner. As observed in Section 1.6, the literature of the band Toeplitz preconditioning lacks of results in non-Hermitian multilevel block case. Actually, at our best knowledge, Theorem 4 is the only available tool for devising spectrally equivalent preconditioners in the non-Hermitian multilevel block case (see Subsection 1.4.1). Even though, Theorem 4 is rather powerful, its assumptions do not seem easy to check. More precisely, a set of important problems to be considered for the practical use of Theorem 4 is the following:

(a) given f regular enough, give conditions such that there exists a trigonometric polynomial g for which the assumptions of Theorem 4 are satisfied;

(b) let us suppose that f satisfies the conditions of the first item; give a constructive way (an algorithm) for defining such a polynomial g.

Due to the difficulty of addressing these problems, instead of looking for a precise spectral localization of $T_n^{-1}(g)T_n(f)$ with g trigonometric polynomial, we just analyze the global asymptotic behavior of the spectrum of $T_n^{-1}(g)T_n(f)$ as $n \to \infty$. As we shall see through numerical experiments, the knowledge of the asymptotic spectral distribution of $\{T_n^{-1}(g)T_n(f)\}_{n\in\mathbb{N}^k}$ can indeed be useful as a guide for designing appropriate preconditioners $T_n(g)$ for $T_n(f)$.

2.1.1 Spectral distribution for preconditioned non-Hermitian families

In the following, we generalize Theorem 7, which concerns the non-preconditioned matrix-family $\{T_n(f)\}_{n\in\mathbb{N}^k}$, to the case of preconditioned matrix-families of the form $\{T_n^{-1}(g)T_n(f)\}_{n\in\mathbb{N}^k}$.

Theorem 13. *Let $f, g \in L^\infty(k,s)$, with $0 \notin \mathrm{Coh}[\mathcal{ENR}(g)]$, and let $h := g^{-1}f$. If $\mathcal{ER}(h)$ has empty interior and does not disconnect the complex plane, that is h is in the Tilli class, then $\{T_n^{-1}(g)T_n(f)\}_{n\in\mathbb{N}^k} \sim_\lambda (h, I_k)$.*

Before to prove Theorem 13, we need two lemmas which involve the notion of sectorial function. Lemma 1 provides simple conditions that ensure a given function $f \in L^1(k,s)$ to be weakly sectorial or sectorial. The proof can be obtained using the following topological properties of convex sets: the separability properties provided by the geometric forms of the Hahn-Banach theorem, see [27, Theorems 1.6 and 1.7]; the result stating that, for any convex set C, the closure $\overline{C}$ and the interior $\mathrm{Int}(C)$ are convex, and $\overline{\mathrm{Int}(C)} = \overline{C}$ whenever $\mathrm{Int}(C)$ is nonempty, see e.g. [27, Exercise 1.7]. Lemma 2 shows under which condition the function $h = g^{-1}f$ belongs to $L^\infty(k,s)$.

Lemma 1. *Let $f \in L^1(k,s)$.*

- *f is weakly sectorial if and only if $0 \notin \mathrm{Int}(\mathrm{Coh}[\mathcal{ENR}(f)])$.*
- *If $0 \notin \overline{\mathrm{Coh}[\mathcal{ENR}(f)]}$ then f is sectorial. Equivalently, if $d(\mathrm{Coh}[\mathcal{ENR}(f)], 0) > 0$, then f is sectorial.*

Note that, for a function $g \in L^\infty(k,s)$, the essential numerical range $\mathcal{ENR}(g)$ is compact and hence $\mathrm{Coh}[\mathcal{ENR}(g)]$ is also compact (we recall that the convex hull of a compact set is compact). Moreover, if $0 \notin \mathrm{Coh}[\mathcal{ENR}(g)]$, $d := d(\mathrm{Coh}[\mathcal{ENR}(g)], 0)$ is positive and, from previous Lemma 1, g is sectorial. The condition $0 \notin \mathrm{Coh}[\mathcal{ENR}(g)]$ also ensures that g is invertible a.e., because, for almost every $x \in I_k$, $\lambda_i(g(x)) \in \mathcal{ER}(g) \subseteq \mathcal{ENR}(g) \subseteq \mathrm{Coh}[\mathcal{ENR}(g)]$ for all $i = 1, \ldots, s$, implying that $\lambda_i(g) \neq 0$ for all $i = 1, \ldots, s$, a.e.

Lemma 2. *Let $f, g \in L^\infty(k,s)$ with $0 \notin \mathrm{Coh}[\mathcal{ENR}(g)]$. Then $h := g^{-1}f \in L^\infty(k,s)$.*

Proof. Since $g \in L^\infty(k,s)$ and $0 \notin \mathrm{Coh}[\mathcal{ENR}(g)]$, the convex hull $\mathrm{Coh}[\mathcal{ENR}(g)]$ is compact, the distance $d := d(\mathrm{Coh}[\mathcal{ENR}(g)], 0)$ is positive, and g is invertible a.e. (recall the discussion before this lemma). We are going to show that

$$\|g^{-1}(x)\| \leq \frac{1}{d}, \quad \text{for a.e. } x \in I_k. \tag{2.1}$$

Since in a matrix the absolute value of each component is bounded from above by the spectral norm, once we have proved (2.1), it follows that $g^{-1} \in L^\infty(k,s)$, and the lemma is proved. Now, by the fact that $d > 0$ and by Lemma 1, g is sectorial. By Theorem 3, first item, there exists a separating line z for $\mathcal{ENR}(g)$ such that $d(z,0) = d$. Let $\omega(z)$ be the rotation number (of modulus 1) for which $\omega(z) \cdot z = \{w \in \mathbb{C} : \mathrm{Re}(w) = d\}$. By applying Remark 1 with g in place of f (see in particular (1.10) and note that $\omega_z = \omega(z)$ because $d(z,0) = d > 0$), we obtain

$$\lambda_{\min}(\mathrm{Re}(\omega(z)g(x))) \geq d, \quad \text{for a.e. } x \in I_k.$$

Hence, by the Fan-Hoffman theorem [19, Proposition III.5.1], for a.e. $x \in I_k$ we have

$$\|g^{-1}(x)\| = \frac{1}{\sigma_{\min}(g(x))} = \frac{1}{\sigma_{\min}(\omega(z)g(x))} \leq \frac{1}{\lambda_{\min}(\mathrm{Re}(\omega(z)g(x)))} \leq \frac{1}{d},$$

and (2.1) is proved. □

For the sake of clarity, we summarize the results of Lemma 1, Theorem 3, the proof of Lemma 2, and the discussion between Lemma 1 and Lemma 2, in Lemma 3. It will turn out to be useful even in next section.

Lemma 3. *Let $g \in L^\infty(k,s)$ with $0 \notin \mathrm{Coh}[\mathcal{ENR}(g)]$. Then,*

- *$\mathrm{Coh}[\mathcal{ENR}(g)]$ is compact, implying that the distance $d := d(\mathrm{Coh}[\mathcal{ENR}(g)], 0)$ is positive;*
- *g is sectorial and invertible a.e., and $g^{-1} \in L^\infty(k,s)$, because $\|g^{-1}(x)\| \leq \frac{1}{d}$ for a.e. $x \in I_k$;*
- *$T_n(g)$ is invertible for all $n \in \mathbb{N}^k$, because $\sigma \geq d$ for every singular values σ of $T_n(g)$;*
- *$\|T_n^{-1}(g)\| \leq \frac{1}{d}$ for all $n \in \mathbb{N}^k$ (consequence of the previous item).*

Aside from previous lemmas, one more result contained in Proposition 8 is needed. It concerns the evaluation of the trace-norm of $T_n(g)T_n(f) - T_n(gf)$ for $f, g \in L^\infty(k,s)$, which is a crucial point for the proof of Theorem 13. As for Proposition 7, fixed $s = 1$, we can find the proof of this result (full for $k = 1$ and sketched for $k > 1$) in [134]. In the following, we report the full proof for $k > 1$, also considering the generalization to $s > 1$.

Proposition 8. *Let $f, g \in L^\infty(k,s)$, then $\|T_n(g)T_n(f) - T_n(gf)\|_1 = o(\widehat{n})$ as $n \to \infty$.*

Proof. Let $g_m : \mathbb{C}^k \to \mathcal{M}_s$, $g_m = [(g_m)_{l,t}]_{l,t=1}^s$, be a k-variate trigonometric polynomial of degree $m = (m_1, \ldots, m_k)$. Let $m^- := (m_1^-, \ldots, m_k^-)$, where m_i^- is the minimum degree among all the polynomials $(g_m)_{l,t}(x)$ with respect to the variable x_i. We choose g_m such that $\|g_m\|_{L^\infty} \le \|g\|_{L^\infty}$ for every m and $\|g_m - g\|_{L^1} \to 0$ as $m^- \to \infty$. The polynomials g_m can be constructed by using the m-th Cesaro sum of g (see [156]) and indeed the linear positive character of the Cesaro operator and Korovkin theory [98, 125] imply the existence of a g_m with the desired properties. Note that, by (1.5) with $p = 1$, the fact that $\|g_m - g\|_{L^1} \to 0$ as $m^- \to \infty$ is equivalent to saying that $\|(g_m)_{l,t} - g_{l,t}\|_{L^1} \to 0$ as $m^- \to \infty$ for all $l, t = 1, \ldots, s$. Now, by adding and subtracting and by using the triangle inequality several times we get

$$\begin{aligned}&\|T_n(g)T_n(f) - T_n(gf)\|_1\\ &\le \|T_n(g)T_n(f) - T_n(g_m)T_n(f)\|_1 + \|T_n(g_m)T_n(f) - T_n(g_m f)\|_1 + \|T_n(g_m f) - T_n(gf)\|_1. \end{aligned} \tag{2.2}$$

Using the linearity of the operator $T_n(\cdot)$, the Hölder inequalities (1.2) and (1.8), and the relations (1.17), (1.18), we obtain

$$\|T_n(g)T_n(f) - T_n(g_m)T_n(f)\|_1 \le \|T_n(g - g_m)\|_1 \|T_n(f)\| \le \frac{\widehat{n}}{(2\pi)^k} \|g_m - g\|_{L^1} \|f\|_{L^\infty} \tag{2.3}$$

$$\|T_n(g_m f) - T_n(gf)\|_1 \le \frac{\widehat{n}}{(2\pi)^k} \|g_m f - gf\|_{L^1} \le \frac{\widehat{n}}{(2\pi)^k} \|g_m - g\|_{L^1} \|f\|_{L^\infty}. \tag{2.4}$$

Moreover, using the relation (1.4) and the inequality $(1 + c)^k \ge 1 + kc$ for $c \ge -1$, Proposition 7 tells us that, for any $n \ge 2m + e$,

$$\begin{aligned}\|T_n(g_m)T_n(f) - T_n(g_m f)\|_1 &\le \operatorname{rank}(T_n(g_m)T_n(f) - T_n(g_m f))\|T_n(g_m)T_n(f) - T_n(g_m f)\|\\ &\le s\widehat{n}\left[1 - \prod_{i=1}^k \left(1 - \frac{2m_i}{n_i}\right)\right](\|T_n(g_m)T_n(f)\| + \|T_n(g_m f)\|)\\ &\le s\widehat{n}\left[1 - \left(1 - \frac{2\|m\|_\infty}{\min_j n_j}\right)^k\right](2\|g_m\|_{L^\infty}\|f\|_{L^\infty})\\ &\le s\widehat{n}k\frac{2\|m\|_\infty}{\min_j n_j}(2\|g\|_{L^\infty}\|f\|_{L^\infty}) = 4sk\|m\|_\infty\|g\|_{L^\infty}\|f\|_{L^\infty}\frac{\widehat{n}}{\min_j n_j}. \end{aligned} \tag{2.5}$$

Substituting (2.3)–(2.5) in (2.2), for each k-tuple m and for each $n \ge 2m + e$ the following inequality holds:

$$\|T_n(g)T_n(f) - T_n(gf)\|_1 \le \widehat{n}\xi(m) + \gamma(m)\frac{\widehat{n}}{\min_j n_j},$$

where $\xi(m) := 2(2\pi)^{-k}\|g_m - g\|_{L^1}\|f\|_{L^\infty}$, $\gamma(m) := 4sk\|m\|_\infty\|g\|_{L^\infty}\|f\|_{L^\infty}$, and we note that $\xi(m) \to 0$ as $m^- \to \infty$. Now, for $\epsilon > 0$, we choose a k-tuple m such that $\xi(m) < \epsilon/2$. For $n \to \infty$ (i.e. for $\min_j n_j \to \infty$) we have $\gamma(m)/\min_j n_j \to 0$ and so we can choose a $\nu \ge 2\|m\|_\infty + 1$ such that $\gamma(m)/\min_j n_j \le \epsilon/2$ for $\min_j n_j \ge \nu$. Then, if $\min_j n_j \ge \nu$, we have $n \ge 2m + e$ and

$$\frac{\|T_n(g)T_n(f) - T_n(gf)\|_1}{\widehat{n}} \le \epsilon.$$

This means that $\dfrac{\|T_n(g)T_n(f) - T_n(gf)\|_1}{\widehat{n}} \to 0$ as $n \to \infty$, i.e. $\|T_n(g)T_n(f) - T_n(gf)\|_1 = o(\widehat{n})$ as $n \to \infty$. □

We are now ready to prove Theorem 13. We will show that, under the assumptions of Theorem 13, the conditions **(c_1)**–**(c_4)** of Theorem 1 are met, for any matrix-sequence $\{T_{n(m)}^{-1}(g)T_{n(m)}(f)\}_m$ extracted from $\{T_n^{-1}(g)T_n(f)\}_{n\in\mathbb{N}^k}$ and such that $\min_j n_j(m) \to \infty$. Actually, to simplify the notation, we suppress the index m and we will talk about a generic matrix-sequence $\{T_n^{-1}(g)T_n(f)\}_n$ such that $\min_j n_j \to \infty$, where it is understood the presence of an underlying index m.

Proof of Theorem 13. By Lemma 3, $d := d(\mathrm{Coh}[\mathcal{ENR}(g)], 0)$ is positive and

$$\|T_n^{-1}(g)\| = \frac{1}{\sigma_{\min}(T_n(g))} \leq \frac{1}{d}.$$

By hypothesis $f \in L^\infty(k,s)$ and by (1.18), it follows that

$$\|T_n^{-1}(g)T_n(f)\| \leq \|f\|_{L^\infty}/d,$$

so that requirement $(\boldsymbol{c_1})$ in Theorem 1 is satisfied. Since $h \in L^\infty(k,s)$ (by Lemma 3) and since $\mathcal{ER}(h)$ has empty interior and does not disconnect the complex plane (by hypothesis), h satisfies the assumptions of Theorem 7 and so $\{T_n(h)\}_n \sim_\lambda (h, I_k)$. The latter implies that

$$\lim_{n\to\infty} \frac{\mathrm{tr}(T_n(h)^N)}{\widehat{n}} = \frac{1}{(2\pi)^k} \int\limits_{I_k} \frac{\mathrm{tr}(h^N(t))}{s} dt. \tag{2.6}$$

Indeed, although the function $\tilde{F}(z) = z^N$ does not belong to $\mathcal{C}_0(\mathbb{C})$, we know that $h \in L^\infty(k,s)$. This has two consequences: first, by (1.18) all the eigenvalues of $T_n(h)$ lie in the compact disk $\overline{D(0, \|h\|_{L^\infty})}$; second, by definition of $\|h\|_{L^\infty}$, also the eigenvalues of $h(t)$ belong to $\overline{D(0, \|h\|_{L^\infty})}$ for a.e. $t \in I_k$. Therefore, by choosing any function $F \in \mathcal{C}_0(\mathbb{C})$ such that $F(z) = z^N$ for all $z \in \overline{D(0, \|h\|_{L^\infty})}$, from $\{T_n(h)\}_n \sim_\lambda (h, I_k)$ we get (2.6). By (2.6) and by using the inequality (1.3), item $(\boldsymbol{c_2})$ in Theorem 1 is proved if we show that

$$\|(T_n^{-1}(g)T_n(f))^N - T_n(h)^N\|_1 = o(\widehat{n}) \tag{2.7}$$

for every nonnegative integer N. If $N = 0$ the result is trivial. For $N = 1$, using Proposition 8 we obtain

$$\begin{aligned} \|T_n^{-1}(g)T_n(f) - T_n(h)\|_1 &= \|T_n^{-1}(g)(T_n(f) - T_n(g)T_n(h))\|_1 \\ &\leq \|T_n^{-1}(g)\|\|T_n(f) - T_n(g)T_n(h)\|_1 \\ &\leq \frac{1}{d}\|T_n(f) - T_n(g)T_n(h)\|_1 = o(\widehat{n}), \end{aligned}$$

so (2.7) is satisfied and we can write $T_n^{-1}(g)T_n(f) = T_n(h) + R_n$ with $\|R_n\|_1 = o(\widehat{n})$. Using this, when $N \geq 2$ we have

$$(T_n^{-1}(g)T_n(f))^N = T_n(h)^N + S_n,$$

where S_n is the sum of all possible (different) combinations of products of j matrices $T_n(h)$ and ℓ matrices R_n, with $j + \ell = N$, $j \neq N$. By using the Hölder inequality (1.2), and taking into account that $R_n = T_n^{-1}(g)T_n(f) - T_n(h)$, for every summand S of S_n we have

$$\begin{aligned} \|S\|_1 &\leq \|T_n(h)\|^j \|R_n\|^{\ell-1}\|R_n\|_1 \\ &\leq \|h\|_{L^\infty}^j (\|f\|_{L^\infty}/d + \|h\|_{L^\infty})^{\ell-1} o(\widehat{n}) \leq C o(\widehat{n}), \end{aligned}$$

where C is some positive constant. So, since the number of summands in S_n is finite, (2.7) holds for every positive integer N, and requirement $(\boldsymbol{c_2})$ in Theorem 1 is then satisfied. Requirement $(\boldsymbol{c_3})$ in Theorem 1 is also satisfied, because by **[GLT1-3]**, the sequences of multilevel block Toeplitz matrices with $L^1(k,s)$ symbols belong to the GLT class together with their algebra. Finally, by taking into account that $\mathcal{ER}(h)$ has empty interior and does not disconnect the complex plane, the last condition $(\boldsymbol{c_4})$ in Theorem 1 is met, and the application of Theorem 1 shows that $\{T_n^{-1}(g)T_n(f)\}_n \sim_\lambda (h, I_k)$. □

2.1.2 Numerical results

In the following we consider a list of numerical examples (both 1-level and 2-level) concerning the eigenvalue localization and the clustering properties of the matrix $T_n^{-1}(g)T_n(f)$, and regarding the effectiveness of the preconditioned GMRES, with preconditioning strategies chosen according to the theoretical indications given in the previous subsection.

Univariate examples

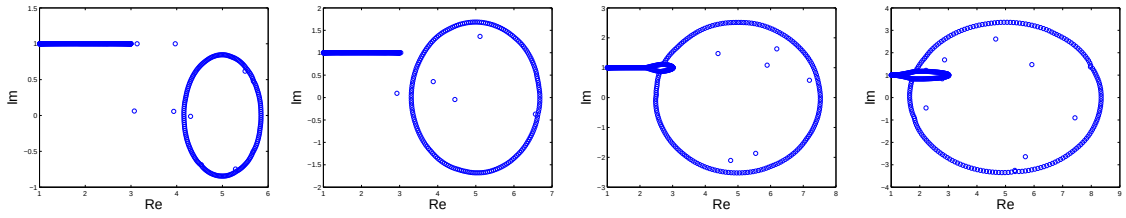
Fixed $s = 2$ and $k = 1$, we consider f and g of the form

$$\begin{aligned} f(x) &= Q(x)A(x)Q(x)^T \\ g(x) &= Q(x)B(x)Q(x)^T \end{aligned} \tag{2.8}$$

where

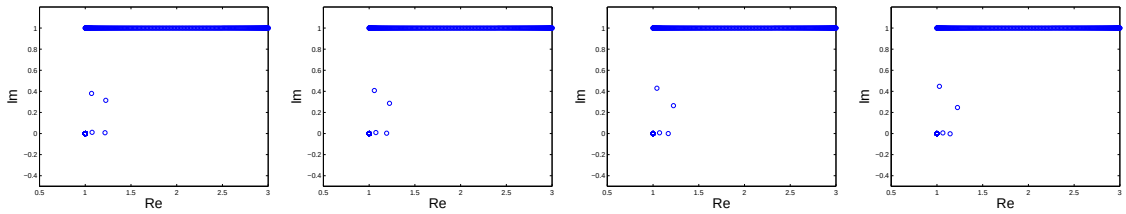
$$Q(x) = \begin{pmatrix} \cos(x) & \sin(x) \\ -\sin(x) & \cos(x) \end{pmatrix},$$

while $A(x)$ and $B(x)$ vary from case to case. For each example, we focus our attention on the spectral behavior of the matrices $T_n(f)$, $T_n^{-1}(g)T_n(f)$ for different sizes n and on the solution of the associated linear system with a random right-hand side. From a computational point of view, to solve such systems, we apply (full or preconditioned) GMRES with tolerance 10^{-6} using the Matlab built-in `gmres` function.



$r = 1$ $r = 2$ $r = 3$ $r = 4$

Figure 2.1: Eigenvalues in the complex plane of $T_{200}(f_1)$ for $r = 1, 2, 3, 4$



$r = 1$ $r = 2$ $r = 3$ $r = 4$

Figure 2.2: Eigenvalues in the complex plane of $T_{200}^{-1}(g_1)T_{200}(f_1)$ for $r = 1, 2, 3, 4$

Case 1. Let us choose $A^{(1)}(x)$ and $B^{(1)}(x)$ as follows

$$A^{(1)}(x) = \begin{pmatrix} 2 + \mathbf{i} + \cos(x) & 0 \\ 1 & 5 + re^{\mathbf{i}x} \end{pmatrix}, \quad B^{(1)}(x) = \begin{pmatrix} 1 & 0 \\ 0 & 5 + re^{\mathbf{i}x} \end{pmatrix},$$

where r is a real positive parameter, and define

$$f_1(x) = Q(x)A^{(1)}(x)Q(x)^T$$
$$g_1(x) = Q(x)B^{(1)}(x)Q(x)^T.$$

Figures 2.1 and 2.2 refer to the eigenvalues in the complex plane of $T_{200}(f_1)$ and $T_{200}^{-1}(g_1)T_{200}(f_1)$ for $r = 1, 2, 3, 4$. The eigenvalues of $T_{200}(f_1)$ are clustered at the union of the ranges of $\lambda_1(f_1(x)) = \lambda_1(A^{(1)}(x)) = A_{1,1}^{(1)}(x)$ and $\lambda_2(f_1(x)) = \lambda_2(A^{(1)}(x)) = A_{2,2}^{(1)}(x)$, which is the essential range of $f_1(x)$. More precisely, the matrix $T_{200}(f_1)$ has two sub-clusters for the eigenvalues: one collects the eigenvalues with real part in $[1, 3]$ and imaginary part around 1 (such eigenvalues recall the behavior of the function $\lambda_1(f_1(x))$); the other, mimicking $\lambda_2(f_1(x))$, is made by a circle centered in 5 with radius r, in agreement with theoretical results. We know that the GMRES in this case is optimal, since the eigenvalues of f_1 have no zeros. Looking at Figure 2.2, if we use the preconditioner $T_{200}(g_1)$, we improve the cluster of the eigenvalues and so the GMRES converges with a constant number of iterations (cf. Table 2.1), which is substantially independent both on n and r.

This example fits with the theoretical results of this section, since f_1 and g_1 are both bounded and $0 \notin \mathrm{Coh}[\mathcal{ENR}(g_1)]$. We stress that g_1 has been chosen so that the essential range of

$$h_1(x) = g_1^{-1}(x)f_1(x) = Q(x) \begin{pmatrix} 2 + \mathbf{i} + \cos(x) & 0 \\ 1/(5 + re^{\mathbf{i}x}) & 1 \end{pmatrix} Q(x)^T,$$

which is given by

$$\mathcal{ER}(h_1) = \mathcal{ER}(\lambda_1(h_1)) \bigcup \mathcal{ER}(\lambda_2(h_1)) = \{t + \mathbf{i} : 1 \le t \le 3\} \bigcup \{1\},$$

is 'compressed' and 'well separated from 0' independently of the value of r. In this way, since Theorem 13 and Remark 3 ensure that the matrix-sequence $\{T_n^{-1}(g_1)T_n(f_1)\}_n$ is weakly clustered at $\mathcal{ER}(h_1)$, we expect a number of preconditioned GMRES iterations independent of r, n and 'small enough'. This is confirmed by the results in Tables 2.1 and 2.2 (a).

Case 2. Let us choose $A^{(2)}(x)$ and $B^{(2)}(x)$ as follows

$$A^{(2)}(x) = \begin{pmatrix} 2+\mathbf{i}+\cos(x) & 0 \\ 1/(x^2-1) & 5+re^{\mathbf{i}x} \end{pmatrix}, \quad B^{(2)}(x) = B^{(1)}(x)$$

and define

$$f_2(x) = Q(x)A^{(2)}(x)Q(x)^T$$
$$g_2(x) = Q(x)B^{(2)}(x)Q(x)^T.$$

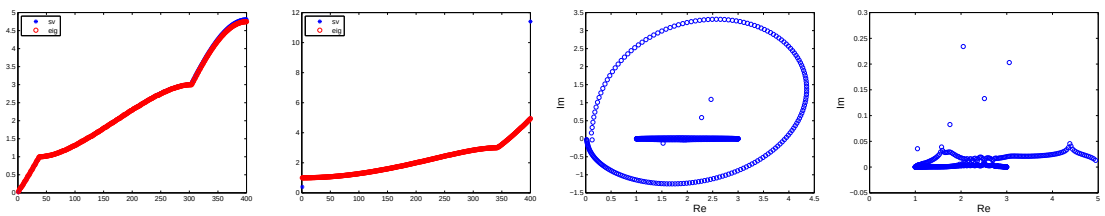
	Iterations	
n	No Prec.	Prec.
50	55	14
100	98	14
200	179	13
400	230	14
800	235	13

Table 2.1: Number of GMRES iterations for $T_n(f_1)$ and $T_n^{-1}(g_1)T_n(f_1)$ fixed $r = 4.8$ and varying n

	Iterations			
r	No Prec.	Prec.	No Prec.	Prec.
1	17	14	17	13
2	22	14	22	13
3	31	14	30	13
4	55	14	53	13
4.8	185	14	183	13
	(a)		(b)	

Table 2.2: (a) number of GMRES iterations for $T_{200}(f_1)$ and $T_{200}^{-1}(g_1)T_{200}(f_1)$ varying r; (b) number of GMRES iterations for $T_{200}(f_2)$ and $T_{200}^{-1}(g_2)T_{200}(f_2)$ varying r

Although this case is not covered by the theory, since f_2 is not bounded, we find that the eigenvalues of $T_{200}^{-1}(g_2)T_{200}(f_2)$ are closely related to the eigenvalues of $g_2^{-1}f_2$. The graphs of such eigenvalues (not reported here) are very similar to those in Figure 2.2. Table 2.2 (b) shows the number of GMRES iterations, setting $n = 200$ and moving the radius of the disk used to define f_2 and g_2.



(a) (b) (c) (d)

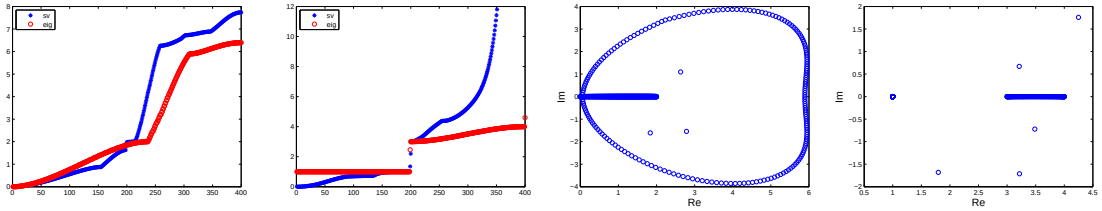
Figure 2.3: (a) Singular values and moduli of the eigenvalues of $T_{200}(f_3)$; (b) Singular values and moduli of the eigenvalues of $T_{200}^{-1}(g_3)T_{200}(f_3)$; (c) Eigenvalues in the complex plane of $T_{200}(f_3)$; (c) Eigenvalues in the complex plane of $T_{200}^{-1}(g_3)T_{200}(f_3)$

Case 3. Let us choose $A^{(3)}(x)$ and $B^{(3)}(x)$ as follows

$$A^{(3)}(x) = \begin{pmatrix} (1-\mathrm{e}^{\mathbf{i}x})\left(1+x^2/\pi^2\right) & 0 \\ 0 & 2+\cos(x) \end{pmatrix}, \quad B^{(3)}(x) = \begin{pmatrix} 1-\mathrm{e}^{\mathbf{i}x} & 0 \\ 0 & 1 \end{pmatrix}$$

and define

$$f_3(x) = Q(x)A^{(3)}(x)Q(x)^T$$
$$g_3(x) = Q(x)B^{(3)}(x)Q(x)^T.$$



(a) (b) (c) (d)

Figure 2.4: (a) Singular values and moduli of the eigenvalues of $T_{200}(f_4)$; (b) Singular values and moduli of the eigenvalues of $T_{200}^{-1}(g_4)T_{200}(f_4)$; (c) Eigenvalues in the complex plane of $T_{200}(f_4)$; (d) Eigenvalues in the complex plane of $T_{200}^{-1}(g_4)T_{200}(f_4)$

Figures 2.3 (a) and (b) show the singular values and the moduli of the eigenvalues of $T_{200}(f_3)$ and $T_{200}^{-1}(g_3)T_{200}(f_3)$, respectively. Let us observe that the singular values and the moduli of the eigenvalues are, for both matrices, almost superposed. Figure 2.3 (c) and (d) refer to the eigenvalues in the complex plane of the same matrices. As already argued for Case 1, even in this case the eigenvalues of $T_{200}(f_3)$ show two different behaviors: a half of the eigenvalues is clustered at $[1,3]$, which is the range of the function $\lambda_2(f_3(x)) = A_{2,2}^{(3)}(x)$, the others mimic $\lambda_1(f_3(x)) = A_{1,1}^{(3)}(x)$, drawing a circle passing near to 0. The closeness of the eigenvalues to 0 is responsible of the non-optimality of the GMRES method when we solve a linear system with matrix $T_{200}(f_3)$. Indeed, as can be observed in Table 2.3 (a), the number

	Iterations			
n	No Prec.	Prec.	No Prec.	Prec.
50	59	15	100	9
100	106	16	200	9
200	192	16	338	9
400	343	16	577	9

(a) (b)

Table 2.3: (a) Number of GMRES iterations for $T_n(f_3)$ and $T_n^{-1}(g_3)T_n(f_3)$ varying n; (b) Number of GMRES iterations for $T_n(f_4)$ and $T_n^{-1}(g_4)T_n(f_4)$ varying n

of GMRES iterations required to reach tolerance 10^{-6} increases with n for $T_n(f_3)$. The preconditioned matrix $T_n^{-1}(g_3)T_n(f_3)$ has eigenvalues far from 0 and bounded in modulus (see Figure 2.3 (d)) and so the preconditioned GMRES converges with a constant number of iterations (see Table 2.3 (a)). This example is not covered by the theory explained in previous subsection, since $\mathrm{Coh}[\mathcal{ENR}(g_3)]$ includes the complex zero. The numerical tests, however, show that there is room for improving the theory, by allowing the symbol of the preconditioner to have eigenvalues assuming zero value.

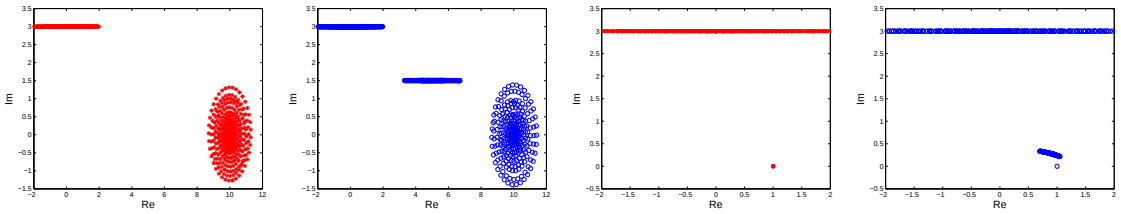
Case 4. Let us choose $A^{(4)}(x)$ and $B^{(4)}(x)$ as follows

$$A^{(4)}(x) = \begin{pmatrix} (1-\mathrm{e}^{\mathbf{i}x})\left(\sin^2(x)+3\right) & 0 \\ x & 1+\cos(x) \end{pmatrix}, \quad B^{(4)}(x) = \begin{pmatrix} 1-\mathrm{e}^{\mathbf{i}x} & 0 \\ 0 & 1+\cos(x) \end{pmatrix}$$

and define

$$f_4(x) = Q(x)A^{(4)}(x)Q(x)^T$$
$$g_4(x) = Q(x)B^{(4)}(x)Q(x)^T.$$

Figures 2.4 (a) and (b) show the singular values and the moduli of the eigenvalues of $T_{200}(f_4)$ and $T_{200}^{-1}(g_4)T_{200}(f_4)$, respectively. For a better resolution, in Figure 2.4 (b) some singular values of order about 10^4 have been cut. Figures 2.4 (c) and (d) refer to the eigenvalues in the complex plane of the same matrices. The reasoning regarding the behavior of the eigenvalues applies as in Case 3. Table 2.3 (b) shows the number of GMRES iterations required to reach the prescribed tolerance varying n. Again the number of iterations increases with n for $T_n(f_4)$, while in the preconditioned case the related iteration count remains constant. For the same reason of the previous example, even in this case Theorem 13 does not apply, but again the numerical results give us hope for improving our tools.

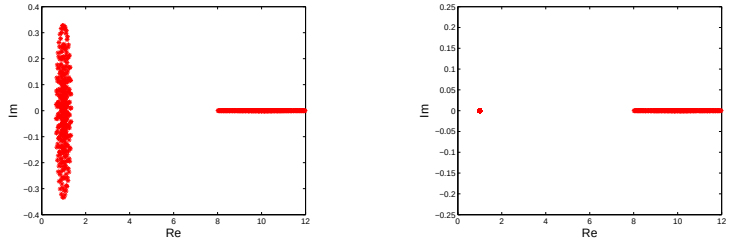


(a) (b) (c) (d)

Figure 2.5: (a) Eigenvalues in the complex plane of $T_n(A^{(5)})$ with $n = (20, 20)$; (b) Eigenvalues in the complex plane of $T_n(f_5)$ with $n = (20, 20)$; (c) Eigenvalues in the complex plane of $T_n^{-1}(B^{(5)})T_n(A^{(5)})$ with $n = (20, 20)$; (d) Eigenvalues in the complex plane of $T_n^{-1}(g_5)T_n(f_5)$ with $n = (20, 20)$

	Counting the outliers	
$n = (n_1, n_2)$	Out.	Out./$\sqrt{n}$
(5,5)	32	6.40
(10,10)	72	7.20
(15,15)	112	7.47
(20,20)	152	7.60
(25,25)	192	7.68
(30,30)	232	7.73

Table 2.4: Number of outliers for both $T_n(f_5)$ and $T_n^{-1}(g_5)T_n(f_5)$ varying n_1 and n_2



(a) (b)

Figure 2.6: (a) Eigenvalues in the complex plane of $T_n(A^{(6)})$ with $n = (20, 20)$; (b) Eigenvalues in the complex plane of $T_n^{-1}(B^{(6)})T_n(A^{(6)})$ with $n = (20, 20)$

Bivariate examples

Here we fix $s = 2$ and $k = 2$, that is we consider $\mathcal{M}_2$-valued symbols of 2 variables. In particular, we extend the definitions of f and g given in (2.8) taking $x = (x_1, x_2)$ and

$$Q(x) = \begin{pmatrix} \cos(x_1 + x_2) & \sin(x_1 + x_2) \\ -\sin(x_1 + x_2) & \cos(x_1 + x_2) \end{pmatrix}.$$

From here onwards, n is a 2-index, that is of type $n = (n_1, n_2)$.

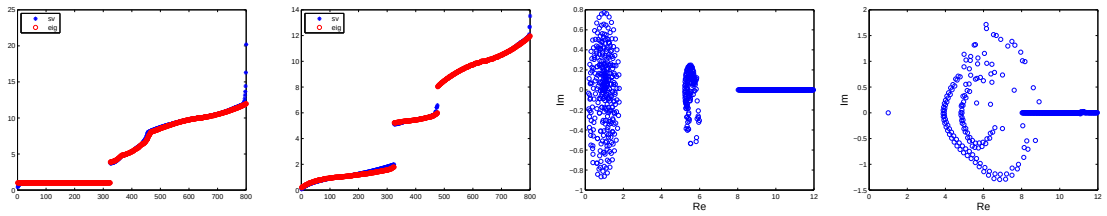
Case 5. This case can be seen as a 2-level extension of Case 1 obtained by choosing

$$A^{(5)}(x) = \begin{pmatrix} 3\mathbf{i} + \cos(x_1) + \cos(x_2) & 0 \\ 0 & 10 + 2(\mathrm{e}^{\mathbf{i}x_1} + \mathrm{e}^{\mathbf{i}x_2}) \end{pmatrix}, \quad B^{(5)}(x) = \begin{pmatrix} 1 & 0 \\ 0 & 10 + 2(\mathrm{e}^{\mathbf{i}x_1} + \mathrm{e}^{\mathbf{i}x_2}) \end{pmatrix}$$

and defining

$$f_5(x) = Q(x)A^{(5)}(x)Q(x)^T$$
$$g_5(x) = Q(x)B^{(5)}(x)Q(x)^T.$$

Let us observe that f_5 and $A^{(5)}$ are similar via the unitary transformation $Q(x)$, then, according to the theory, the associated 2-level block Toeplitz matrices are distributed in the sense of the eigenvalues in the same way. As shown in Figures 2.5 (a) and (b), in which $n = (n_1, n_2) = (20, 20)$, both the eigenvalues of $T_n(f_5)$ and $T_n(A^{(5)})$ are divided in two sub-clusters, one at the range of $A^{(5)}_{1,1}(x) = \lambda_1(f_5(x))$, the other at the range of $A^{(5)}_{2,2}(x) = \lambda_2(f_5(x))$. Interestingly enough, for $T_n(A^{(5)})$ the clusters are of strong type, while in the case of $T_n(f_5)$ the spectrum presents outliers with real part in $(3, 7)$ and imaginary part equal to 1.5. Table 2.4 shows that the number of outliers seems to behave as $o(\widehat{n})$ or, more specifically, as $O(\sqrt{\widehat{n}})$ (notice that this estimate is in line with the analysis in [136]). Analogous results are obtained in the comparison between $T_n^{-1}(g_5)T_n(f_5)$ and $T_n^{-1}(B^{(5)})T_n(A^{(5)})$, as shown in Figures 2.5 (c) and (d). Refer again to Table 2.4 for the number of outliers of $T_n^{-1}(g_5)T_n(f_5)$ varying n_1 and n_2 (this number is exactly the same as the number of outliers of $T_n(f_5)$). The eigenvalues of the symbol f_5 have no zeros, so the number of GMRES iterations required to reach tolerance 10^{-6} in solving the system associated to $T_n(f_5)$ is optimal, that is it does not depend on n. However, as shown in Table 2.5 (a), when preconditioning with $T_n(g_5)$, we preserve the optimality with a smaller number of iterations.



(a) (b) (c) (d)

Figure 2.7: (a) Singular values and moduli of the eigenvalues of $T_n(f_6)$ with $n = (20, 20)$; (b) Singular values and moduli of the eigenvalues of $T_n^{-1}(g_6)T_n(f_6)$ with $n = (20, 20)$; (c) Eigenvalues in the complex plane of $T_n(f_6)$ with $n = (20, 20)$; (d) Eigenvalues in the complex plane of $T_n^{-1}(g_6)T_n(f_6)$ with $n = (20, 20)$

	Iterations			
$n = (n_1, n_2)$	No Prec.	Prec.	No Prec.	Prec.
(5,5)	21	15	21	14
(10,10)	33	24	50	16
(15,15)	39	24	69	16
(20,20)	42	25	103	16
	(a)		(b)	

Table 2.5: (a) Number of GMRES iterations for $T_n(f_5)$ and $T_n^{-1}(g_5)T_n(f_5)$ varying n_1 and n_2; (b) Number of GMRES iterations for $T_n(f_6)$ and $T_n^{-1}(g_6)T_n(f_6)$ varying n_1 and n_2

Case 6. Let us choose $A^{(6)}$ and $B^{(6)}$ as

$$A^{(6)}(x) = \begin{pmatrix} 1 - (\mathrm{e}^{\mathbf{i}x_1} + \mathrm{e}^{\mathbf{i}x_2})/2 & 0 \\ 1/(x_1^2 + x_2^2 - 2) & 10 + \cos(x_1) + \cos(x_2) \end{pmatrix}, \quad B^{(6)}(x) = \begin{pmatrix} 1 - (\mathrm{e}^{\mathbf{i}x_1} + \mathrm{e}^{\mathbf{i}x_2})/2 & 0 \\ 0 & 1 \end{pmatrix}$$

and define

$$f_6(x) = Q(x)A^{(6)}(x)Q(x)^T$$
$$g_6(x) = Q(x)B^{(6)}(x)Q(x)^T.$$

The remark pointed out in the previous example about the outliers applies also in this case (compare Figures 2.6 (a) and (b) with Figures 2.7 (c) and (d)). In particular, we have found that the outliers behave again like $O(\sqrt{\widehat{n}})$. The singular values and the moduli of the eigenvalues of $T_n(f_6)$ are bounded, as shown in Figure 2.7 (a). More precisely, a half of the eigenvalues of $T_n(f_6)$ is clustered at $[8, 12]$, that is in the range of $A^{(6)}_{2,2}(x) = \lambda_2(f_6(x))$, the remaining part behaves as $A^{(6)}_{1,1}(x) = \lambda_1(f_6(x))$ (see Figure 2.7 (c)). Figures 2.7 (b) and (d) refer to the singular values and the eigenvalues of $T_n^{-1}(g_6)T_n(f_6)$. Let us observe that, although $0 \in \mathrm{Coh}[\mathcal{ENR}(g_6)]$, the spectrum of $T_n^{-1}(g_6)T_n(f_6)$ is essentially determined by the spectrum of the function $g_6^{-1}f_6$. Table 2.5 (b) highlights once again that the GMRES with preconditioner $T_n(g_6)$ converges faster than its non-preconditioned version.

From previous examples, we conclude that the proposed preconditioning approaches for non-Hermitian problems are numerically effective and confirm the theoretical findings.

2.2 Preconditioned HSS for multilevel block Toeplitz matrices

In this section, we perform a spectral analysis of the PHSS scheme (Section 1.7) applied to linear systems of the form $A_n x = b$, in the case where the coefficient matrix $A_n = T_n(f)$ and the preconditioner $P_n = T_n(g)$ are multilevel block Toeplitz matrices. Our aim is to compute the asymptotic eigenvalue distribution as $n \to \infty$ of the PHSS iteration matrix (1.28), which, in this framework, takes the form

$$\begin{aligned} M_n(\alpha) :=& \left(\alpha I + \mathbf{i}\, T_n^{-1}(g)\mathrm{Im}(T_n(f))\right)^{-1} \left(\alpha I - T_n^{-1}(g)\mathrm{Re}(T_n(f))\right) \\ & \left(\alpha I + T_n^{-1}(g)\mathrm{Re}(T_n(f))\right)^{-1} \left(\alpha I - \mathbf{i}\, T_n^{-1}(g)\mathrm{Im}(T_n(f))\right) \\ =& \left(\alpha T_n(g) + \mathbf{i}\,\mathrm{Im}(T_n(f))\right)^{-1} \left(\alpha T_n(g) - \mathrm{Re}(T_n(f))\right) \\ & \left(\alpha T_n(g) + \mathrm{Re}(T_n(f))\right)^{-1} \left(\alpha T_n(g) - \mathbf{i}\,\mathrm{Im}(T_n(f))\right). \end{aligned} \tag{2.9}$$

To reach this goal, we first prove Theorems 15,17, concerning the spectral distribution of matrix-sequences obtained from a combination of some algebraic operations on multilevel block Toeplitz matrices. After that, we will be able to compute the eigenvalue distribution of the algebraic combination of Toeplitz matrices which show up in the PHSS method. Finally, we will show through numerical experiments how the knowledge of the eigenvalue distribution of $M_n(\alpha)$ can be used in practice either to provide a guess for the asymptotic spectral radius of $M_n(\alpha)$ or to find efficient PHSS preconditioners $T_n(g)$ for linear systems with coefficient matrix $T_n(f)$.

2.2.1 Spectral distribution results for some algebraic combinations

When the involved operations on Toeplitz matrices are not just the simple inversion and multiplication considered in Theorem 13 we need to extend our tools: this is what we are going to do in this subsection. Before focusing on the preliminary steps needed for proving Theorem 15, we point out that, thanks to the notion of GLT sequences (see Section 1.5), an even more powerful version of Theorem 15, reported in Theorem 14, is already available for computing the singular value distribution of algebraic combinations of multilevel Toeplitz matrices.

Theorem 14. [133] *Let $f_1, \ldots, f_m \in L^1(k, s)$ and let $\phi(f_1, \ldots, f_m) : I_k \to \mathcal{M}_s$,*

$$\phi(f_1, \ldots, f_m) := \sum_{i=1}^{r} \prod_{j=1}^{q_i} g_{ij}^{\nu_{ij}},$$

where $r, q_1, \ldots, q_r$ are positive integers and, for all i and j, $\nu_{ij} \in \{-1, +1\}$, $g_{ij} \in \{f_1, \ldots, f_m\}$, and g_{ij} is invertible a.e. whenever $\nu_{ij} = -1$. Then, setting

$$A_n := \phi(T_n(f_1), \ldots, T_n(f_m)) := \sum_{i=1}^{r} \prod_{j=1}^{q_i} T_n^{\nu_{ij}}(g_{ij}),$$

we have $\{A_n\}_{n\in\mathbb{N}^k} \sim_\sigma (\phi(f_1, \ldots, f_m), I_k)$.

Roughly speaking, Theorem 14 says that, when A_n is the matrix obtained by applying some algebraic operations to certain Toeplitz matrices $T_n(f_1), \ldots, T_n(f_m)$, the function $\phi(f_1, \ldots, f_m)$ obtained from the same operations applied to the generating functions $f_1, \ldots, f_m$ describes the singular value distribution

of the matrix-family $\{A_n\}_{n\in\mathbb{N}^k}$, in the sense of Definition 3. The proof of Theorem 14 easily follows from properties **[GLT1-3]**. Concerning the eigenvalue distribution of $\{A_n\}_{n\in\mathbb{N}^k}$, we are going to prove in Theorem 15 a result analogous to the one in Theorem 14. For the proof of Theorem 15, we need to strengthen a little bit the hypotheses of Theorem 14. In particular, we will assume that $\phi(f_1,\dots,f_m)$ belongs to the Tilli class. We will also need the following proposition.

Proposition 9. *Let $f,g\in L^\infty(k,s)$ with $0\notin \mathrm{Coh}[\mathcal{ENR}(g)]$, then $\|T_n(f)T_n^{-1}(g)-T_n(fg^{-1})\|_1=o(\widehat{n})$ as $n\to\infty$.*

Proof. By Lemma 3, our assumptions imply that $d:=d(\mathrm{Coh}[\mathcal{ENR}(g)],0)$ is positive, every $T_n(g)$ is invertible with inverse matrix $T_n^{-1}(g)$ satisfying the inequality $\|T_n^{-1}(g)\|\le\frac{1}{d}$, g is invertible a.e., and $g^{-1}\in L^\infty(k,s)$. In particular, $fg^{-1}\in L^\infty(k,s)$. Using Proposition 8 and the Hölder inequality (1.2) for the Schatten 1-norm, we obtain

$$\begin{aligned}\|T_n(f)T_n^{-1}(g)-T_n(fg^{-1})\|_1&=\|(T_n(f)-T_n(fg^{-1})T_n(g))T_n^{-1}(g)\|_1\\&\le\|T_n(f)-T_n(fg^{-1})T_n(g)\|_1\|T_n^{-1}(g)\|\\&\le\|T_n(f)-T_n(fg^{-1})T_n(g)\|_1\frac{1}{d}=o(\widehat{n}),\end{aligned}$$

which proves the thesis. □

We are now ready to prove Theorem 15.

Theorem 15. *Let $f_1,\dots,f_m\in L^\infty(k,s)$ and let $\phi(f_1,\dots,f_m):I_k\to\mathcal{M}_s$,*

$$\phi(f_1,\dots,f_m):=\sum_{i=1}^{r}\prod_{j=1}^{q_i}g_{ij}^{\nu_{ij}},\tag{2.10}$$

where $r,q_1,\dots,q_r$ are positive integers and, for all i and j, $\nu_{ij}\in\{-1,+1\}$, $g_{ij}\in\{f_1,\dots,f_m\}$, and $0\notin\mathrm{Coh}[\mathcal{ENR}(g_{ij})]$ whenever $\nu_{ij}=-1$. Assume that $\phi(f_1,\dots,f_m)$ belongs to the Tilli class. Then, setting

$$A_n:=\phi(T_n(f_1),\dots,T_n(f_m)):=\sum_{i=1}^{r}\prod_{j=1}^{q_i}T_n^{\nu_{ij}}(g_{ij}),\tag{2.11}$$

we have $\{A_n\}_{n\in\mathbb{N}^k}\sim_\lambda(\phi(f_1,\dots,f_m),I_k)$.

Proof. Let $\phi(x):=\phi(f_1(x),\dots,f_m(x))$, $x\in I_k$, and let $\{A_{n(m)}\}_m$ be any matrix-sequence extracted from the family $\{A_n\}_{n\in\mathbb{N}^k}$. In the following, the index m remains hidden and it is understood that $n\to\infty$ when $m\to\infty$. We prove that $\{A_n\}_n\sim_\lambda(\phi,I_k)$ by showing that the conditions **(c_1)**–**(c_4)** in Theorem 1 are met with $f=\phi$. Consider the following sets:

$$\begin{aligned}M&:=\{g_{ij}:\ i=1,\dots,r,\ j=1,\dots,q_i\},\\E&:=\{g_{ij}:\ i=1,\dots,r,\ j=1,\dots,q_i,\ \nu_{ij}=-1\}.\end{aligned}$$

By hypothesis and by Lemma 3, for every $g_{ij}\in E$ the distance $d_{ij}:=d(\mathrm{Coh}[\mathcal{ENR}(g_{ij})],0)$ is positive and

$$\|T_n^{-1}(g_{ij})\|\le\frac{1}{d_{ij}}.\tag{2.12}$$

Moreover, for every $g_{ij}\in M\backslash E$ we have (see (1.18))

$$\|T_n(g_{ij})\|\le\|g_{ij}\|_{L^\infty}.\tag{2.13}$$

By the sub-multiplicative property of the spectral norm, we obtain

$$\|A_n\|=\left\|\sum_{i=1}^{r}\prod_{j=1}^{q_i}T_n^{\nu_{ij}}(g_{ij})\right\|\le\sum_{i=1}^{r}\prod_{j=1}^{q_i}\|T_n^{\nu_{ij}}(g_{ij})\|.$$

Combining the previous inequality with (2.12) when $g_{ij}\in E$ and (2.13) when $g_{ij}\in M\backslash E$, we conclude that condition **(c_1)** in Theorem 1 is met.

Now we observe that $g_{ij}^{-1} \in L^\infty(k,s)$ for every $g_{ij} \in E$ (see Lemma 3) and so $\phi \in L^\infty(k,s)$. Recalling that ϕ is in the Tilli class by hypothesis, the application of Theorem 7 shows that $\{T_n(\phi)\}_n \sim_\lambda (\phi, I_k)$. The latter implies that, for all integers $N \geq 0$,

$$\lim_{n\to\infty} \frac{\mathrm{tr}(T_n^N(\phi))}{s\widehat{n}} = \frac{1}{(2\pi)^k} \int_{I_k} \frac{\mathrm{tr}(\phi^N(x))}{s}\, dx. \tag{2.14}$$

Indeed, although the function $\tilde{F}(z) = z^N$ does not belong to $\mathcal{C}_0(\mathbb{C})$, we know that $\phi \in L^\infty(k,s)$. This has two consequences: first, using $\|T_n(\phi)\| \leq \|\phi\|_{L^\infty}$, all the eigenvalues of $T_n(\phi)$ lie in the compact disk $\overline{D(0, \|\phi\|_{L^\infty})}$; second, by definition of $\|\phi\|_{L^\infty}$ (see (1.7)), also the eigenvalues of ϕ belong to $\overline{D(0, \|\phi\|_{L^\infty})}$ a.e. Therefore, by choosing any function $F \in \mathcal{C}_0(\mathbb{C})$ such that $F(z) = z^N$ for all $z \in \overline{D(0, \|\phi\|_{L^\infty})}$, from $\{T_n(\phi)\}_n \sim_\lambda (\phi, I_k)$ we get (2.14). By (2.14) and by using the inequality (1.3), condition $(\boldsymbol{c_2})$ in Theorem 1 is proved if we show that

$$\|A_n^N - T_n^N(\phi)\|_1 = \left\| \left(\sum_{i=1}^r \prod_{j=1}^{q_i} T_n^{\nu_{ij}}(g_{ij}) \right)^N - T_n^N \left(\sum_{i=1}^r \prod_{j=1}^{q_i} g_{ij}^{\nu_{ij}} \right) \right\|_1 = o(\widehat{n}), \tag{2.15}$$

for every positive integer N. We first prove (2.15) for $N = 1$. Using the linearity of the operator $T_n(\cdot)$, we can bound the quantity in (2.15) as follows:

$$\left\| \sum_{i=1}^r \prod_{j=1}^{q_i} T_n^{\nu_{ij}}(g_{ij}) - T_n \left(\sum_{i=1}^r \prod_{j=1}^{q_i} g_{ij}^{\nu_{ij}} \right) \right\|_1 \leq \sum_{i=1}^r \left\| \prod_{j=1}^{q_i} T_n^{\nu_{ij}}(g_{ij}) - T_n \left(\prod_{j=1}^{q_i} g_{ij}^{\nu_{ij}} \right) \right\|_1 .$$

Therefore, it is sufficient to prove that, for all $i = 1, \ldots, r$,

$$\left\| \prod_{j=1}^{q_i} T_n^{\nu_{ij}}(g_{ij}) - T_n \left(\prod_{j=1}^{q_i} g_{ij}^{\nu_{ij}} \right) \right\|_1 = o(\widehat{n}). \tag{2.16}$$

For each i, we prove (2.16) by induction on q_i. If $q_i = 1$, (2.16) holds (obviously) if $\nu_{i1} = 1$, while in the case $\nu_{i1} = -1$ it holds by Proposition 9 applied with $f = 1$ and $g = g_{i1}$. For $q_i \geq 2$, by the inductive step we have $\| \prod_{j=1}^{q_i-1} T_n^{\nu_{ij}}(g_{ij}) - T_n(\prod_{j=1}^{q_i-1} g_{ij}^{\nu_{ij}})\|_1 = o(\widehat{n})$, i.e., $\prod_{j=1}^{q_i-1} T_n^{\nu_{ij}}(g_{ij}) = T_n(\prod_{j=1}^{q_i-1} g_{ij}^{\nu_{ij}}) + R_n$ with $\|R_n\|_1 = o(\widehat{n})$. As a consequence,

$$\begin{aligned} \left\| \prod_{j=1}^{q_i} T_n^{\nu_{ij}}(g_{ij}) - T_n \left(\prod_{j=1}^{q_i} g_{ij}^{\nu_{ij}} \right) \right\|_1 &= \left\| \left(\prod_{j=1}^{q_i-1} T_n^{\nu_{ij}}(g_{ij}) \right) T_n^{\nu_{iq_i}}(g_{iq_i}) - T_n \left(\prod_{j=1}^{q_i} g_{ij}^{\nu_{ij}} \right) \right\|_1 \\ &= \left\| T_n \left(\prod_{j=1}^{q_i-1} g_{ij}^{\nu_{ij}} \right) T_n^{\nu_{iq_i}}(g_{iq_i}) + R_n T_n^{\nu_{iq_i}}(g_{iq_i}) - T_n \left(\prod_{j=1}^{q_i} g_{ij}^{\nu_{ij}} \right) \right\|_1 . \end{aligned} \tag{2.17}$$

Using Proposition 8 or Proposition 9 (depending on the value of ν_{iq_i}), we obtain

$$T_n \left(\prod_{j=1}^{q_i-1} g_{ij}^{\nu_{ij}} \right) T_n^{\nu_{iq_i}}(g_{iq_i}) = T_n \left(\prod_{j=1}^{q_i} g_{ij}^{\nu_{ij}} \right) + R_n' \tag{2.18}$$

with $\|R_n'\|_1 = o(\widehat{n})$. Substituting (2.18) in (2.17) and using the Hölder inequality for the Schatten 1-norm as well as (2.12) or (2.13) (again depending on the value of ν_{iq_i}), we obtain

$$\begin{aligned} \left\| \prod_{j=1}^{q_i} T_n^{\nu_{ij}}(g_{ij}) - T_n \left(\prod_{j=1}^{q_i} g_{ij}^{\nu_{ij}} \right) \right\|_1 &= \|R_n T_n^{\nu_{iq_i}}(g_{iq_i}) + R_n'\|_1 \leq \|R_n\|_1 \|T_n^{\nu_{iq_i}}(g_{iq_i})\| + \|R_n'\|_1 \\ &= o(\widehat{n}). \end{aligned}$$

This concludes the proof of (2.16) and shows that (2.15) holds for $N = 1$. Hence, we can write

$$A_n = T_n(\phi) + R_n'' \tag{2.19}$$

with $\|R_n''\|_1 = o(\widehat{n})$ and $\|R_n''\|$ uniformly bounded with respect to n (the latter is true because $R_n'' = A_n - T_n(\phi)$ and we have seen that $\|A_n\|$, $\|T_n(\phi)\|$ are uniformly bounded). Using (2.19), when $N \geq 2$ we have

$$A_n^N = T_n^N(\phi) + S_n,$$

where S_n is the sum of all possible (different) combinations of products of t matrices $T_n(\phi)$ and ℓ matrices R_n'', with $t + \ell = N$, $t \neq N$. For every summand S of S_n we have

$$\|S\|_1 \leq \|T_n(\phi)\|^t \|R_n''\|^{\ell-1} \|R_n''\|_1 = o(\widehat{n}).$$

Since the number of summands in S_n is finite, (2.15) holds for every positive integer N, and condition $(\boldsymbol{c_2})$ in Theorem 1 is then satisfied.

Condition $(\boldsymbol{c_3})$ is also satisfied, because the sequences of multilevel block Toeplitz matrices generated by a $L^1(k,s)$ function, as well as their corresponding algebra, form a subset of the GLT class ([**GLT1-3**]). Finally, by taking into account that $\mathcal{ER}(\phi)$ has empty interior and does not disconnect the complex plane, the last condition $(\boldsymbol{c_4})$ in Theorem 1 is satisfied, and the application of Theorem 1 shows that $\{A_n\}_n \sim_\lambda (\phi, I_k)$. Since this is true for any matrix-sequence $\{A_n\}_n$ extracted from the family $\{A_n\}_{n\in\mathbb{N}^k}$, we conclude that $\{A_n\}_{n\in\mathbb{N}^k} \sim_\lambda (\phi, I_k)$. □

Remark 10. It is important to emphasize that, since the matrix product is not commutative, in the case $s > 1$ the order in which the matrix-valued functions $g_{ij}^{\nu_{ij}}$ appear in the product in (2.10) must be the same as the order in which the Toeplitz matrices $T_n^{\nu_{ij}}(g_{ij})$ appear in the product in (2.11). Nevertheless, in the case $s = 1$, the function $\phi(f_1, \ldots, f_m)$ is not affected by the order of the (scalar-valued) functions $g_{ij}^{\nu_{ij}}$. In this case, different orderings of the Toeplitz matrices $T_n^{\nu_{ij}}(g_{ij})$ in (2.11) give rise to matrix-families $\{A_n\}_{n\in\mathbb{N}^k}$ with the same symbol $\phi(f_1, \ldots, f_m)$. Of course, similar considerations also apply to Theorem 14.

We end this subsection by providing a simple (but useful) extension of Theorem 15. For the proof, we will make use of Theorem 16 and of Lemma 4 below. We recall that the direct sum of p matrices $X_1, \ldots, X_p$ is defined as the block-diagonal matrix

$$X_1 \oplus \cdots \oplus X_p := \operatorname*{diag}_{j=1,\ldots,p} X_j = \begin{bmatrix} X_1 & & \\ & \ddots & \\ & & X_p \end{bmatrix}.$$

Note that, if X_i is of size $m_1^{(i)} \times m_2^{(i)}$, then $X_1 \oplus \cdots \oplus X_p$ is of size $(\sum_{i=1}^p m_1^{(i)}) \times (\sum_{i=1}^p m_2^{(i)})$.

In Theorem 16, we prove the linearity of the Toeplitz operator $T_n(\cdot)$ with respect to the direct sum, modulo permutation transformations which depend only on the dimensions of the involved matrices. The proof of Theorem 16 is based on a distributive property of the tensor product with respect to the direct sum; see [75, Lemma 4].

Theorem 16. *Let $t := (t_1, \ldots, t_p) \in \mathbb{N}^p$, let $n \in \mathbb{N}^k$ and let $f_v \in L^1(k, t_v)$, $v = 1, \ldots, p$. Then, there exists a permutation matrix $Q_{\widehat{n},t}$, depending* only *on $\widehat{n}$ and t, such that*

$$T_n(f_1 \oplus \cdots \oplus f_p) = Q_{\widehat{n},t}[T_n(f_1) \oplus \cdots \oplus T_n(f_p)]Q_{\widehat{n},t}^T.$$

Proof. Let $s := \sum_{v=1}^p t_v$. We first notice that the direct-sum function $f_1 \oplus \cdots \oplus f_p : I_k \to \mathcal{M}_s$,

$$(f_1 \oplus \cdots \oplus f_p)(x) = f_1(x) \oplus \cdots \oplus f_p(x) = \begin{bmatrix} f_1(x) & & \\ & \ddots & \\ & & f_p(x) \end{bmatrix},$$

belongs to $L^1(k,s)$, and its Fourier coefficients (see (1.14)) are given by

$$(\widehat{\oplus_{v=1}^p f_v})_j := \frac{1}{(2\pi)^k} \int_{I_k} (f_1 \oplus \cdots \oplus f_p)(x) \mathrm{e}^{-\mathbf{i}\langle j,x\rangle}\, dx = (\widehat{f_1})_j \oplus \cdots \oplus (\widehat{f_p})_j \in \mathcal{M}_s, \qquad j \in \mathbb{Z}^k$$

(we recall that the integrals are computed componentwise). Therefore, using the definition (1.15), we get

$$\begin{aligned} T_n(f_1 \oplus \cdots \oplus f_p) &= \sum_{|j_1|<n_1} \cdots \sum_{|j_k|<n_k} \left[J_{n_1}^{(j_1)} \otimes \cdots \otimes J_{n_k}^{(j_k)} \right] \otimes (\widehat{\oplus_{v=1}^{p} f_v})_j \\ &= \sum_{|j_1|<n_1} \cdots \sum_{|j_k|<n_k} \left[J_{n_1}^{(j_1)} \otimes \cdots \otimes J_{n_k}^{(j_k)} \right] \otimes ((\widehat{f}_1)_j \oplus \cdots \oplus (\widehat{f}_p)_j). \end{aligned}$$

Due to Lemma 4 in [75], there exists a permutation matrix $Q_{\widehat{n},t}$, depending only on $\widehat{n}$ and t, such that, for all $j \in \mathbb{Z}^k$,

$$\begin{aligned} &\left[J_{n_1}^{(j_1)} \otimes \cdots \otimes J_{n_k}^{(j_k)} \right] \otimes ((\widehat{f}_1)_j \oplus \cdots \oplus (\widehat{f}_p)_j) \\ &= Q_{\widehat{n},t} \left[\left(J_{n_1}^{(j_1)} \otimes \cdots \otimes J_{n_k}^{(j_k)} \otimes (\widehat{f}_1)_j \right) \oplus \cdots \oplus \left(J_{n_1}^{(j_1)} \otimes \cdots \otimes J_{n_k}^{(j_k)} \otimes (\widehat{f}_p)_j \right) \right] Q_{\widehat{n},t}^T. \end{aligned}$$

Hence

$$\begin{aligned} T_n(f_1 \oplus \cdots \oplus f_p) &= \sum_{|j_1|<n_1} \cdots \sum_{|j_k|<n_k} \left[J_{n_1}^{(j_1)} \otimes \cdots \otimes J_{n_k}^{(j_k)} \right] \otimes ((\widehat{f}_1)_j \oplus \cdots \oplus (\widehat{f}_p)_j) \\ &= \sum_{|j_1|<n_1} \cdots \sum_{|j_k|<n_k} Q_{\widehat{n},t} \left[\left(J_{n_1}^{(j_1)} \otimes \cdots \otimes J_{n_k}^{(j_k)} \otimes (\widehat{f}_1)_j \right) \oplus \cdots \right. \\ &\qquad \left. \cdots \oplus \left(J_{n_1}^{(j_1)} \otimes \cdots \otimes J_{n_k}^{(j_k)} \otimes (\widehat{f}_p)_j \right) \right] Q_{\widehat{n},t}^T \\ &= Q_{\widehat{n},t} \left\{ \sum_{|j_1|<n_1} \cdots \sum_{|j_k|<n_k} \left[\left(J_{n_1}^{(j_1)} \otimes \cdots \otimes J_{n_k}^{(j_k)} \otimes (\widehat{f}_1)_j \right) \oplus \cdots \right. \right. \\ &\qquad \left. \left. \cdots \oplus \left(J_{n_1}^{(j_1)} \otimes \cdots \otimes J_{n_k}^{(j_k)} \otimes (\widehat{f}_p)_j \right) \right] \right\} Q_{\widehat{n},t}^T \\ &= Q_{\widehat{n},t} \left\{ \left[\sum_{|j_1|<n_1} \cdots \sum_{|j_k|<n_k} \left(J_{n_1}^{(j_1)} \otimes \cdots \otimes J_{n_k}^{(j_k)} \otimes (\widehat{f}_1)_j \right) \right] \oplus \cdots \right. \\ &\qquad \left. \cdots \oplus \left[\sum_{|j_1|<n_1} \cdots \sum_{|j_k|<n_k} \left(J_{n_1}^{(j_1)} \otimes \cdots \otimes J_{n_k}^{(j_k)} \otimes (\widehat{f}_p)_j \right) \right] \right\} Q_{\widehat{n},t}^T \\ &= Q_{\widehat{n},t} [T_n(f_1) \oplus \cdots \oplus T_n(f_p)] Q_{\widehat{n},t}^T. \end{aligned}$$

□

Lemma 4 concerns the eigenvalue distribution of a sequence of direct sum of matrices and follows from Definition 3 and from the fact that, if $X_1, X_2, \ldots, X_p$ are square matrices of size $m_1, m_2, \ldots, m_p$, respectively, then the eigenvalues of $X_1 \oplus X_2 \oplus \cdots \oplus X_p$ are

$$\lambda_i(X_v), \quad i = 1, \ldots, m_v, \quad v = 1, \ldots, p.$$

Lemma 4. *Let $h_v : G \to \mathcal{M}_{t_v}$, $v = 1, \ldots, p$, be measurable functions, defined on a measurable set $G \subset \mathbb{R}^k$, with $0 < m_k(G) < \infty$. Let $\{A_{n,1}\}_n, \ldots, \{A_{n,p}\}_n$ be matrix-sequences, with $A_{n,v}$ of size $d_{n,v}$ tending to infinity as $n \to \infty$. Assume that, for all $v = 1, \ldots, p$, $\{A_{n,v}\}_n \sim_\lambda h_v$ and $d_{n,v}/d_n \to t_v/s$ as $n \to \infty$, where $d_n := \sum_{v=1}^p d_{n,v}$ and $s := \sum_{v=1}^p t_v$. Then*

$$\{A_{n,1} \oplus \cdots \oplus A_{n,p}\}_n \sim_\lambda h_1 \oplus \cdots \oplus h_p.$$

Theorem 17. *Let $f_1, \ldots, f_m \in L^\infty(k, s)$. Take an integer $p \geq 1$, a p-index $t := (t_1, \ldots, t_p)$ and a constant matrix C (independent of x) such that, for all $u = 1, \ldots, m$, we have*

$$f_u = C \left(\bigoplus_{v=1}^{p} f_{u,v} \right) C^{-1} \quad \text{a.e.},$$

where $f_{u,v} : I_k \to \mathcal{M}_{t_v},\ v = 1, \dots, p$, *are some functions. Let* $\phi(f_1, \dots, f_m) : I_k \to \mathcal{M}_s$,

$$\phi(f_1, \dots, f_m) := \sum_{i=1}^{r} \prod_{j=1}^{q_i} g_{ij}^{\nu_{ij}} = C \left[\bigoplus_{v=1}^{p} \left(\sum_{i=1}^{r} \prod_{j=1}^{q_i} g_{ij,v}^{\nu_{ij}} \right) \right] C^{-1} = C \left[\bigoplus_{v=1}^{p} \phi(f_{1,v}, \dots, f_{m,v}) \right] C^{-1}, \quad (2.20)$$

where $r, q_1, \dots, q_r$ *are positive integers and, for all* i *and* j, $\nu_{ij} \in \{-1, +1\}$, $g_{ij} \in \{f_1, \dots, f_m\}$, $g_{ij,v} \in \{f_{1,v}, \dots, f_{m,v}\}$, *and* $0 \notin \bigcup_{v=1}^{p} \mathrm{Coh}[\mathcal{ENR}(g_{ij,v})]$ *whenever* $\nu_{ij} = -1$. *Assume that each 'component'* $\phi(f_{1,v}, \dots, f_{m,v})$ *belongs to the Tilli class. Then, setting*

$$A_n := \phi(T_n(f_1), \dots, T_n(f_m)) := \sum_{i=1}^{r} \prod_{j=1}^{q_i} T_n^{\nu_{ij}}(g_{ij}), \quad (2.21)$$

we have $\{A_n\}_{n \in \mathbb{N}^k} \sim_\lambda (\phi(f_1, \dots, f_m), I_k)$.

Proof. Define $\tilde{f}_u := f_{u,1} \oplus \cdots \oplus f_{u,p}$, for $u = 1, \dots, m$. The hypotheses imply that $\sum_{v=1}^{p} t_v = s$ and $\tilde{f}_u \in L^\infty(k, s)$. Moreover, since $\tilde{f}_u = C^{-1} f_u C$ a.e., for all $u = 1, \dots, m$ and $n \in \mathbb{N}^k$ we have

$$T_n(\tilde{f}_u) = (I_{\widehat{n}} \otimes C)^{-1} T_n(f_u)(I_{\widehat{n}} \otimes C), \quad (2.22)$$

where $I_{\widehat{n}}$ is the identity matrix of order $\widehat{n}$. This follows from the definitions of $T_n(\tilde{f}_u)$, $T_n(f_u)$, see (1.15), the linearity of the integral (1.14) involved in the definition of the Fourier coefficients, and the properties of the tensor product of matrices. Now, by applying Theorem 16 we obtain

$$T_n(\tilde{f}_u) = Q_{\widehat{n},t}[T_n(f_{u,1}) \oplus \cdots \oplus T_n(f_{u,p})] Q_{\widehat{n},t}^T, \quad (2.23)$$

where $Q_{\widehat{n},t}$ is a permutation matrix depending only on $\widehat{n}$ and $t = (t_1, \dots, t_p)$. Putting together (2.22) and (2.23) we get

$$T_n(f_u) = (I_{\widehat{n}} \otimes C) Q_{\widehat{n},t}[T_n(f_{u,1}) \oplus \cdots \oplus T_n(f_{u,p})] Q_{\widehat{n},t}^T (I_{\widehat{n}} \otimes C)^{-1},$$

which holds for all $u = 1, \dots, m$ and $n \in \mathbb{N}^k$. Therefore, looking at (2.21), for all i, j we have

$$T_n^{\nu_{ij}}(g_{ij}) = (I_{\widehat{n}} \otimes C) Q_{\widehat{n},t}[T_n^{\nu_{ij}}(g_{ij,1}) \oplus \cdots \oplus T_n^{\nu_{ij}}(g_{ij,p})] Q_{\widehat{n},t}^T (I_{\widehat{n}} \otimes C)^{-1}$$

and so

$$A_n = \phi(T_n(f_1), \dots, T_n(f_m)) = (I_{\widehat{n}} \otimes C) Q_{\widehat{n},t} \left[\bigoplus_{v=1}^{p} \phi(T_n(f_{1,v}), \dots, T_n(f_{m,v})) \right] Q_{\widehat{n},t}^T (I_{\widehat{n}} \otimes C)^{-1}, \quad (2.24)$$

where (cf. (2.20))

$$\phi(T_n(f_{1,v}), \dots, T_n(f_{m,v})) := \sum_{i=1}^{r} \prod_{j=1}^{q_i} T_n^{\nu_{ij}}(g_{ij,v}), \qquad v = 1, \dots, p.$$

Since $0 \notin \bigcup_{v=1}^{p} \mathrm{Coh}[\mathcal{ENR}(g_{ij,v})]$ whenever $\nu_{ij} = -1$, by Theorem 15 we obtain

$$\{\phi(T_n(f_{1,v}), \dots, T_n(f_{m,v}))\}_{n \in \mathbb{N}^k} \sim_\lambda (\phi(f_{1,v}, \dots, f_{m,v}), I_k), \qquad v = 1, \dots, p.$$

The application of Lemma 4 concludes the proof. We just remark that, with the notation of Lemma 4, we have $A_{n,v} = \phi(T_n(f_{1,v}), \dots, T_n(f_{m,v}))$, $h_v = \phi(f_{1,v}, \dots, f_{m,v}) : I_k \to \mathcal{M}_{t_v}$, $d_{n,v} = \hat{n} t_v$ and $d_n = \widehat{n} s$. Also note that, by (2.20) and (2.24),

$$\phi(f_1, \dots, f_m) \sim \bigoplus_{v=1}^{p} \phi(f_{1,v}, \dots, f_{m,v}) \quad \text{a.e.,} \qquad A_n \sim \bigoplus_{v=1}^{p} \phi(T_n(f_{1,v}), \dots, T_n(f_{m,v})).$$

□

Theorem 17 shows that Theorem 15 continues to hold if the hypothesis '$\mathcal{ER}(\phi(f_1, \dots, f_m))$ has empty interior and does not disconnect the complex plane' (i.e., '$\phi(f_1, \dots, f_m)$ belongs to the Tilli class') is replaced by the weaker assumption that 'its individual 'components' $\mathcal{ER}(\phi(f_{1,v}, \dots, f_{m,v}))$, $v = 1, \dots, p$, have empty interior and do not disconnect the complex plane' (i.e., that 'every $\phi(f_{1,v}, \dots, f_{m,v})$ belongs to the Tilli class'). Note that Theorem 15 is obtained from Theorem 17 by choosing $t = s$ (a p-index with $p = 1$) and $C = I_s$ (the identity matrix of order s). This explicitly shows that Theorem 17 is an actual extension of Theorem 15.

Remark 11. Appropriate versions of Theorems 15,17 continue to hold even when, in the expression of $\phi(f_1, \dots, f_m)$, other operations are allowed such as unary multiplication for a complex scalar β and unary conjugate transposition. In fact, since $\beta T_n(f) = T_n(\beta f)$, $T_n^*(f) = T_n(f^*)$ and $T_n^{-*}(f) = T_n^{-1}(f^*)$, these operations are reduced to operations on the generating function.

2.2.2 PHSS applied to multilevel block Toeplitz matrices

Before starting, in Proposition 10 we prove that, whenever f is sectorial, the HSS and PHSS methods can be applied to a properly scaled version of the Toeplitz linear system $T_n(f)x = b$. This shows that the applicability of HSS and PHSS in the context of multilevel block Toeplitz matrices is actually maximal. Indeed, if f is not sectorial, then $T_n(f)$ could be either singular for some n or its condition number, as a function of n, may be erratic and with exponentially growing subsequences [23]. Consequently, the sectoriality requirement for f is mild or even minimal, if we want to ensure the invertibility of the coefficient matrix $T_n(f)$.

Proposition 10. *Let $f \in L^1(k,s)$ be sectorial. Then, there exists a complex number ω such that $\mathrm{Re}(\omega T_n(f))$ is positive definite for all $n \in \mathbb{N}^k$. In particular, the HSS and PHSS methods can be applied to the linear systems with coefficient matrix $\omega T_n(f) = T_n(\omega f)$.*

Proof. It is enough to combine Definition 2 and Proposition 1. Since f is sectorial, we can find a separating line z and a rotation number ω such that

$$\mathcal{ENR}(\omega f) = \omega \cdot \mathcal{ENR}(f) \subseteq \{w \in \mathbb{C} : \mathrm{Re}(w) \geq d(z,0)\}$$

and the eigenvalue of minimum modulus of $\mathrm{Re}(\omega f)$ is not a.e. equal to $d(z,0)$. Since the matrix $\mathrm{Re}(\omega f)$ is HPSD a.e. (see Remark 1), it follows that the minimal eigenvalue of $\mathrm{Re}(\omega f)$ is not a.e. equal to 0. Hence, by Proposition 1 and by the linearity of $T_n(\cdot)$,

$$T_n(\mathrm{Re}(\omega f)) = T_n\left(\frac{\omega f + (\omega f)^*}{2}\right) = \frac{T_n(\omega f) + T_n((\omega f)^*)}{2} = \frac{T_n(\omega f) + T_n^*(\omega f)}{2} = \mathrm{Re}(\omega T_n(f))$$

is HPD for all $n \in \mathbb{N}^k$. □

Now, let $f, g \in L^1(k,s)$, let $A_n := T_n(f)$, $P_n := T_n(g)$, and assume that $\mathrm{Re}(f)$ and g are HPSD a.e., with minimal eigenvalues $\lambda_{\min}(\mathrm{Re}(f))$ and $\lambda_{\min}(g)$ not a.e. equal to 0. In this way, Proposition 1, together with the equality $\mathrm{Re}(T_n(f)) = T_n(\mathrm{Re}(f))$, ensures that $\mathrm{Re}(A_n)$ and P_n are HPD for all $n \in \mathbb{N}^k$. Hence, we can apply to the linear system $A_n x = b$ the PHSS method with preconditioner P_n. The resulting PHSS iteration matrix is, cf. (2.9),

$$\begin{aligned} M_n(\alpha) := &\left(\alpha T_n(g) + \mathbf{i}\,\mathrm{Im}(T_n(f))\right)^{-1}\left(\alpha T_n(g) - \mathrm{Re}(T_n(f))\right) \\ &\left(\alpha T_n(g) + \mathrm{Re}(T_n(f))\right)^{-1}\left(\alpha T_n(g) - \mathbf{i}\,\mathrm{Im}(T_n(f))\right). \end{aligned} \tag{2.25}$$

It is now easy to see that $M_n(\alpha)$ can be written in the form (2.21). For example, for the first factor in (2.25) we have

$$\left(\alpha T_n(g) + \mathbf{i}\,\mathrm{Im}(T_n(f))\right)^{-1} = \left(\alpha T_n(g) + \mathbf{i}\,T_n(\mathrm{Im}(f))\right)^{-1} = T_n^{-1}(\alpha g + \mathbf{i}\,\mathrm{Im}(f)).$$

By performing similar calculations for the other factors, we obtain

$$\begin{aligned} M_n(\alpha) &= T_n^{-1}(\alpha g + \mathbf{i}\,\mathrm{Im}(f))T_n(\alpha g - \mathrm{Re}(f))T_n^{-1}(\alpha g + \mathrm{Re}(f))T_n(\alpha g - \mathbf{i}\,\mathrm{Im}(f)) \\ &= \phi(T_n(\alpha g + \mathbf{i}\,\mathrm{Im}(f)), T_n(\alpha g - \mathrm{Re}(f)), T_n(\alpha g + \mathrm{Re}(f)), T_n(\alpha g - \mathbf{i}\,\mathrm{Im}(f))), \end{aligned} \tag{2.26}$$

where $\phi(f_1, f_2, f_3, f_4) := f_1^{-1} f_2 f_3^{-1} f_4$. Therefore, the application of Theorems 15,17 yields the eigenvalue distribution and the symbol of the matrix-family $\{M_n(\alpha)\}_{n \in \mathbb{N}^k}$. This is reported in Theorem 18, together with other relevant properties of the PHSS in the Toeplitz setting. For the proof of one of these properties, we need the following lemma.

Lemma 5. *Let $0 < \lambda_1 \leq \ldots \leq \lambda_m$. Then, for all $\alpha > 0$,*

$$\max_{i=1,\ldots,m} \left|\frac{\alpha - \lambda_i}{\alpha + \lambda_i}\right| = \max\left(\left|\frac{\alpha - \lambda_1}{\alpha + \lambda_1}\right|, \left|\frac{\alpha - \lambda_m}{\alpha + \lambda_m}\right|\right)$$

and

$$\min_{\alpha > 0}\left[\max\left(\left|\frac{\alpha - \lambda_1}{\alpha + \lambda_1}\right|, \left|\frac{\alpha - \lambda_m}{\alpha + \lambda_m}\right|\right)\right] = \max\left(\left|\frac{\alpha - \lambda_1}{\alpha + \lambda_1}\right|, \left|\frac{\alpha - \lambda_m}{\alpha + \lambda_m}\right|\right)\Bigg|_{\alpha = \sqrt{\lambda_1 \lambda_m}} = \frac{\sqrt{\frac{\lambda_m}{\lambda_1}} - 1}{\sqrt{\frac{\lambda_m}{\lambda_1}} + 1}.$$

From now on, if $h : I_k \to \mathcal{M}_s$ is any measurable matrix-valued function whose eigenvalues are real a.e., we will denote by m_h and M_h the minimum and the maximum of $\mathcal{ER}(h)$. In other words,

$$m_h := \operatorname*{ess\,inf}_{x\in I_k} \lambda_{\min}(h(x)), \qquad M_h := \operatorname*{ess\,sup}_{x\in I_k} \lambda_{\max}(h(x)).$$

Theorem 18. *Let $f, g \in L^1(k,s)$, assume that $\mathrm{Re}(f)$, g are HPSD a.e., with minimal eigenvalues $\lambda_{\min}(\mathrm{Re}(f))$, $\lambda_{\min}(g)$ not a.e. equal to 0, and let $\alpha > 0$. Then, setting $A_n := T_n(f)$, $P_n := T_n(g)$, the following results hold.*

1. *$\mathrm{Re}(A_n)$, P_n are HPD for all $n \in \mathbb{N}^k$.*
2. *The PHSS applied to A_n with preconditioner P_n and parameter α has iteration matrix*

$$M_n(\alpha) = T_n^{-1}(\alpha g + \mathbf{i}\,\mathrm{Im}(f))T_n(\alpha g - \mathrm{Re}(f))T_n^{-1}(\alpha g + \mathrm{Re}(f))T_n(\alpha g - \mathbf{i}\,\mathrm{Im}(f)) \tag{2.27}$$

and

$$\rho(M_n(\alpha)) \le \sigma_n(\alpha) = \max_{\lambda\in\Lambda(P_n^{-1}\mathrm{Re}(A_n))} \left|\frac{\alpha-\lambda}{\alpha+\lambda}\right|. \tag{2.28}$$

3. *The eigenvalue distribution relation*

$$\{M_n(\alpha)\}_{n\in\mathbb{N}^k} \sim_\lambda (\phi_\alpha(f,g), I_k), \tag{2.29}$$

with

$$\phi_\alpha(f,g) := (\alpha g + \mathbf{i}\,\mathrm{Im}(f))^{-1}(\alpha g - \mathrm{Re}(f))(\alpha g + \mathrm{Re}(f))^{-1}(\alpha g - \mathbf{i}\,\mathrm{Im}(f)), \tag{2.30}$$

holds whenever $f, g \in L^\infty(k,s)$ and at least one between the following conditions is met:

(a) $\phi_\alpha(f,g)$ is in the Tilli class, $0 \notin \mathrm{Coh}[\mathcal{ENR}(\alpha g + \mathbf{i}\,\mathrm{Im}(f))] \bigcup \mathrm{Coh}[\mathcal{ENR}(\alpha g + \mathrm{Re}(f))]$;

(b) there is a p-index $t := (t_1, \dots, t_p)$ and a constant matrix C (independent of x) such that

$$f = C(f_1 \oplus \cdots \oplus f_p)C^{-1} \quad \text{a.e.,} \quad \text{with } f_v : I_k \to \mathcal{M}_{t_v} \text{ for } v = 1, \dots, p, \tag{2.31}$$

$$g = C(g_1 \oplus \cdots \oplus g_p)C^{-1} \quad \text{a.e.,} \quad \text{with } g_v : I_k \to \mathcal{M}_{t_v} \text{ for } v = 1, \dots, p, \tag{2.32}$$

every

$$\phi_\alpha(f_v, g_v) := (\alpha g_v + \mathbf{i}\,\mathrm{Im}(f_v))^{-1}(\alpha g_v - \mathrm{Re}(f_v))(\alpha g_v + \mathrm{Re}(f_v))^{-1}(\alpha g_v - \mathbf{i}\,\mathrm{Im}(f_v)),$$

$v = 1, \dots, p$, is in the Tilli class, and

$$0 \notin \left(\bigcup_{v=1}^p \mathrm{Coh}[\mathcal{ENR}(\alpha g_v + \mathbf{i}\,\mathrm{Im}(f_v))]\right) \bigcup \left(\bigcup_{v=1}^p \mathrm{Coh}[\mathcal{ENR}(\alpha g_v + \mathrm{Re}(f_v))]\right). \tag{2.33}$$

Note that (b) contains (a) as a particular case (take $p = 1$, $t = s$ and $C = I_s$ in (b) to obtain (a)). However, we stated these conditions separately, because (a) is simpler.

4. *Assume that g is HPD a.e. and let $h := g^{-1}\mathrm{Re}(f) : I_k \to \mathcal{M}_s$. Then,*

$$\lambda_{\min}(P_n^{-1}\mathrm{Re}(A_n)) \to m_h, \qquad \lambda_{\max}(P_n^{-1}\mathrm{Re}(A_n)) \to M_h. \tag{2.34}$$

As a consequence, for the values α_n^*, κ_n, $\sigma_n(\alpha_n^*)$ in (1.32)–(1.34) we have

$$\alpha_n^* \to \alpha^* := \sqrt{m_h M_h}, \qquad \kappa_n \to \kappa := \frac{M_h}{m_h}, \qquad \sigma_n(\alpha_n^*) \to \sigma^* := \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}, \tag{2.35}$$

provided that the definitions of α^*, κ, σ^* have meaning.[1] *In particular, recalling (2.28), we obtain*

$$\limsup_{n\to\infty} \rho(M_n(\alpha_n^*)) \le \lim_{n\to\infty} \sigma_n(\alpha_n^*) = \sigma^*. \tag{2.36}$$

Moreover,

$$\min_{\alpha>0} \|\rho(\phi_\alpha(f,g))\|_\infty \le \sigma^*, \tag{2.37}$$

where $\|\rho(\phi_\alpha(f,g))\|_\infty := \|\rho(\phi_\alpha(f,g))\|_{L^\infty(I_k)} = \operatorname{ess\,sup}_{x\in I_k} \rho(\phi_\alpha(f(x),g(x)))$ is the infinity norm of the spectral radius of the symbol $\phi_\alpha(f,g)$.

[1] While the limit relations (2.34) are always true, the relations (2.35) follow from (2.34), provided that no indeterminate form occurs. The only critical case is when $m_h = 0$ and $M_h = \infty$; in this case, the definitions of κ and σ^* have meaning ($\kappa = \infty$, $\sigma^* = 1$), but the definition of α^* has not (we encounter the indeterminate form $0 \cdot \infty$), and so we do not know if α_n^* converges to something or not. In all the other cases, when $m_h \neq 0$ or $M_h < \infty$, no indeterminate form occurs and the definitions of α^*, κ, σ^* have meaning.

Proof. Item 1 follows from Proposition 1. Item 2 follows from (2.26) and Theorem 10. The proof that (2.29) holds whenever $f, g \in L^\infty(k,s)$ and 3(a) is fulfilled follows from the expression of $M_n(\alpha)$ given in (2.27) and from Theorem 15 applied with

$$f_1 := \alpha g + \mathbf{i}\,\mathrm{Im}(f), \quad f_2 := \alpha g - \mathrm{Re}(f), \quad f_3 := \alpha g + \mathrm{Re}(f), \quad f_4 := \alpha g + \mathrm{Re}(f), \tag{2.38}$$

and

$$\phi(f_1, f_2, f_3, f_4) := f_1^{-1} f_2 f_3^{-1} f_4 = \phi_\alpha(f, g). \tag{2.39}$$

The proof that (2.29) holds whenever $f, g \in L^\infty(k,s)$ and 3(b) is fulfilled follows from the expression of $M_n(\alpha)$ given in (2.27) and from Theorem 17 applied with f_1, f_2, f_3, f_4, $\phi(f_1, f_2, f_3, f_4)$ as in (2.38)–(2.39). We just remark that, from (2.31)–(2.32), we get

$$\begin{aligned}
\alpha g \pm \mathbf{i}\,\mathrm{Im}(f) &= C\left[(\alpha g_1 \pm \mathbf{i}\,\mathrm{Im}(f_1)) \oplus \cdots \oplus (\alpha g_p \pm \mathbf{i}\,\mathrm{Im}(f_p))\right] C^{-1}, \\
\alpha g \pm \mathrm{Re}(f) &= C\left[(\alpha g_1 \pm \mathrm{Re}(f_1)) \oplus \cdots \oplus (\alpha g_p \pm \mathrm{Re}(f_p))\right] C^{-1}.
\end{aligned}$$

Now, assume that g is HPD a.e. and define $h := g^{-1}\mathrm{Re}(f)$ as in item 4. The limit relations (2.34) follow by combining the results of Theorem 9 and Proposition 2 (the proof is the same as in the scalar case so we do not report the details). It only remains to prove the inequality (2.37). The steps to prove (2.37) are analogous to the ones used for the proof of the upper bound (2.28).

By performing some manipulations on the expression of $\phi_\alpha(f,g)$ in (2.30) we see that, a.e.,

$$\begin{aligned}
\phi_\alpha(f,g) = g^{-1/2}\,&(\alpha I_s + \mathbf{i}\,g^{-1/2}\mathrm{Im}(f)g^{-1/2})^{-1}(\alpha I_s - g^{-1/2}\mathrm{Re}(f)g^{-1/2}) \\
&(\alpha I_s + g^{-1/2}\mathrm{Re}(f)g^{-1/2})^{-1}(\alpha I_s - \mathbf{i}\,g^{-1/2}\mathrm{Im}(f)g^{-1/2})\,g^{1/2}.
\end{aligned}$$

By similarity and by recalling that $\Lambda(AB) = \Lambda(BA)$ for all square matrices A, B of the same size, we obtain that, a.e.,

$$\rho(\phi_\alpha(f,g)) = \rho(\tilde{\phi}_\alpha(f,g)),$$

where

$$\begin{aligned}
\tilde{\phi}_\alpha(f,g) := &(\alpha I_s - g^{-1/2}\mathrm{Re}(f)g^{-1/2})(\alpha I_s + g^{-1/2}\mathrm{Re}(f)g^{-1/2})^{-1} \\
&(\alpha I_s - \mathbf{i}\,g^{-1/2}\mathrm{Im}(f)g^{-1/2})(\alpha I_s + \mathbf{i}\,g^{-1/2}\mathrm{Im}(f)g^{-1/2})^{-1}.
\end{aligned}$$

Now we observe that the matrix $(\alpha I_s - \mathbf{i}\,g^{-1/2}\mathrm{Im}(f)g^{-1/2})(\alpha I_s + \mathbf{i}\,g^{-1/2}\mathrm{Im}(f)g^{-1/2})^{-1}$ is unitary. This can be proved by direct computation, showing that the product of this matrix with its conjugate transpose is I_s (when verifying this, take into account that all the functions of the Hermitian matrix $g^{-1/2}\mathrm{Im}(f)g^{-1/2}$ commute). Using the unitary invariance of the spectral norm, the fact that the matrix $(\alpha I_s - g^{-1/2}\mathrm{Re}(f)g^{-1/2})(\alpha I_s + g^{-1/2}\mathrm{Re}(f)g^{-1/2})^{-1}$ is Hermitian, and the similarity relation $g^{-1/2}\mathrm{Re}(f)g^{-1/2} \sim h$, we obtain that, a.e.,

$$\begin{aligned}
\rho(\phi_\alpha(f,g)) &= \rho(\tilde{\phi}_\alpha(f,g)) \le \|\tilde{\phi}_\alpha(f,g)\| \\
&= \|(\alpha I_s - g^{-1/2}\mathrm{Re}(f)g^{-1/2})(\alpha I_s + g^{-1/2}\mathrm{Re}(f)g^{-1/2})^{-1}\| \\
&= \rho((\alpha I_s - g^{-1/2}\mathrm{Re}(f)g^{-1/2})(\alpha I_s + g^{-1/2}\mathrm{Re}(f)g^{-1/2})^{-1}) \\
&= \max_{\lambda\in\Lambda(h)} \left|\frac{\alpha-\lambda}{\alpha+\lambda}\right| = \max\left(\left|\frac{\alpha-\lambda_{\min}(h)}{\alpha+\lambda_{\min}(h)}\right|, \left|\frac{\alpha-\lambda_{\max}(h)}{\alpha+\lambda_{\max}(h)}\right|\right),
\end{aligned}$$

where in the last passage we invoked Lemma 5. Therefore,

$$\begin{aligned}
\|\rho(\phi_\alpha(f,g))\|_\infty &\le \left\|\max\left(\left|\frac{\alpha-\lambda_{\min}(h)}{\alpha+\lambda_{\min}(h)}\right|, \left|\frac{\alpha-\lambda_{\max}(h)}{\alpha+\lambda_{\max}(h)}\right|\right)\right\|_{L^\infty(I_k)} \\
&= \max\left(\left|\frac{\alpha-m_h}{\alpha+m_h}\right|, \left|\frac{\alpha-M_h}{\alpha+M_h}\right|\right),
\end{aligned}$$

where it is understood that $\frac{\alpha-M_h}{\alpha+M_h} = -1$ if $M_h = \infty$. In conclusion, (2.37) holds if $m_h = 0$ or $M_h = \infty$, because in these cases $\sigma^* = 1$, and (2.37) also holds in the case $0 < m_h \le M_h < \infty$, because

$$\begin{aligned}
\min_{\alpha>0} \|\rho(\phi_\alpha(f,g))\|_\infty &\le \min_{\alpha>0}\left[\max\left(\left|\frac{\alpha-m_h}{\alpha+m_h}\right|, \left|\frac{\alpha-M_h}{\alpha+M_h}\right|\right)\right] \\
&= \max\left(\left|\frac{\alpha-m_h}{\alpha+m_h}\right|, \left|\frac{\alpha-M_h}{\alpha+M_h}\right|\right)\Bigg|_{\alpha=\alpha^*} = \sigma^*,
\end{aligned}$$

where we invoked again Lemma 5. □

Proposition 11 gives a sufficient condition to met conditions in items 3(a) and 3(b). To prove it we need Lemma 6. In the remaining part of this subsection, if I, J are two intervals in $\mathbb{R}$, we denote by $I \times J$ the rectangle in $\mathbb{C}$ defined as

$$I \times J := \{z \in \mathbb{C} : \ \mathrm{Re}(z) \in I, \ \mathrm{Im}(z) \in J\}.$$

Lemma 6. *Let $f : G \subseteq \mathbb{R}^k \to \mathcal{M}_s$ be measurable and let $\mathcal{ENR}(f)$ be its essential numerical range. Then*

$$\mathcal{ENR}(f) \subseteq [m_{\mathrm{Re}(f)}, M_{\mathrm{Re}(f)}] \times [m_{\mathrm{Im}(f)}, M_{\mathrm{Im}(f)}].$$

In particular, $\mathrm{Coh}[\mathcal{ENR}(f)] \subseteq [m_{\mathrm{Re}(f)}, M_{\mathrm{Re}(f)}] \times [m_{\mathrm{Im}(f)}, M_{\mathrm{Im}(f)}]$.

Proof. Let r be a complex number such that $r \notin [m_{\mathrm{Re}(f)}, M_{\mathrm{Re}(f)}] \times [m_{\mathrm{Im}(f)}, M_{\mathrm{Im}(f)}]$. We show that $r \notin \mathcal{ENR}(f)$, i.e., that

$$\exists\, \epsilon > 0 : \ m_k\{t \in G : \ \exists\, v \in \mathbb{C}^s \text{ with } \|v\|_2 = 1 \text{ such that } v^* f(t) v \in D(r, \epsilon)\} = 0.$$

In other words, we have to prove the following result:

$$\exists\, \epsilon > 0 : \ \text{for a.e. } t \in G \text{ we have } v^* f(t) v \notin D(r, \epsilon) \text{ for all } v \in \mathbb{C}^s \text{ with } \|v\|_2 = 1. \tag{2.40}$$

Choose $\epsilon > 0$ such that $D(r, \epsilon) \bigcap [m_{\mathrm{Re}(f)}, M_{\mathrm{Re}(f)}] \times [m_{\mathrm{Im}(f)}, M_{\mathrm{Im}(f)}]$ is empty; note that such an ϵ exists because $r \notin [m_{\mathrm{Re}(f)}, M_{\mathrm{Re}(f)}] \times [m_{\mathrm{Im}(f)}, M_{\mathrm{Im}(f)}]$ and $[m_{\mathrm{Re}(f)}, M_{\mathrm{Re}(f)}] \times [m_{\mathrm{Im}(f)}, M_{\mathrm{Im}(f)}]$ is closed. By definition of $m_{\mathrm{Re}(f)}, M_{\mathrm{Re}(f)}, m_{\mathrm{Im}(f)}, M_{\mathrm{Im}(f)}$, for a.e. $t \in G$ we have

$$m_{\mathrm{Re}(f)} \le \lambda_{\min}(\mathrm{Re}(f(t))) \le \lambda_{\max}(\mathrm{Re}(f(t))) \le M_{\mathrm{Re}(f)},$$

$$m_{\mathrm{Im}(f)} \le \lambda_{\min}(\mathrm{Im}(f(t))) \le \lambda_{\max}(\mathrm{Im}(f(t))) \le M_{\mathrm{Im}(f)},$$

and so, by the minimax principle,

$$v^* f(t) v = v^* \mathrm{Re}(f(t)) v + \mathrm{i}\, v^* \mathrm{Im}(f(t)) v \in [m_{\mathrm{Re}(f)}, M_{\mathrm{Re}(f)}] \times [m_{\mathrm{Im}(f)}, M_{\mathrm{Im}(f)}] \qquad \forall\, v \in \mathbb{C}^s \text{ with } \|v\|_2 = 1;$$

in particular, $v^* f(t) v \notin D(r, \epsilon)$ for all $v \in \mathbb{C}^s$ with $\|v\|_2 = 1$. This concludes the proof of (2.40). □

Proposition 11. *Under the hypotheses of Theorem 18, if $m_g > 0$ it holds that*

(a) $0 \notin \mathrm{Coh}[\mathcal{ENR}(\alpha g + \mathbf{i}\,\mathrm{Im}(f))] \bigcup \mathrm{Coh}[\mathcal{ENR}(\alpha g + \mathrm{Re}(f))]$*, hence the last condition in item 3(a) of Theorem 18 is met;*

(b) in any decomposition of f, g of the form (2.31)–(2.32) *in which C is a unitary matrix, the last condition* (2.33) *in item 3(b) of Theorem 18 is met.*

Proof. Since g is Hermitian a.e.

$$\mathrm{Re}(\alpha g + \mathrm{i}\,\mathrm{Im}(f)) = \alpha g \qquad \text{a.e.},$$

$$\mathrm{Re}(\alpha g + \mathrm{Re}(f)) = \alpha g + \mathrm{Re}(f) \qquad \text{a.e.}$$

Since $\mathrm{Re}(f)$ is HPSD a.e., by the minimax principle we have $m_{\alpha g + \mathrm{Re}(f)} \ge m_{\alpha g} = \alpha m_g$. Thus, Lemma 6 implies that

$$\mathrm{Coh}[\mathcal{ENR}(\alpha g + \mathrm{i}\,\mathrm{Im}(f))] \bigcup \mathrm{Coh}[\mathcal{ENR}(\alpha g + \mathrm{Re}(f))] \subseteq [\alpha m_g, +\infty) \times \mathbb{R}, \tag{2.41}$$

and in particular, since m_g is assumed to be positive (and α is positive as well), 0 cannot belong to the union in the left-hand side. This proves part (a).

To prove part (b), we note that, if

$$f = C(f_1 \oplus \cdots \oplus f_p) C^{-1} \qquad \text{a.e.},$$

$$g = C(g_1 \oplus \cdots \oplus g_p) C^{-1} \qquad \text{a.e.}$$

are decompositions of f, g in which the matrix C is unitary, then $C^{-1} = C^*$ and:

- $\mathrm{Re}(f_1) \oplus \cdots \oplus \mathrm{Re}(f_p) = \mathrm{Re}(f_1 \oplus \cdots \oplus f_p) = \mathrm{Re}(C^* f C) = C^* \mathrm{Re}(f) C$ a.e.; in particular, $\mathrm{Re}(f_1), \ldots, \mathrm{Re}(f_p)$ are HPSD a.e., because $\mathrm{Re}(f)$ is HPSD a.e. by hypothesis;

- $g_1 \oplus \cdots \oplus g_p = C^* g C$ a.e., so $g_1, \ldots, g_p$ are Hermitian a.e. because g is Hermitian a.e. by hypothesis;
- $m_{g_v} \geq m_g$ for all $v = 1, \ldots, p$, because the eigenvalues of g are $\lambda_i(g_v)$, $i = 1, \ldots, t_v$, $v = 1, \ldots, p$, where t_v is the size of g_v.

Therefore, by replicating the above argument which allowed us to obtain (2.41), we see that $\forall\, v = 1, \ldots, p$

$$\mathrm{Coh}[\mathcal{ENR}(\alpha g_v + \mathrm{i}\,\mathrm{Im}(f_v))] \bigcup \mathrm{Coh}[\mathcal{ENR}(\alpha g_v + \mathrm{Re}(f_v))] \subseteq [\alpha m_{g_v}, +\infty) \times \mathbb{R} \subseteq [\alpha m_g, +\infty) \times \mathbb{R},$$

and, recalling that $m_g > 0$, we see that part (b) is proved. □

Remark 12. With Remark 2 in mind, assuming the validity of (2.29), we expect that

$$\lim_{n\to\infty} \rho(M_n(\alpha)) = \|\rho(\phi_\alpha(f, g))\|_\infty. \tag{2.42}$$

Consequently, the optimal PHSS preconditioner $T_n(g)$ and the optimal parameter α for $T_n(f)$ are found by determining the couple (g, α) that minimizes $\|\rho(\phi_\alpha(f, g))\|_\infty$, subject to the constraints that $\alpha > 0$ and g is HPSD with $\lambda_{\min}(g)$ not a.e. equal to 0. Although it may seem that this heuristic idea will hardly work, the numerical experiments in the next subsection show that, in practice, it can be quite successful.

2.2.3 Numerical results

This subsection contains numerical examples that illustrate the effectiveness of the provided theoretical analysis and of the PHSS with preconditioner chosen according to our eigenvalue distribution results. In particular, we include an example coming from the approximation of PDEs (Case 11), while one of the numerical experiments (Case 9) also shows that the upper bound $\sigma_n(\alpha)$ may be a crude estimate for $\rho(M_n(\alpha))$ and that the best parameter α_n^* for $\sigma_n(\alpha)$ need not to be the best parameter for $\rho(M_n(\alpha))$. This confirms that the effects of the imaginary part $\mathrm{Im}(T_n(f))$ in the PHSS iteration matrix (2.9) can be significant.

Univariate examples

Fixed $s = 2$ and $k = 1$, we will consider matrix-valued functions $f, g \in L^\infty(1, 2)$ of the form

$$f(x) = Q(x)A(x)Q^*(x),$$
$$g(x) = Q(x)B(x)Q^*(x),$$

where $A(x)$, $B(x)$, $Q(x)$ belong to $L^\infty(1, 2)$ and $Q(x)$ is unitary a.e. In the examples, we focus our attention on the spectral behavior of the PHSS iteration matrices $M_n(\alpha)$ for different sizes n, on the minimization of $\|\rho(\phi_\alpha(f, g))\|_\infty$ mentioned in Remark 12, and on the solution of the linear system with coefficient matrix $T_n(f)$ and random right-hand side. From a computational point of view, to solve such systems, we implemented both the HSS and the PHSS with preconditioner $T_n(g)$, using a tolerance of 10^{-6} and an initial guess equal to the zero vector. The choice of the parameter α will be specified each time.

In the first case considered below, we will be concerned with the validation of the theory developed so far, with particular attention to the eigenvalue distribution result (2.29). We will consider the clustering of $M_n(\alpha)$ at the essential range $\mathcal{ER}(\phi_\alpha(f, g))$ as the main visual indicator that (2.29) holds; see Remark 3 for a motivation. The other considered indicator is the limit relation (2.42); see Remark 12. If the assumptions of item 3 in Theorem 18 are fulfilled, we shall see that $\{M_n(\alpha)\}_n$ is clustered at $\mathcal{ER}(\phi_\alpha(f, g))$ and (2.42) holds. In the considered example, it turns out that the same is true even if the assumptions in Theorem 18 are violated. This shows that the minimization of $\|\rho(\phi_\alpha(f, g))\|_\infty$ to find optimal PHSS preconditioners $T_n(g)$ and optimal parameters α for $T_n(f)$ (cf. Remark 12) makes sense even if we are not sure that (2.29) holds. In the second two cases considered below, we will focus on this minimization problem and we will see that the couple (g, α) obtained in this way leads to efficient PHSS methods for solving linear systems with coefficient matrix $T_n(f)$.

Case 7. Let

$$A(x) := \begin{pmatrix} 2 + \mathbf{i} + \cos x & 0 \\ 0 & 1 - \beta \mathrm{e}^{\mathbf{i}x} \end{pmatrix}, \quad B(x) := \begin{pmatrix} 1 & 0 \\ 0 & 1 - \beta \cos x \end{pmatrix},$$

where $0 < \beta \leq 1$, and define

$$f(x) := Q(x)A(x)Q^*(x),$$
$$g(x) := Q(x)B(x)Q^*(x).$$

Due to the assumption $\beta \in (0, 1]$, the matrices $g(x)$ and

$$\mathrm{Re}(f(x)) = Q(x) \begin{pmatrix} 2+\cos x & 0 \\ 0 & 1-\beta\cos x \end{pmatrix} Q^*(x)$$

are HPD for a.e. $x \in I_1 = (-\pi, \pi)$. Using Proposition 1, it follows that $\mathrm{Re}(T_n(f)) = T_n(\mathrm{Re}(f))$ and $T_n(g)$ are HPD for all n. Therefore, we can use the PHSS with preconditioner $T_n(g)$ for solving a linear system with coefficient matrix $T_n(f)$.

If $0 < \beta < 1$, then $m_g = 1 - \beta > 0$, and so, by Proposition 11, all the hypotheses of item 3(a) in Theorem 18 are satisfied except, possibly, the assumption that $\phi_\alpha(f, g)$ is in the Tilli class. By direct computation,

$$\begin{aligned}\phi_\alpha(f(x), g(x)) &= Q(x)\phi_\alpha(A(x), B(x))Q^*(x) \\ &= Q(x) \begin{pmatrix} \frac{\alpha-\mathbf{i}}{\alpha+\mathbf{i}} \cdot \frac{\alpha-2-\cos x}{\alpha+2+\cos x} & 0 \\ 0 & \frac{\alpha-1}{\alpha+1} \cdot \frac{\alpha+\frac{\mathbf{i}\beta\sin x}{1-\beta\cos x}}{\alpha-\frac{\mathbf{i}\beta\sin x}{1-\beta\cos x}} \end{pmatrix} Q^*(x).\end{aligned}$$

This implies that

$$\begin{aligned}\mathcal{ER}(\phi_\alpha(f, g)) &= \mathcal{ER}(\lambda_1(\phi_\alpha(f, g))) \cup \mathcal{ER}(\lambda_2(\phi_\alpha(f, g))) \\ &= \mathcal{ER}(\phi_\alpha(A, B)_{1,1}) \cup \mathcal{ER}(\phi_\alpha(A, B)_{2,2}).\end{aligned}$$

A necessary (but in general not sufficient) condition for $\phi_\alpha(f, g)$ to be in the Tilli class is that both $\phi_\alpha(A, B)_{1,1}$ and $\phi_\alpha(A, B)_{2,2}$ are in the Tilli class. Expressing $\phi_\alpha(A, B)_{1,1}$ and $\phi_\alpha(A, B)_{2,2}$ in polar form, we have:

(i) $\phi_\alpha(A, B)_{1,1} = \psi_\alpha \mathrm{e}^{-2\mathbf{i}\varphi_\alpha}$, where $\varphi_\alpha := \arctan\frac{1}{\alpha}$ is constant and $\psi_\alpha(x) := \frac{\alpha-2-\cos x}{\alpha+2+\cos x}$ depends on x (so it describes a line segment with slope $\tan(-2\varphi_\alpha)$). Since $\psi_\alpha(x)$ is even and monotone increasing over $[0, \pi]$ (its derivative is nonnegative on this interval), it follows that

$$\inf_{x\in I_1} \psi_\alpha(x) = \psi_\alpha(0) = \frac{\alpha-3}{\alpha+3}, \qquad \sup_{x\in I_1} \psi_\alpha(x) = \psi_\alpha(\pi) = \frac{\alpha-1}{\alpha+1}.$$

Consequently, $\mathcal{ER}(\phi_\alpha(A, B)_{1,1})$ is the (closed) line segment in $\mathbb{C}$ whose extremes are the points $\frac{\alpha-3}{\alpha+3}\mathrm{e}^{-2\mathbf{i}\varphi_\alpha}$, $\frac{\alpha-1}{\alpha+1}\mathrm{e}^{-2\mathbf{i}\varphi_\alpha}$. Note that this segment passes through the origin when $\alpha \in [1, 3]$;

(ii) $\phi_\alpha(A, B)_{2,2} = \frac{\alpha-1}{\alpha+1}\mathrm{e}^{2\mathbf{i}\theta_\alpha}$, where $\theta_\alpha(x) := \arctan\left(\frac{\beta\sin x}{\alpha(1-\beta\cos x)}\right)$ depends on x and $\frac{\alpha-1}{\alpha+1}$ is constant (so it describes an arc of a circle centered at the origin and with constant radius). If $\beta = 1$, the function $\xi_\alpha(x) := \frac{\beta\sin x}{\alpha(1-\beta\cos x)}$ assumes all values in $\mathbb{R}$ for $x \in I_1$, implying that $\mathcal{ER}(\phi_\alpha(A, B)_{2,2})$ is the circle centered at 0 and with radius $\left|\frac{\alpha-1}{\alpha+1}\right|$. If $\beta \in (0, 1)$, then $\xi_\alpha(x)$ is odd and bounded over I_1, vanishes at the points $x = -\pi, 0, \pi$, and its derivative is nonnegative if and only if $-\arccos\beta \leq x \leq \arccos\beta$. Hence,

$$\inf_{x\in I_1} \xi_\alpha(x) = \xi_\alpha(-\arccos\beta) = -\frac{\beta}{\alpha\sqrt{1-\beta^2}}, \qquad \sup_{x\in I_1} \xi_\alpha(x) = \xi_\alpha(\arccos\beta) = \frac{\beta}{\alpha\sqrt{1-\beta^2}},$$

and so $\mathcal{ER}(\phi_\alpha(A, B)_{2,2})$ is the (closed) arc centered at 0 and with radius $\left|\frac{\alpha-1}{\alpha+1}\right|$, which passes through the real point $\frac{\alpha-1}{\alpha+1}$ (for $x = 0$) and whose extremes are given by

$$\frac{\alpha-1}{\alpha+1}\mathrm{e}^{-2\mathbf{i}\arctan\frac{\beta}{\alpha\sqrt{1-\beta^2}}}, \ \frac{\alpha-1}{\alpha+1}\mathrm{e}^{2\mathbf{i}\arctan\frac{\beta}{\alpha\sqrt{1-\beta^2}}}.$$

Figure 2.8 shows the essential ranges of $\phi_\alpha(A, B)_{1,1}$ and $\phi_\alpha(A, B)_{2,2}$. From (i) we deduce that $\mathcal{ER}(\phi_\alpha(A, B)_{1,1})$ does not disconnect the complex plane and has empty interior, i.e., $\phi_\alpha(A, B)_{1,1}$ is in the Tilli class. Concerning (ii), if $\beta = 1$, then $\mathcal{ER}(\phi_\alpha(A, B)_{2,2})$ disconnects the complex plane and

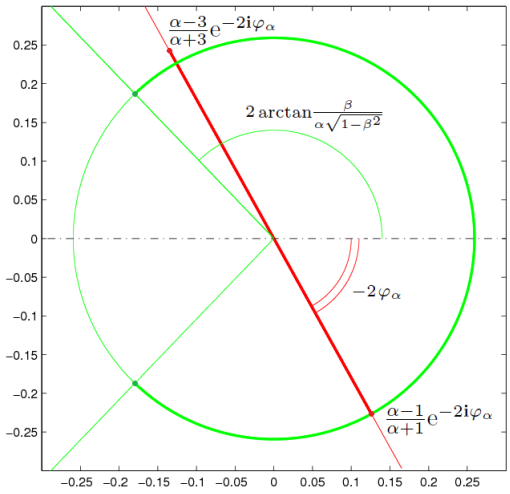


Figure 2.8: Essential ranges of $\phi_\alpha(A,B)_{1,1}$ (red) and $\phi_\alpha(A,B)_{2,2}$ (green). The figure refers to the case $\alpha \in (1,3]$ and $\beta \in (0,1)$.

$\phi_\alpha(A,B)_{2,2}$ is not in the Tilli class; otherwise, if $\beta \in (0,1)$, then $\phi_\alpha(A,B)_{2,2}$ is in the Tilli class. We then conclude that both $\phi_\alpha(A,B)_{1,1}$ and $\phi_\alpha(A,B)_{2,2}$ are in the Tilli class when $\beta \in (0,1)$, while for $\beta = 1$, $\phi_\alpha(A,B)_{2,2}$, and hence also $\phi_\alpha(f,g)$, is not in the Tilli class. From the analysis in (i)–(ii), we also obtain that

$$\|\rho(\phi_\alpha(f,g))\|_\infty = \max\left(\left|\frac{\alpha-1}{\alpha+1}\right|, \left|\frac{\alpha-3}{\alpha+3}\right|\right),$$

which is independent of β.

The previous discussion does not depend on the unitary matrix $Q(x)$. Let us now consider two different choices for $Q(x)$.

Subcase 1.1. Let $Q(x) = Q$ be a constant transformation. Using item 3(b) in Theorem 18 together with Proposition 11, we have $\{M_n(\alpha)\}_n \sim_\lambda (\phi_\alpha(f,g), I_1)$ if $\phi_\alpha(A,B)_{1,1}$ and $\phi_\alpha(A,B)_{2,2}$ are in the Tilli class, and this is verified for $\beta \in (0,1)$.

In the numerical tests shown in Figures 2.9–2.11, we fixed

$$Q = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix} \tag{2.43}$$

and we chose $\alpha = \alpha^* = \sqrt{m_h M_h} = \sqrt{3}$, where

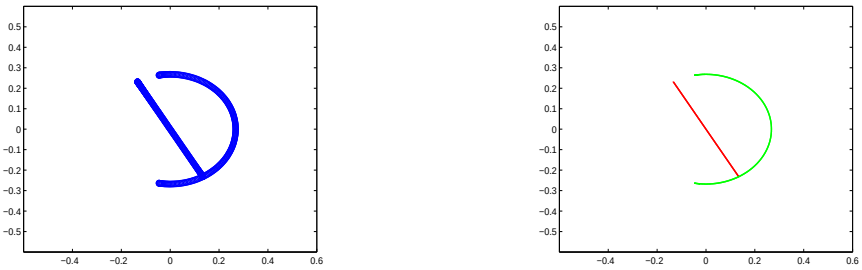
$$h(x) := g^{-1}(x)\mathrm{Re}(f(x)) = Q\begin{pmatrix} 2+\cos x & 0 \\ 0 & 1 \end{pmatrix} Q^*$$

is independent of β (cf. item 4 in Theorem 18). We considered three choices of the parameter β:

- $\beta = 0.9$. Both $\phi_\alpha(A,B)_{1,1}$ and $\phi_\alpha(A,B)_{2,2}$ are in the Tilli class. Figures 2.9(a) and 2.9(b) refer to the eigenvalues in the complex plane of $M_{200}(\alpha)$ and to the eigenvalue functions of $\phi_\alpha(f,g)$, respectively. Note that also $\phi_\alpha(f,g)$ is in the Tilli class. As expected, the eigenvalues of $M_{200}(\alpha)$ are distributed as $\lambda_1(\phi_\alpha(f,g)) = \phi_\alpha(A,B)_{1,1}$ and $\lambda_2(\phi_\alpha(f,g)) = \phi_\alpha(A,B)_{2,2}$, and in fact they are clustered at the union of the ranges of $\lambda_1(\phi_\alpha(f,g))$ and $\lambda_2(\phi_\alpha(f,g))$, which is the essential range of $\phi_\alpha(f,g)$;

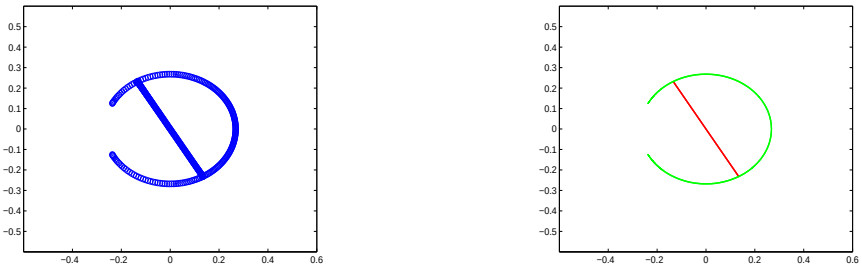
- $\beta = 0.99$. Both $\phi_\alpha(A,B)_{1,1}$ and $\phi_\alpha(A,B)_{2,2}$ are in the Tilli class, but $\phi_\alpha(f,g)$ is not (see Figure 2.10(b)). In spite of this, as predicted by item 3(b) in Theorem 18, the eigenvalues of $M_{200}(\alpha)$ are distributed as the eigenvalue functions of $\phi_\alpha(f,g)$ (see Figure 2.10(a));

- $\beta = 1$. In this case $\phi_\alpha(A,B)_{1,1}$ is in the Tilli class, but $\phi_\alpha(A,B)_{2,2}$ is not. We cannot apply neither item 3(a) nor item 3(b) in Theorem 18, but, as shown in Figures 2.11(a) and 2.11(b), the eigenvalues of $M_{200}(\alpha)$ are distributed as the eigenvalue functions of $\phi_\alpha(f,g)$. Therefore we guess that, even in this case, $\{M_n(\alpha)\}_n \sim_\lambda (\phi_\alpha(f,g), I_1)$.



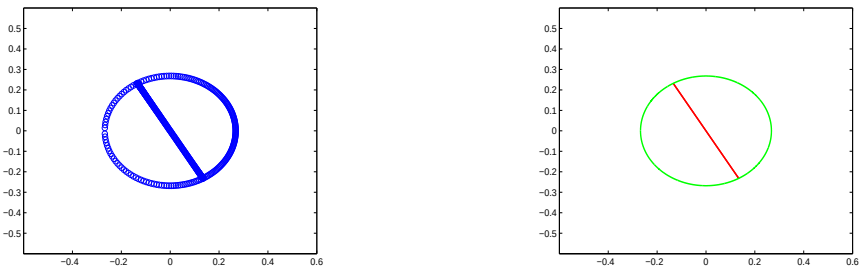
(a) (b)

Figure 2.9: (a) Eigenvalues in the complex plane of $M_n(\alpha)$ fixed $\alpha = \sqrt{3}$, $\beta = 0.9$ and $n = 200$; (b) Essential ranges of $\lambda_1(\phi_\alpha(f,g))$ (red) and $\lambda_2(\phi_\alpha(f,g))$ (green) for $\alpha = \sqrt{3}$, $\beta = 0.9$ and $n = 200$.



(a) (b)

Figure 2.10: (a) Eigenvalues in the complex plane of $M_n(\alpha)$ fixed $\alpha = \sqrt{3}$, $\beta = 0.99$ and $n = 200$; (b) Essential ranges of $\lambda_1(\phi_\alpha(f,g))$ (red) and $\lambda_2(\phi_\alpha(f,g))$ (green) for $\alpha = \sqrt{3}$, $\beta = 0.99$ and $n = 200$.

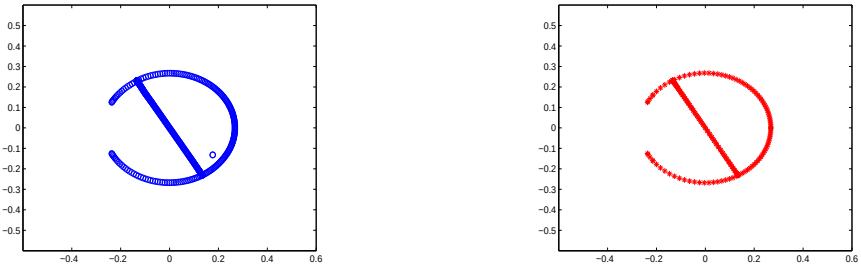


(a) (b)

Figure 2.11: (a) Eigenvalues in the complex plane of $M_n(\alpha)$ fixed $\alpha = \sqrt{3}$, $\beta = 1$ and $n = 200$; (b) Essential ranges of $\lambda_1(\phi_\alpha(f,g))$ (red) and $\lambda_2(\phi_\alpha(f,g))$ (green) for $\alpha = \sqrt{3}$, $\beta = 1$ and $n = 200$.

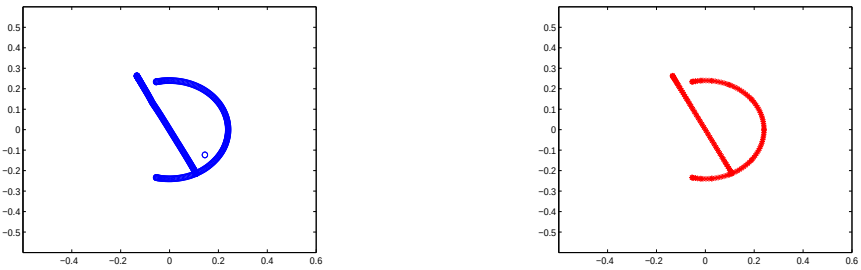
	$\alpha = \sqrt{3}$		$\alpha = 3$	
n	$\beta = 0.9$	$\beta = 1$	$\beta = 0.9$	$\beta = 1$
50	0.2679	0.2679	0.5000	0.5000
100	0.2679	0.2679	0.5000	0.5000
200	0.2679	0.2679	0.5000	0.5000
400	0.2679	0.2679	0.5000	0.5000
	$\|\rho(\phi_\alpha(f,g))\|_\infty = \frac{\sqrt{3}-1}{\sqrt{3}+1}$		$\|\rho(\phi_\alpha(f,g))\|_\infty = \frac{1}{2}$	

Table 2.6: Spectral radius of the PHSS iteration matrix fixed $Q(x) = Q$ as in (2.43).



(a) (b)

Figure 2.12: (a) Eigenvalues in the complex plane of $M_n(\alpha)$ fixed $\alpha = \sqrt{3}$, $\beta = 0.99$ and $n = 200$; (b) Essential range of $\phi_\alpha(f, g)$ for $\alpha = \sqrt{3}$, $\beta = 0.99$ and $n = 200$.



(a) (b)

Figure 2.13: (a) Eigenvalues in the complex plane of $M_n(\alpha)$ fixed $\alpha = \sqrt{3} - 0.1$, $\beta = 0.9$ and $n = 200$; (b) Essential range of $\phi_\alpha(f, g)$ fixed $\alpha = \sqrt{3} - 0.1$, $\beta = 0.9$ and $n = 200$.

Table 2.6 shows the spectral radius of the PHSS iteration matrix $M_n(\alpha)$ for different values of α and β, fixed $Q(x) = Q$ as in (2.43). In all the considered cases, the limit relation (2.42) holds. Note that this could be expected for $\alpha = \sqrt{3}$, simply looking at Figures 2.9 and 2.11.

Subcase 1.2. Let

$$Q(x) = \begin{pmatrix} \cos x & \sin x \\ -\sin x & \cos x \end{pmatrix}. \tag{2.44}$$

This unitary transformation depends on x and so we cannot apply item 3(b) of Theorem 18 as we did in Subcase 1.1. However, we can apply item 3(a) whenever $\phi_\alpha(f, g)$ is in the Tilli class. We know from the discussion before Subcase 1.1 that $\phi_\alpha(f, g)$ is not in the Tilli class for $\beta = 1$. For $\beta \in (0, 1)$, to understand when $\phi_\alpha(f, g)$ is not in the Tilli class, we study when the essential ranges of $\phi_\alpha(A, B)_{1,1}$ and $\phi_\alpha(A, B)_{2,2}$ have two intersections, like in Figures 2.10(b) and 2.11(b). Recalling the discussion in (i)–(ii), see also Figure 2.8, they have two intersections when the following conditions are satisfied:

1. $\frac{\alpha-3}{\alpha+3} \leq -\frac{\alpha-1}{\alpha+1} < \frac{\alpha-1}{\alpha+1} \iff \alpha \in (1, \sqrt{3}]$;
2. $2 \arctan \frac{\beta}{\alpha\sqrt{1-\beta^2}} \geq -2\varphi_\alpha + \pi \ \wedge \ -2 \arctan \frac{\beta}{\alpha\sqrt{1-\beta^2}} \leq -2\varphi_\alpha$, which, assuming $\alpha \geq 1$ and recalling the identities $\varphi_\alpha = \arctan \frac{1}{\alpha}$ and $\arctan \alpha + \arctan \frac{1}{\alpha} = \frac{\pi}{2}$, is equivalent to $\beta \geq \eta(\alpha) := \frac{\alpha^2}{\sqrt{1+\alpha^4}}$.

We then conclude that, for $\beta \in (0, 1)$, $\phi_\alpha(f, g)$ is not in the Tilli class if and only if $\alpha \in (1, \sqrt{3}]$ and $\beta \in [\eta(\alpha), 1)$.

In the numerical experiments shown in Figures 2.12–2.13, we considered two different choices of α and β:

- $\alpha = \alpha^* = \sqrt{3}$ (the best asymptotic parameter suggested by Theorem 18) and $\beta = 0.99$. Since $\eta(\sqrt{3}) \approx 0.95 < 0.99$, for this choice of α and β, both $\phi_\alpha(A, B)_{1,1}$ and $\phi_\alpha(A, B)_{2,2}$ are in the Tilli class, but $\phi_\alpha(f, g)$ is not (see Figure 2.12(b)). Although Theorem 18 cannot be applied, a comparison between Figures 2.12(a) and 2.12(b) shows that the eigenvalues of $M_{200}(\alpha)$ are clustered at the essential range of $\phi_\alpha(f, g)$. So, we guess that $\{M_n(\alpha)\}_n \sim_\lambda (\phi_\alpha(f, g), I_1)$.

	$\alpha = 0.5$		$\alpha = 2$	
n	$\beta = 0.3$	$\beta = 0.7$	$\beta = 0.3$	$\beta = 0.7$
50	0.7141	0.7141	0.3333	0.3333
100	0.7142	0.7142	0.3333	0.3333
200	0.7143	0.7143	0.3333	0.3333
400	0.7143	0.7143	0.3333	0.3333
	$\|\rho(\phi_\alpha(f,g))\|_\infty = \frac{5}{7}$		$\|\rho(\phi_\alpha(f,g))\|_\infty = \frac{1}{3}$	

Table 2.7: Spectral radius of the PHSS iteration matrix fixed $Q(x)$ as in (2.44).

- $\alpha = \sqrt{3} - 0.1$, $\beta = 0.9$. Since $\eta(\sqrt{3} - 0.1) \approx 0.94 > 0.9$, we know that $\phi_\alpha(f,g)$ is in the Tilli class. This is confirmed by Figure 2.13(b). Furthermore, as predicted by the theory, the eigenvalues of $M_{200}(\alpha)$ are distributed as the eigenvalue functions of $\phi_\alpha(f,g)$ (see Figure 2.13(a)).

Table 2.7 shows the spectral radius of the PHSS iteration matrix $M_n(\alpha)$ for different values of α and β, fixed $Q(x)$ as in (2.44). Even in this case, the limit relation (2.42) holds.

Case 8. Let

$$\begin{aligned} f(x) &:= Q(x) \begin{pmatrix} 1 & 0 \\ 0 & |x| + 1 + \mathbf{i} \end{pmatrix} Q^*(x), \\ g(x) &:= Q(x) \begin{pmatrix} 1 & 0 \\ 0 & a + b\cos x \end{pmatrix} Q^*(x), \end{aligned} \tag{2.45}$$

where a, b are real parameters such that $a > |b| \geq 0$. We note that $g(x)$ and

$$\mathrm{Re}(f(x)) = Q(x) \begin{pmatrix} 1 & 0 \\ 0 & |x| + 1 \end{pmatrix} Q^*(x)$$

are HPD for all $x \in I_1$. Therefore, setting $A_n := T_n(f)$ and $P_n := T_n(g)$, the following results hold (cf. Theorem 18).

i. $\mathrm{Re}(A_n)$, P_n are HPD for all $n \in \mathbb{N}$, so we can apply to A_n the PHSS with preconditioner P_n.

ii. For the PHSS iteration matrix $M_n(\alpha)$ in (2.27) we have

$$\rho(M_n(\alpha)) \leq \sigma_n(\alpha),$$

where $\sigma_n(\alpha)$ is given in (2.28).

iii. Defining

$$h(x) := g^{-1}(x)\mathrm{Re}(f(x)) = Q(x) \begin{pmatrix} 1 & 0 \\ 0 & \dfrac{|x| + 1}{a + b\cos x} \end{pmatrix} Q^*(x)$$

and

$$\xi := \inf_{x \in I_1} \frac{|x| + 1}{a + b\cos x} = \min_{x \in [0,\pi]} \frac{|x| + 1}{a + b\cos x}, \quad \eta := \sup_{x \in I_1} \frac{|x| + 1}{a + b\cos x} = \max_{x \in [0,\pi]} \frac{|x| + 1}{a + b\cos x}, \tag{2.46}$$

we have

$$m_h = \min(1, \xi) > 0, \qquad M_h = \max(1, \eta) < \infty.$$

Moreover,

$$\lambda_{\min}(P_n^{-1} A_n) \to m_h, \qquad \lambda_{\max}(P_n^{-1} A_n) \to M_h,$$

and for the values α_n^*, κ_n, $\sigma_n(\alpha_n^*)$ in (1.32)–(1.34) the limit relations (2.35) holds, i.e.,

$$\alpha_n^* \to \alpha^* := \sqrt{m_h M_h} = \sqrt{\min(1, \xi)\max(1, \eta)}, \tag{2.47}$$

$$\kappa_n \to \kappa := \frac{M_h}{m_h} = \frac{\max(1, \eta)}{\min(1, \xi)}, \tag{2.48}$$

$$\sigma_n(\alpha_n^*) \to \sigma^* := \frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1}. \tag{2.49}$$

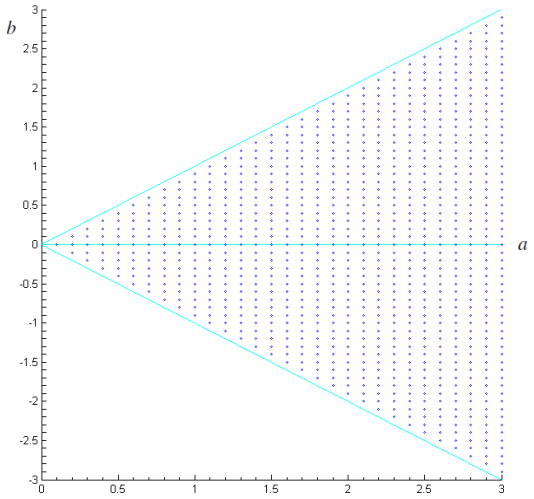


Figure 2.14: Graphical representation of $\mathcal{D}'$.

Suppose now that

$$\{M_n(\alpha)\}_n \sim_\lambda (\phi_\alpha(f,g), I_1). \tag{2.50}$$

Since $m_g = \min(1, a - |b|) > 0$, (2.50) certainly holds if $\phi_\alpha(f,g)$ belongs to the Tilli class, but we have seen in Case 7 that (2.50) may hold even if this condition is not satisfied. Assuming (2.50), we expect that, for large n,

$$\rho(M_n(\alpha)) \approx \|\rho(\phi_\alpha(f,g))\|_\infty; \tag{2.51}$$

see Remark 12, in particular equation (2.42). In this case, the quantity $\|\rho(\phi_\alpha(f,g))\|_\infty$ is an estimate of the asymptotic PHSS convergence rate.

Motivated by this observation, in this example we consider the problem of determining a, b, α (i.e. g, α) that minimize the $\|\rho(\phi_\alpha(f,g))\|_\infty$. By direct computation, we obtain

$$\phi_\alpha(f(x), g(x)) = Q(x) \begin{pmatrix} \dfrac{\alpha-1}{\alpha+1} & 0 \\ 0 & \dfrac{\alpha(a+b\cos x)-\mathbf{i}}{\alpha(a+b\cos x)+\mathbf{i}} \cdot \dfrac{\alpha - \dfrac{|x|+1}{a+b\cos x}}{\alpha + \dfrac{|x|+1}{a+b\cos x}} \end{pmatrix} Q^*(x)$$

and

$$\rho(\phi_\alpha(f(x), g(x))) = \max\left(\left|\frac{\alpha-1}{\alpha+1}\right|, \left| \frac{\alpha - \dfrac{|x|+1}{a+b\cos x}}{\alpha + \dfrac{|x|+1}{a+b\cos x}} \right| \right). \tag{2.52}$$

Noting that for $\alpha > 0$ and $M \geq m \geq 0$ we have $\max_{t\in[m,M]} \left|\frac{\alpha-t}{\alpha+t}\right| = \max\left(\left|\frac{\alpha-m}{\alpha+m}\right|, \left|\frac{\alpha-M}{\alpha+M}\right| \right)$, from (2.52) we get

$$\|\rho(\phi_\alpha(f,g))\|_\infty = \max\left(\left|\frac{\alpha-1}{\alpha+1}\right|, \left|\frac{\alpha-\xi}{\alpha+\xi}\right|, \left|\frac{\alpha-\eta}{\alpha+\eta}\right| \right), \tag{2.53}$$

where $\xi := \xi(a,b)$ and $\eta := \eta(a,b)$ are defined in (2.46). We plan to minimize the quantity $\|\rho(\phi_\alpha(f,g))\|_\infty$ over the set

$$\mathcal{D} := \{(a,b,\alpha) : \ (a,b) \in \mathcal{D}', \ \alpha > 0\} = \mathcal{D}' \times (0,\infty),$$

where

$$\mathcal{D}' := \{(a,b) : \ a = 0.1,\, 0.2,\, 0.3,\, \ldots,\, 3, \ b = -a + 0.1,\, \ldots, a - 0.1\}$$

(the analytic minimization of $\|\rho(\phi_\alpha(f,g))\|_\infty$ over $\{(a,b,\alpha) : \ a > |b| \geq 0, \ \alpha > 0\}$ is too complicated). Figure 2.14 shows the set $\mathcal{D}'$. From Lemma 5 and from (2.53) we have

$$\min_{\alpha>0} \|\rho(\phi_\alpha(f,g))\|_\infty = \|\rho(\phi_{\alpha^*}(f,g))\|_\infty = \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1} = \sigma^*, \tag{2.54}$$

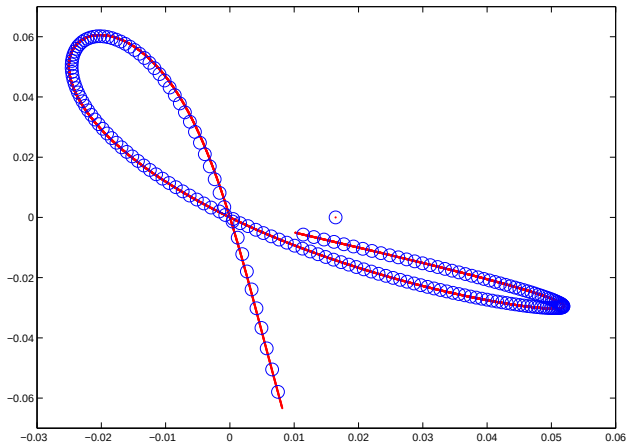


Figure 2.15: Eigenvalues in the complex plane of $M_{200}(\hat{\alpha}^*)$ (blue) and essential range of $\phi_{\hat{\alpha}^*}(f,\hat{g})$ (red) fixed $Q(x) = I_2$.

n	HSS ($\alpha = \sqrt{\pi+1}$)	PHSS ($g = \hat{g},\ \alpha = \hat{\alpha}^* \approx 1.033$)
50	0.3410	0.0625
100	0.3410	0.0632
200	0.3410	0.0636
400	0.3410	0.0638

Table 2.8: Spectral radius of the HSS and PHSS iteration matrices fixed $Q(x) = I_2$.

where $\alpha^* := \alpha^*(a,b),\ \kappa := \kappa(a,b),\ \sigma^* := \sigma^*(a,b)$ are given in (2.47)–(2.49). Hence, in this example, the inequality (2.37) holds as an equality independently of (a,b). Using the computer, we obtain

$$\min_{(a,b,\alpha)\in\mathcal{D}} \|\rho(\phi_\alpha(f,g))\|_\infty = \min_{(a,b)\in\mathcal{D}'} \sigma^*(a,b) = \sigma^*(\hat{a},\hat{b}) =: \hat{\sigma}^* \approx 0.064, \tag{2.55}$$

where

$$(\hat{a},\hat{b}) := (2.6,-1.5). \tag{2.56}$$

Summarizing: the minimum of $\|\rho(\phi_\alpha(f,g))\|_\infty$ over $\mathcal{D}$ is obtained for $(a,b,\alpha) = (\hat{a},\hat{b},\hat{\alpha}^*)$, where $\hat{\alpha}^* := \alpha^*(\hat{a},\hat{b}) \approx 1.033$. The minimizing couple (g,α) is then $(\hat{g},\hat{\alpha}^*)$, where $\hat{g} := g_{\hat{a},\hat{b}}$ is the function g obtained from (2.45) for $(a,b) = (\hat{a},\hat{b})$. The minimum value $\|\rho(\phi_{\hat{\alpha}^*}(f,\hat{g}))\|_\infty$ is given by (2.55) and coincides with the minimum $\hat{\sigma}^*$ of the best asymptotic upper bound σ^* over $\mathcal{D}'$.

Remark 13. Recalling (2.36) and taking into account the limit relation $\alpha_n^* \to \alpha^*$, we expect that

$$\limsup_{n\to\infty} \rho(M_n(\alpha^*)) \le \sigma^* = \|\rho(\phi_{\alpha^*}(f,g))\|_\infty. \tag{2.57}$$

In particular, for $(a,b) = (\hat{a},\hat{b})$, (2.57) becomes

$$\limsup_{n\to\infty} \rho(M_n(\hat{\alpha}^*)) \le \hat{\sigma}^* = \|\rho(\phi_{\hat{\alpha}^*}(f,\hat{g}))\|_\infty \approx 0.064.$$

We numerically verified for $Q(x)$ equal to the 2×2 identity matrix I_2 that the previous inequality holds as an equality and with 'lim sup' replaced by 'lim':

$$\lim_{n\to\infty} \rho(M_n(\hat{\alpha}^*)) = \|\rho(\phi_{\hat{\alpha}^*}(f,\hat{g}))\|_\infty \tag{2.58}$$

(see Table 2.8). We also verified that $M_{200}(\hat{\alpha}^*)$ is clustered at $\mathcal{ER}(\phi_{\hat{\alpha}^*}(f,\hat{g}))$; see Figure 2.15. This indicates that, for $(a,b,\alpha) = (\hat{a},\hat{b},\hat{\alpha}^*)$, (2.50) holds and hence (2.58) is not surprising; see the discussion between (2.50)–(2.51) and recall (2.42).

As shown in Tables 2.8–2.9, the PHSS method corresponding to the optimal choice $(g,\alpha) = (\hat{g},\hat{\alpha}^*)$ outperforms the HSS method with the best asymptotic parameter $\sqrt{\pi+1}$ suggested by Theorem 18, which is obtained from (2.35) in the case where $g = I_2$ and $h = \mathrm{Re}(f)$ (note that $m_{\mathrm{Re}(f)} = 1,\ M_{\mathrm{Re}(f)} = \pi+1$).

n	HSS ($\alpha = \sqrt{\pi+1}$)	PHSS ($g = \hat{g},\ \alpha = \hat{\alpha}^* \approx 1.033$)
50	14	6
100	14	6
200	14	6
400	14	6

Table 2.9: Number of HSS and PHSS iterations for $T_n(f)$ fixed $Q(x) = I_2$.

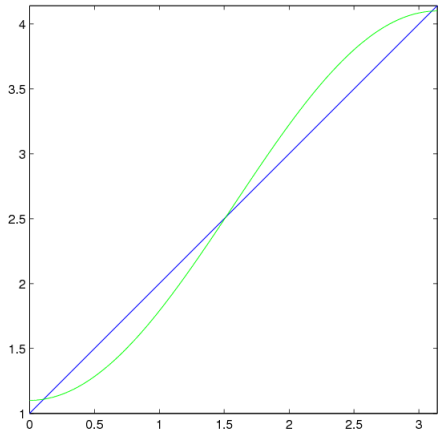


Figure 2.16: Graphs of the functions $x \mapsto |x| + 1$ and $x \mapsto 2.6 - 1.5\cos x$ over $[0, \pi]$.

Remark 14. Suppose that the matrix $Q(x)$ is a trigonometric polynomial, for instance

$$Q(x) = \begin{pmatrix} \cos x & \sin x \\ -\sin x & \cos x \end{pmatrix}, \quad Q(x) = \begin{pmatrix} -\cos(2x) & -\sin(2x) \\ -\sin(2x) & \cos(2x) \end{pmatrix}, \quad Q(x) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad \dots$$

In this case, the function $g(x)$ defined in (2.45) is a trigonometric polynomial as well. Consequently, the preconditioner P_n is banded and a linear system with matrix P_n is easily solvable. More precisely, P_n is a block-banded matrix, and for such matrices there exist versions of the Gaussian elimination with linear cost with respect to n.

Remark 15. In this example, the problem of minimizing the quantity $\|\rho(\phi_\alpha(f,g))\|_\infty$ over the set $\mathcal{D} = \mathcal{D}' \times (0,\infty)$ was equivalent to the problem of minimizing the best asymptotic upper bound σ^* in (2.49) over $\mathcal{D}'$. The reason is that the minimum of $\|\rho(\phi_\alpha(f,g))\|_\infty$ over all $\alpha > 0$, obtained in (2.54) for $\alpha = \alpha^*$, turned out to be σ^* independently of (a,b); the subsequent minimization of $\|\rho(\phi_{\alpha^*}(f,g))\|_\infty$ over $\mathcal{D}'$ coincided with the minimization of σ^* over the same set. In general, however, it may happen that the minimization of $\|\rho(\phi_\alpha(f,g))\|_\infty$ is not at all equivalent to the minimization of the best asymptotic upper bound σ^*, because the inequality (2.37) might be strict. We will see an example on this subject in Case 9.

We also observe that, since

$$\sigma^* = \frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1}$$

is an increasing function of

$$\kappa = \frac{\max(1,\eta)}{\min(1,\xi)},$$

the minimization of σ^* over $\mathcal{D}'$ is equivalent to the minimization of κ over the same set. The couple $(\hat{a}, \hat{b})$ in (2.56) is then the minimizer of both κ and σ^* over $\mathcal{D}'$. Recalling the definitions of ξ and η in (2.46), we note that κ is a sort of measure of the relative approximation between $|x| + 1$ and $a + b\cos x$. The minimization of κ (and σ^*) is therefore equivalent to finding the trigonometric polynomial of the form $a + b\cos x$ that best approximates the function $|x|+1$ according to the measure κ. In Figure 2.16 we report the graphs of $|x| + 1$ and $\hat{a} + \hat{b}\cos x$, from which we see that $\hat{a} + \hat{b}\cos x$ is a quite good approximation of $|x|+1$. Another reasonable approximation of $|x|+1$ is given by the polynomial $\bar{a}+\bar{b}\cos x$ that interpolates $|x| + 1$ at the endpoints $x = 0$ and $x = \pi$. In this case we have $(\bar{a}, \bar{b}) = (1 + \frac{\pi}{2}, -\frac{\pi}{2}) \approx (2.57, -1.57)$ (note that $(\bar{a}, \bar{b})$ is close to $(\hat{a}, \hat{b}) = (2.6, -1.5)$) and $\sigma^*(\bar{a}, \bar{b}) \approx 0.080$ is close to $\sigma^*(\hat{a}, \hat{b}) \approx 0.064$.

n	HSS ($\alpha = \sqrt{\pi+1}$)	PHSS ($g = \tilde{g},\ \alpha = \tilde{\alpha} = 1$)	PHSS ($g = \tilde{g},\ \alpha = \tilde{\alpha}^* \approx 1.272$)
50	0.3395	0.0030	0.2232
100	0.3403	0.0030	0.2242
200	0.3407	0.0030	0.2247
400	0.3409	0.0030	0.2250

Table 2.10: Spectral radius of the HSS and PHSS iteration matrices fixed $Q(x) = I_2$.

n	HSS ($\alpha = \sqrt{\pi+1}$)	PHSS ($g = \tilde{g},\ \alpha = \tilde{\alpha} = 1$)
50	14	5
100	13	4
200	13	4
400	13	4

Table 2.11: Number of HSS and PHSS iterations for $T_n(f)$ fixed $Q(x) = I_2$.

Case 9. As pointed out in Remark 15, in general the minimization of $\|\rho(\phi_\alpha(f,g))\|_\infty$ when (g,α) varies in some set $\mathcal{D}' \times (0,\infty)$ is not equivalent to the minimization of $\sigma^* := \sigma^*(g)$ when g varies in $\mathcal{D}'$. This is true because $\min_{\alpha>0} \|\rho(\phi_\alpha(f,g))\|_\infty$ may be strictly less than $\sigma^*(g)$ for some $g \in \mathcal{D}'$. Actually, it may also happen that

$$\min_{\alpha>0} \|\rho(\phi_\alpha(f,g))\|_\infty \ll \sigma^*(g).$$

In this example, we will see that, after choosing a discrete set $\mathcal{I} \subset (0,\infty)$ in which varying the parameter α, we have

$$\min_{(g,\alpha)\in\mathcal{D}'\times\mathcal{I}} \|\rho(\phi_\alpha(f,g))\|_\infty \approx 0.003 \ll \min_{g\in\mathcal{D}'} \sigma^*(g) \approx 0.064.$$

In addition, the function $\tilde{g}$ for which the left minimum is attained is different from the function $\hat{g}$ for which the right minimum is attained, and the parameter $\tilde{\alpha}$ yielding the left minimum together with $\tilde{g}$ is different from $\tilde{\alpha}^* := \alpha^*(\tilde{g})$. Note that, despite the fact that both 0.003 and 0.064 are small, their ratio 0.064/0.003 is about 21, meaning that $\min_{g\in\mathcal{D}'} \sigma^*(g)$ is relatively much larger than $\min_{(g,\alpha)\in\mathcal{D}'\times\mathcal{I}} \|\rho(\phi_\alpha(f,g))\|_\infty$.

Let

$$\begin{aligned} f(x) &:= Q(x) \begin{pmatrix} 1+0.64\,\mathbf{i} & -1.5\,\mathbf{i} \\ -1.5\,\mathbf{i} & |x|+1+\mathbf{i} \end{pmatrix} Q^*(x), \\ g(x) &:= Q(x) \begin{pmatrix} 1 & 0 \\ 0 & a+b\cos x \end{pmatrix} Q^*(x), \end{aligned} \tag{2.59}$$

where a, b are real parameters such that $a > |b| \geq 0$. We note that $g(x)$ and

$$\mathrm{Re}(f(x)) = Q(x) \begin{pmatrix} 1 & 0 \\ 0 & |x|+1 \end{pmatrix} Q^*(x)$$

are exactly the same as in Case 8. Therefore, setting $A_n := T_n(f)$ and $P_n := T_n(g)$, the facts i–iii hold *unchanged*. In particular, the value σ^* is still given by (2.49), implying that the minimization of $\sigma^* = \sigma^*(a,b)$ over the same set $\mathcal{D}'$ considered in Case 8 yields again the result that we have seen in (2.55):

$$\min_{(a,b)\in\mathcal{D}'} \sigma^*(a,b) = \sigma^*(\hat{a},\hat{b}) =: \hat{\sigma}^* \approx 0.064,$$

where $(\hat{a},\hat{b}) := (2.6,-1.5)$ as in (2.56). On the other hand, the minimization of $\|\rho(\phi_\alpha(f,g))\|_\infty$ over $\mathcal{D} = \mathcal{D}' \times \mathcal{I},\ \mathcal{I} := \{\alpha \in (0,\infty) :\ \alpha = 0.1,\ 0.2,\ 0.3,\ \ldots,\ 1.2\}$, yields

$$\min_{(a,b,\alpha)\in\mathcal{D}\times\mathcal{I}} \|\rho(\phi_\alpha(f,g))\|_\infty = \|\rho(\phi_\alpha(f,g))\|_\infty|_{(a,b,\alpha)=(\tilde{a},\tilde{b},\tilde{\alpha})} = \|\rho(\phi_{\tilde{\alpha}}(f,\tilde{g}))\|_\infty \approx 0.003,$$

where $(\tilde{a},\tilde{b},\tilde{\alpha}) := (1.6,0,1)$ and $\tilde{g} := g_{\tilde{a},\tilde{b}}$ is the function g obtained from (2.59) for $(a,b) = (\tilde{a},\tilde{b})$.

In the following numerical experiments we fixed $Q(x) = I_2$. Table 2.10 shows that the spectral radius of the PHSS iteration matrix $M_n(\tilde{\alpha})$, corresponding to the optimal choice $(a,b,\alpha) = (\tilde{a},\tilde{b},\tilde{\alpha})$, converges to $\|\rho(\phi_{\tilde{\alpha}}(f,\tilde{g}))\|_\infty \approx 0.003$ as $n \to \infty$. This is in accordance with the theoretical prediction

in Remark 12 if $\{M_n(\tilde{\alpha})\}_n \sim_\lambda (\phi_{\tilde{\alpha}}(f,\tilde{g}), I_1)$, so we are led to believe that this eigenvalue distribution relation really holds. From Tables 2.10–2.11, we see that the PHSS obtained with $(a,b,\alpha) = (\tilde{a},\tilde{b},\tilde{\alpha})$ is much faster than the HSS with the best asymptotic parameter $\sqrt{m_{\mathrm{Re}(f)}M_{\mathrm{Re}(f)}} = \sqrt{\pi+1}$ suggested by Theorem 18.

Remark 16. Fix $g = \tilde{g}$ and let $M_n(\alpha)$ be the corresponding PHSS iteration matrix. Table 2.10 (last two columns) shows that, for $\alpha = \tilde{\alpha} = 1$ and $\alpha = \tilde{\alpha}^* \approx 1.272$, we have

$$\lim_{n\to\infty} \rho(M_n(\alpha)) = \|\rho(\phi_\alpha(f,\tilde{g}))\|_\infty \approx \begin{cases} 0.003 & \text{if } \alpha = \tilde{\alpha}, \\ 0.225 & \text{if } \alpha = \tilde{\alpha}^*. \end{cases}$$

On the other hand, the best upper bound $\sigma_n(\tilde{\alpha}_n^*)$ for $\rho(M_n(\alpha))$ converges to $\tilde{\sigma}^* := \sigma^*(\tilde{g}) \approx 0.34$, and $\tilde{\alpha}_n^* \to \tilde{\alpha}^* \approx 1.272$. Therefore, for large n,

- the best upper bound $\sigma_n(\tilde{\alpha}_n^*) \approx 0.34$ is much larger than $\rho(M_n(\tilde{\alpha})) \approx 0.003$, implying that $\sigma_n(\tilde{\alpha}) \geq 0.34$ is a terribly crude estimate for $\rho(M_n(\tilde{\alpha}))$;

- the best parameter $\tilde{\alpha}_n^*$ that minimizes $\sigma_n(\alpha)$ is approximately equal to $\tilde{\alpha}^* \approx 1.272$. Hence, $\tilde{\alpha}_n^*$ is rather far from the actual minimizer $\tilde{\alpha} = 1$ of $\rho(M_n(\alpha))$ over $\mathcal{I}$, and it is also far from minimizing $\rho(M_n(\alpha))$, because $\rho(M_n(\tilde{\alpha}_n^*)) \approx 0.225 \gg 0.003 \approx \rho(M_n(\tilde{\alpha}))$.

This remark shows that, in general, there are situations in which $\sigma_n(\alpha)$ is not at all a good estimate for $\rho(M_n(\alpha))$ and, moreover, the best parameter α_n^* that minimizes $\sigma_n(\alpha)$ is far from minimizing $\rho(M_n(\alpha))$. In these situations, the effects of the imaginary parts in the PHSS iteration matrix $M_n(\alpha)$ are significant.

Bivariate examples

Here we show the practical effectiveness of the proposed approach in a 2-level setting. The results are of interest because of the known theoretical barriers mentioned is Section 1.6 concerning multilevel structures, for which it is not possible, in general, to find superlinear/optimal preconditioners chosen in matrix algebras.

Case 10. Let

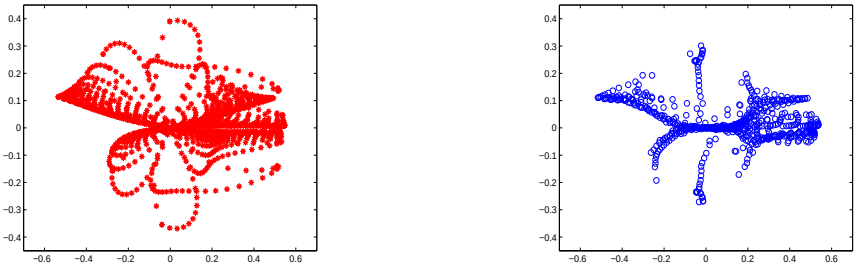
$$\begin{aligned} f(x_1,x_2) &:= \frac{1}{x_1^2+x_2^2+1} + 3\mathrm{i}\cos(x_1+x_2), \\ g(x_1,x_2) &:= a + b\cos(x_1) + c\cos(x_2), \end{aligned}$$

where a,b,c are real parameters such that $a > |b| + |c|$. Under this condition $T_n(g)$ is HPD. Since $\mathrm{Re}(f(x_1,x_2)) = \frac{1}{x_1^2+x_2^2+1}$ is positive for every $(x_1,x_2) \in I_2 = (-\pi,\pi)^2$, the matrix $\mathrm{Re}(T_n(f)) = T_n(\mathrm{Re}(f))$ is HPD as well. Therefore, we can apply the PHSS method with preconditioner $T_n(g)$ for solving linear systems with coefficient matrix $T_n(f)$.

In the following numerical experiments, we fixed a,b,c such that $g(x_1,x_2)$ interpolates $\mathrm{Re}(f(x_1,x_2))$ in $(0,0),(\pi,\pi),(\pi,0)$. An easy computation gives $(a,b,c) \approx (0.524, 0.454, 0.022)$ and it can be checked that these values of a,b,c satisfy the condition $a > |b| + |c|$. As for the choice of α, we fixed the best asymptotic parameter $\alpha = \alpha^* = \sqrt{m_h M_h} \approx 0.329$ suggested by Theorem 18, where

$$h(x_1,x_2) := g(x_1,x_2)^{-1}\mathrm{Re}(f(x_1,x_2)) \approx \frac{1}{(x_1^2+x_2^2+1)(0.524+0.454\cos(x_1)+0.022\cos(x_2))}.$$

In Figure 2.17(a) we report the essential range of $\phi_{\alpha^*}(f,g)$ on I_2, while Figure 2.17(b) refers to the eigenvalues of the PHSS iteration matrix $M_n(\alpha^*)$ with $n = (25,25)$. A comparison between these two figures shows that the eigenvalues of $M_n(\alpha^*)$ are clustered at the essential range of $\phi_{\alpha^*}(f,g)$.



(a) (b)

Figure 2.17: (a) Essential range of $\phi_{\alpha^*}(f,g)$; (b) Eigenvalues in the complex plane of $M_n(\alpha^*)$ with $n=(25,25)$.

Furthermore, as shown in Table 2.12, the spectral radius of the PHSS iteration matrix $M_n(\alpha^*)$ converges to $\|\rho(\phi_{\alpha^*}(f,g))\|_\infty \approx 0.547$, as $n \to \infty$. Hence we might guess that the distribution relation $\{M_n(\alpha^*)\}_{n\in\mathbb{N}^2} \sim_\lambda (\phi_{\alpha^*}(f,g), I_2)$ holds even for this 2-level example.

$n = (n,n)$	$\rho(M_n(\alpha^*))$
(10,10)	0.4947
(15,15)	0.5149
(20,20)	0.5275
(25,25)	0.5352
	$\|\rho(\phi_{\alpha^*}(f,g))\|_\infty \approx 0.547$

Table 2.12: Spectral radius of the PHSS iteration matrix fixed $\alpha = \alpha^* \approx 0.329$.

$n = (n,n)$	HSS ($\alpha = \sqrt{(2\pi^2+1)^{-1}}$)	PHSS ($\alpha = 0.329 \approx \alpha^*$)
(10,10)	25	14
(15,15)	25	15
(20,20)	25	14
(25,25)	24	14

Table 2.13: Number of HSS and PHSS iterations for $T_n(f)$.

Table 2.13 reports the number of HSS and PHSS iterations. We see from the table that the PHSS with parameter $\alpha = \alpha^*$ is faster than the HSS with the best asymptotic parameter $\alpha = \sqrt{m_{\mathrm{Re}(f)} M_{\mathrm{Re}(f)}}$ suggested by Theorem 18.

In the final example, we consider a problem coming from a differential context, whose approximation leads to multilevel Toeplitz structures with weakly sectorial multivariate symbols. We show that also in this setting the performances of the PHSS technique are really very good.

More in detail let us take into consideration the convection diffusion equation

$$\begin{cases} -\Delta u + \beta \cdot \nabla u + \gamma u = g(x), & \text{on } \Omega = [0,1]^d, \\ \\ \text{Dirichlet BCs on } \partial\Omega, \end{cases} \tag{2.60}$$

where $\beta := (\beta_1, \ldots, \beta_d)$, with $\beta_1, \ldots, \beta_d, \gamma$ constants, $\gamma \geq 0$. By resorting to second-order centered finite differences, we find linear systems with coefficient matrices $A_n = T_n(f_n)$, $n := (n, \ldots, n) \in \mathbb{N}^d$: here the sequence of symbols $\{f_n\}_n$, depending on the step-size $h = \frac{1}{n+1}$ and therefore on the matrix size, is such that

$$f_n(x) = \sum_{j=1}^{d} (2 - 2\cos(x_j) - \mathbf{i}\, h\beta_j \sin(x_j) + h^2\gamma), \quad x = (x_1, \ldots, x_d) \in I_d = (-\pi, \pi)^d.$$

In the case where there exists an index j such that $h\beta_j$ is not small, the symbol has a non-trivial imaginary part, but still its real part $\sum_{j=1}^{d}(2 - 2\cos(x_j)) + h^2\gamma$ is nonnegative and not identically zero.

Therefore (see Proposition 1) we are in the framework of the present section with $k = d$ and $s = 1$. This situation is typical of singularly perturbed problems, in which the convection term is not negligible or even dominating.

	$n = 10$		$n = 30$		$n = 50$	
β	HSS	PHSS	HSS	PHSS	HSS	PHSS
0.1	36	2	173	2	370	2
1	36	2	173	2	370	2
5	34	2	173	2	370	2
10	30	2	171	2	370	2
100	27	2	94	2	292	2
1000	31	2	28	2	38	2

Table 2.14: Number of HSS and PHSS iterations for $T_n(f_n)$ fixed $n = 10, 30, 50$.

Case 11. As a specific numerical test we consider problem (2.60) with $d = 2$ and $\beta := (\beta, \beta)$, where $\beta, \gamma \in \mathbb{R}$ and $\gamma \geq 0$. According to the above discussion, the resulting coefficient matrices $T_n(f_n)$, $n = (n, n)$ are 2-level Toeplitz matrices and, more specifically, we have

$$T_n(f_n) = \begin{bmatrix} T_0 & T_{-1} & & & \\ T_1 & T_0 & T_{-1} & & \\ & \ddots & \ddots & \ddots & \\ & & T_1 & T_0 & T_{-1} \\ & & & T_1 & T_0 \end{bmatrix} \in \mathcal{M}_{n^2}$$

where

$$T_0 = \begin{bmatrix} 4+\gamma h^2 & -1+h\beta & & \\ -1-h\beta & \ddots & \ddots & \\ & \ddots & \ddots & -1+h\beta \\ & & -1-h\beta & 4+\gamma h^2 \end{bmatrix} \in \mathcal{M}_n,$$

$$T_{-1} = \begin{bmatrix} -1+h\beta & & \\ & \ddots & \\ & & -1+h\beta \end{bmatrix} \in \mathcal{M}_n, \quad T_1 = \begin{bmatrix} -1-h\beta & & \\ & \ddots & \\ & & -1-h\beta \end{bmatrix} \in \mathcal{M}_n,$$

and $h = \frac{1}{n+1}$. As already observed, $\mathrm{Re}(f_n)$ is nonnegative and not identically zero, then from Proposition 1 it follows that $\mathrm{Re}(T_n(f_n))$ is positive definite.

In the numerical test we fixed $\gamma = 1$ and chose the preconditioner as $P_n = \mathrm{Re}(T_n(f_n))$. It is clear that in this case the best parameter suggested by Theorem 10 is $\alpha = 1$. Table 2.14 refers to the number of HSS and PHSS iterations varying β in $\{0.1, 1, 5, 10, 100, 1000\}$ and fixed $n = 10, 30, 50$, respectively. Note that the value $\beta = 1000$ gives rise to a convenction-dominated problem. While the proposed PHSS is a robust method with number of iterations equal to 2, independent of β and n, the HSS is not optimal in the matrix size, independently of the choice of the convection parameter β.

We observe that the first line of the iteration (1.27) is a trivial equation since the coefficient matrix is a multiple of the identity. Concerning the second line in (1.27), the equation $\left(\alpha I + P_n^{-1}\mathbf{i}\,\mathrm{Im}(A_n)\right) x^{(k+1)} = \left(\alpha I - P_n^{-1}\mathrm{Re}(A_n)\right) x^{(k+\frac{1}{2})} + P_n^{-1} b$ is solved inexactly by using few GM RES steps. As proved in [17, Theorem 3.2], the eigenvalues of $P_n^{-1}\,\mathrm{Im}(A_n)$ are strongly clustered at 0 and contained in a proper interval $[-\delta, \delta]$, with $\delta > 0$ independent of h, and so the resulting GMRES is very fast.

Chapter 3

Preconditioning techniques in image deblurring

In this chapter we focus on two regularization techniques for structured matrices arising in the context of the image deblurring problem which is briefly illustrated in Section 3.1. Before to describe them in more detail, in Section 3.2 we recall what a regularization technique is. The first technique that we propose is a structure preserving preconditioning aimed to accelerate the convergence of iterative regularization methods without spoiling the restoration (Section 3.3). The second one consists in a diagonal regularization preconditioner containing the Fourier coefficients of the observed image or of an approximation of the true image aimed to provide sparse reconstructions. We will interpret it as a regularization matrix for the Tikhonov method in the Fourier domain whose penalty term approximates the 1-norm. The advantage is that the resulting linear system is diagonal and the regularization parameter can be easily estimated (Section 3.4).

3.1 Problem setting: image deblurring

Let us consider the linear system

$$Ax = b, \tag{3.1}$$

where $A \in \mathbb{R}^{n\times n}$ and $x, b \in \mathbb{R}^n$. In the image deblurring context, A is the matrix associated to the blur operator, $x \in \mathbb{R}^n$ is an approximation of the true image $\overline{x} \in \mathbb{R}^n$ of an unknown object (more precisely, x is a stack-ordered vector of an images with n pixels) and $b \in \mathbb{R}^n$ is the detected image affected by blur and corrupted by a noise $\eta \in \mathbb{R}^n$, that is, $b = A\overline{x} + \eta$. Image deblurring consists in computing an approximation of the true image $\overline{x}$ by means of an appropriate solution of (3.1).

3.1.1 Structure of the blurring matrix

The structure of the matrix A depends both on the Point Spread Function (PSF), and the strategy adopted to deal with boundary artifacts [85, 109, 132, 115]. We assume the blurring model to be space-invariant, which means that the same blur will occur on all over the image domain. For the sake of notational simplicity, we consider a square PSF $H \in \mathbb{R}^{n\times n}$. We suppose that the position of the PSF centre is known. Thus, H can be depicted in this way

$$H = \begin{pmatrix} h_{-m_1,-m_2} & \cdots & & h_{-m_1,0} & & \cdots & h_{-m_1,n_2} \\ \vdots & \ddots & & \vdots & & & \vdots \\ & & h_{-1,-1} & h_{-1,0} & h_{-1,1} & & \\ h_{0,-m_2} & \cdots & h_{0,-1} & h_{0,0} & h_{0,1} & \cdots & h_{0,n_2} \\ & & h_{1,-1} & h_{1,0} & h_{1,1} & & \\ \vdots & & & \vdots & & \ddots & \vdots \\ h_{n_1,-m_2} & \cdots & & h_{n_1,0} & & \cdots & h_{n_1,n_2} \end{pmatrix}_{n\times n},$$

where $h_{0,0}$ is the central coefficient and $m_1 + n_1 + 1 = m_2 + n_2 + 1 = n$.

Given the pixels h_{j_1,j_2} of the PSF, it is possible to associate the generating function $f : \mathbb{R}^2 \to \mathbb{C}$ as follows

$$f(x_1, x_2) = \sum_{j_1=-m_1}^{n_1} \sum_{j_2=-m_2}^{n_2} h_{j_1,j_2} \mathrm{e}^{\mathbf{i}(j_1 x_1 + j_2 x_2)} = \sum_{j_1,j_2=-n+1}^{n-1} h_{j_1,j_2} \mathrm{e}^{\mathbf{i}(j_1 x_1 + j_2 x_2)} \ , \tag{3.2}$$

with the assumption that $h_{j_1,j_2} = 0$ if the corresponding pixel is not detected, that is, if the element (h_{j_1,j_2}) does not belong to the matrix H [50]. Note that h_{j_1,j_2} are the Fourier coefficients of $f \in \Pi_{n-1}$, where $\Pi_k = \mathrm{span}\{\mathrm{e}^{\mathbf{i}(j_1 x_1 + j_2 x_2)}, j_1, j_2 = -k, \ldots, k\}$, so that the generating function f contains the same information of H.

Boundary conditions (BCs) try to capture and to include into the deblurring model the unknown behaviour of the signal outside the Field Of View (FOV) in which the detection is made [85]. Indeed, the information inside the FOV contained in the detected image b is generally not complete to restore the true image even in the (unrealistic) noiseless case. Among the BCs schemes, we consider the following zero Dirichlet, periodic, reflective (also called Neumann), and anti-reflective ones.

- *Zero Dirichlet BCs.* In this model the image outside the FOV is null, that is, zero pixel-valued. The blurring matrix A is a BTTB. The zero Dirichlet BCs can be useful for some applications in astronomy, where an empty dark space surrounds a well located object. On the other hand, it gives rise to high ringing effects close to the boundary of the restored image in other classical imaging applications. We highlight that we do not have fast trigonometric transforms for diagonalizing BTTB matrices. This represents an important drawback in using Zero BCs with filtering methods like classical Tikhonov.

- *Periodic BCs.* In this model the image inside the FOV is periodically repeated outside the FOV. So, considering for simplicity the 1D case, they impose that

  $$x_{1-j} = x_{n+1-j} \qquad \text{and} \qquad x_{n+j} = x_j, \qquad j = 1, \ldots, p,$$

  where p is the parameter related to number of pixels outside the FOV that are taken into account. For multidimensional problems it is enough to apply the same assumption in every direction. The 2D corresponding blurring matrix A is a BCCB. Periodic BCs are computational favourable since the matrix A can be easily diagonalized by DFT. Clearly, if we are not in case the image is periodic outside the FOV (and this happens quite often), we have again ringing effects in restoration.

- *Reflective BCs.* In this model the image inside the FOV is periodically reflected, as well as there were a vertical mirror, along each edge. That way, the pixel values across the boundary are extended so that the continuity of the image is preserved at the boundary. Formally, in the 1D case, these BCs impose that

  $$x_{1-j} = x_j \qquad \text{and} \qquad x_{n+j} = x_{n+1-j}, \qquad j = 1, \ldots, p.$$

  In the 2D case the same assumption on the image outside the FOV is done firstly in one direction and then in the other direction. The corresponding blurring matrix A is a Block Toeplitz plus Hankel with Toeplitz plus Hankel blocks, which can be diagonalized by Discrete Cosine Transform (DCT) when the PSF is symmetric [109].

- *Anti-Reflective BCs.* In this model the image inside the FOV is periodically anti-reflected, as well as there were two perpendicular mirrors, one horizontal and one vertical, along each edge. That way, the pixel values across the boundary are extended so that the continuity of the image and the continuity of the normal derivatives are both preserved at the boundary. Formally, in the 1D case, these BCs impose that

  $$x_{1-j} = 2x_1 - x_{j+1} \qquad \text{and} \qquad x_{n+j} = 2x_n - x_{n-j}, \qquad j = 1, \ldots, p.$$

  In the 2D case the matrix A is block Toeplitz-plus-Hankel with Toeplitz-plus-Hankel blocks plus a structured low-rank correction matrix, which can be diagonalized by Anti-Reflective Transform (ART) when the PSF is symmetric [132].

We notice that in all these four cases A has a Toeplitz structure, given by the shift–invariant structure of the continuous operator, plus a correction depending on the BCs. Even if, as said, in some cases the BCs suffer from the lack of fast trigonometric transforms, for all of them the matrix-vector multiplication can be always computed in an efficient way by exploiting the structure Toeplitz+correction. Indeed,

the multiplication can be made by means of FFTs (accessing only the PSF) on an appropriately padded image of larger size. Reflective BCs and Anti-Reflective BCs are even cheaper than the others in some cases since they require only real operations instead of complex ones [3]. For a detailed discussion on BCs and the associated blurring matrices referee to [85, 64].

3.2 Ill-posed problems and regularization methods

It is well-known that the image deblurring is an ill-posed problem and then that the linear system (3.1) is very ill-conditioned. As a consequence, the direct solution of (3.1) is dominated by the noise and a regularization technique has to be applied. Before to explain what we mean with regularization methods, let us start with the definition of well- and ill-posed problems (see [68]).

Definition 16 (Hadamard)**.** A problem is *well-posed* if

(1) it admits solution,

(2) the solution is unique,

(3) the solution depends continuously on the data.

If at least one of Hadamard's conditions is not satisfied then the problem is *ill-posed*. The violation of (1) and (2) is usually less damaging than the violation of (3). Indeed, since in the applications we work with perturbed data, we can relax the notion of solution and, if the solution is not unique, we can impose additional constraints, e.g. choosing the solution of minimal norm. On the contrary, violation of (3) creates considerable numerical problems because the discretization of a continuous ill-posed problem is very ill-conditioned on a large subspace and then the numerical methods used for well-posed problems become unstable. Since we cannot make stable an inherently unstable problem, all that one has expect to do is to recover partial information about the solution as stably as possible. In this direction, there are the so-called *regularization methods*. Several regularization techniques have been proposed in literature like Tikhonov's method and its generalizations, Truncated Singular Values Decomposition (TSVD) [13], iterative regularization methods (Landweber, conjugate gradient, etc.) [18, 68].

The image deblurring is an ill-posed problem since it does not fulfill Hadamards conditions of well-posedness in Definition 16. As already mentioned, a very common approach to guarantee that the solution of (3.1) exists and that it is unique, is to compute a weak solution in the sense of the least-squares formulation (the solution of minimum norm), where the restored image is obtained by solving

$$x_{\mathrm{LS}} = \arg\min_{x \in S} \|x\|_2^2, \qquad S = \arg\min_{z \in \mathbb{R}^n} \|Az - b\|_2. \tag{3.3}$$

This remedy to the violation of (1) and (2) in Definition 16. However, in the continuous image deblurring problem point (3) in Definition 16 is not fulfilled, therefore we have to resort to a regularization strategy. This difficulty can be easily observed in a finite dimensional context as well by comparing the solution of (3.3) and (3.1). Assuming that A has full rank, the solution of (3.3) and that of (3.1) coincide. More precisely, we obtain $x_{\mathrm{LS}} = \bar{x} - A^{-1}\eta$. Therefore, the nonzero components (even only quasi-negligible) of η in the subspace generated by the eigenvectors corresponding to small eigenvalues of A will be dramatically amplified. Since

1. the small eigenvalues of A quickly tend to zero as the size of A tends to infinity,

2. the subspace associated to these degenerating eigenvalues has large dimension (usually more than one half of the global size of A),

if a regularization strategy is not employed, a small perturbation of the observed image can lead to a substantial, not admissible alteration in the deblurred image. In the more general case, considering the least-squares solution of (3.1), we have to use the Moore-Penrose pseudo-inverse of A and the singular value decomposition instead of the spectral decomposition, but the same considerations hold unchanged.

Because of the large dimensions of the linear systems involved, iterative methods are typically used to compute the approximation of $\bar{x}$ since they require only matrix-vector products. In Subsection 3.2.2 we introduce iterative regularization methods like Landweber, CG, etc. Since, their convergence can be slow, it is important to define strategies to accelerate the approximation process without losing the quality of the restored image. With this aim, in Subsection 3.2.3 we discuss the regularization preconditioning. A part from iterative regularization techniques, we consider also the Tikhonov regularization, cf. Subsection 3.2.1.

3.2.1 Tikhonov method

First of all, let us recall that if we deal with the reconstruction of images with $n_1 \times n_2$ pixels, the unknown x in (3.1) is a stack-ordered vector of length $n = n_1 n_2$. Since we will make extensive use of superscripts and subscripts, throughout this chapter, we denote the diagonal matrix of the eigenvalues of $M \in \mathbb{R}^{n\times n}$ by Λ_M instead of $\Lambda(M)$.

The widely used Tikhonov method is given by

$$\min_{x\in\mathbb{R}^n} \left\{ \|Ax - b\|_2^2 + \alpha \mathcal{R}(x) \right\}, \tag{3.4}$$

where $\alpha > 0$ is called *regularization parameter*, $\|Ax - b\|_2^2$ is the *data fitting term*, while $\mathcal{R}(x)$ is the *penalty term*. The term $\mathcal{R}(x)$ should nearly disappear for x close to the true image, and it should penalize only noise components. Furthermore, $\mathcal{R}(x)$ can take into account known properties of the true solution like continuity or sparsity.
A common choice is

$$\mathcal{R}(x) = \|Lx\|_2^2, \tag{3.5}$$

where L guarantees that $\mathcal{N}(A) \bigcap \mathcal{N}(L) = \{0\}$. The method (3.4)-(3.5) is called *generalized Tikhonov method* and is equivalent to the following linear system

$$(A^*A + \alpha L^*L)x = A^*b. \tag{3.6}$$

As observed in Section 3.1, when imposing periodic boundary conditions the matrix A in (3.1) is a BCCB matrix and then can be spectrally decomposed by DFT as follows

$$A = F^*\Lambda_A F, \tag{3.7}$$

where F is given in (1.22).

If even L^*L in (3.6) can be diagonalized through F, that is

$$L^*L = F^*\Lambda_{L^*L}F, \tag{3.8}$$

the generalized Tikhonov method becomes

$$\min_{x\in\mathbb{R}^n} \left\{ \|Ax - b\|_2^2 + \alpha\|Lx\|_2^2 \right\} \iff F^*(\Lambda_A^*\Lambda_A + \alpha\Lambda_{L^*L})Fx = F^*\Lambda_A^*Fb\,,$$

and hence

$$x = F^*(\Lambda_A^*\Lambda_A + \alpha\Lambda_{L^*L})^{-1}\Lambda_A^*Fb.$$

Choice of L. The matrix L can be chosen as the identity matrix or derived by discretizations of some derivative operators (with appropriate boundary conditions [61]). In these cases L does not depend on A. A possible alternative is to define L via A, e.g. in the form

$$L^*L = \left(I - \frac{A^*A}{\rho(A)^2} \right)^p, \tag{3.9}$$

with some exponent p, e.g. $p = 1$ (see [90, 92]).

Choice of α. Finding a good choice for the regularization parameter α is a difficult task. Actually, we cannot determine the optimal α, but only give an estimation of it. Many techniques have been proposed in literature aimed to estimate the regularization parameter α. A largely used method is the Generalized Cross Validation (GCV), see [84]. For the generalized Tikhonov method, the GCV determines the value of α that minimizes the GCV functional

$$G_{\text{Tik}}(\alpha) = \frac{\|(I - AA_{\text{reg}}^\dagger)b\|_2^2}{(\text{tr}(I - AA_{\text{reg}}^\dagger))^2}, \tag{3.10}$$

where $A_{\text{reg}}^\dagger = (A^*A + \alpha L^*L)^{-1}A^*$ is the Tikhonov regularized pseudoinverse. When A and L^*L can be both diagonalized through F the GCV functional becomes

$$G_{\text{Tik}}(\alpha) = \frac{\|(I - \Lambda_A(\Lambda_A^*\Lambda_A + \alpha\Lambda_{L^*L})^{-1}\Lambda_A^*)\widehat{b}\|_2^2}{(\text{tr}(I - \Lambda_A(\Lambda_A^*\Lambda_A + \alpha\Lambda_{L^*L})^{-1}\Lambda_A^*))^2} = \frac{\sum_{i=1}^n (\mu_i\widehat{b}_i)^2}{(\sum_{i=1}^n \mu_i)^2}, \tag{3.11}$$

where $\mu_i = \lambda_i(L^*L)/(|\lambda_i(A)|^2 + \alpha\lambda_i(L^*L))$ and $\widehat{b} = Fb$, cf. [85]. Therefore, the evaluation of G_{Tik} at a point α requires only $O(n)$ operations assuming that the eigenvalues of A and of L^*L have been already computed.

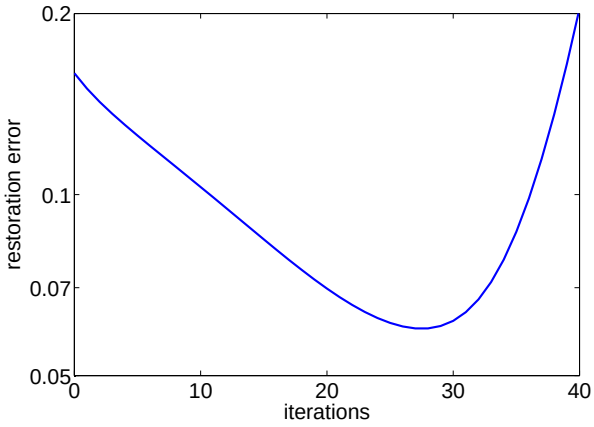


Figure 3.1: Semi-convergence: RREs vs number of iterations

3.2.2 Iterative regularization methods

The solution x_{LS} of the least-square problem (3.3) is such that $\nabla\phi(x_{\mathrm{LS}}) = 0$, where $\phi(z) = \|Az - b\|_2^2$. Moreover, $\nabla\phi(z) = 2A^*Az - 2A^*b$ and hence x_{LS} is solution of the normal equations

$$A^*Ax = A^*b. \tag{3.12}$$

The classical iterative regularization algorithm for least-squares problems is the Landweber method. This is the simplest gradient descent algorithm for solving problem (3.3). Let x_k be the approximate solution computed at the k-th iteration, then the Landweber method is defined by

$$x_{k+1} = x_k + \tau A^*(b - Ax_k),$$

where x_0 is given and τ is a relaxation parameter satisfying $0 < \tau < 2/\|A^*A\|$ for some induced norm $\|\cdot\|$. With this choice of τ the method is convergent and it satisfies the classical semi-convergence property (see [68]). A semi-convergent method starts reconstructing the low-frequency components of the solution; then, as the iteration progresses, the high-frequency components are reconstructed together with the noise components (see [18]). Hence the method must be stopped before it starts to reconstruct the noise. In Figure 3.1 we plot a typical example of the behavior of the Relative Reconstruction Error (RRE) defined as

$$\mathrm{RRE} := \frac{\|x_k - \bar{x}\|_2}{\|\bar{x}\|_2} \tag{3.13}$$

vs the number of iterations for an iterative regularization method (not necessarily the Landweber method). As we can see from Figure 3.1, the RRE reaches a minimum after a suitable number of iterations. Therefore, an early stopping is required. A classical criterium is the *discrepancy principle* (see [68]) which proceeds as follows: stop at the first iteration m that satisfies the condition

$$\|Ax_m - b\|_2 < \gamma\|\eta\|_2, \tag{3.14}$$

where x_m is the approximation provided by the method at the m-th iteration and $\gamma \geq 1$.

In practical applications, a frequent choice for the initial guess is $x_0 = 0$ because, in the case of non-uniqueness, this choice provides the minimal norm solution.

Other gradient descent algorithms for solving problem (3.3) satisfy the semi-convergence property and can be effectively used (see [80]). The most popular is probably the CG method applied to the linear system (3.12), known as CGLS. It is again an iterative regularizing method and usually it converges faster than the Landweber method.

3.2.3 Regularization preconditioners

Due to ill-conditioning of the image deblurring problem, the number of iterations required by a CG-like method to obtain a satisfactory result can be large and a preconditioning technique is required to increase the rate of convergence. In this context, we speak of *regularization preconditioning*. Such kind of preconditioning arises from the fact that for ill-conditioned linear systems related to ill-posed

problems, classical preconditioning may leads to wrong results. Indeed, if the preconditioner is a too close approximation of A, then it strongly "inherits" the ill-conditioning of A. In this case, the first iterations of a preconditioned CG-like method are already highly influenced by the noise of the input data, and the preconditioner gives rise to high instability, mixing up the subspace where the components of the noise are negligible with respect to the components of the signal (signal subspace), with the subspace where the components of the noise are not negligible (noise subspace). In order to avoid instability, the preconditioner should speed up the convergence only in the signal subspace. In other words, a regularizing preconditioner has to be able to approximate A in the signal subspace and filters the noisy components, simultaneously. It is clear that, one has the non-trivial task of choosing a regularization parameter, which distinguishes noise subspace from signal subspace.

Note that if the blurring matrix A is not positive (semi-)definite, to deal with CG-like methods instead of system (3.1), we can solve the system of the normal equations (3.12). In doing this, all iterative methods become more stable, i.e. less sensitive with respect to data noise, but their rate of convergence slows down, then a regularization preconditioning is needed as well.

In the past twenty years, regularization circulant preconditioners have been extensively investigated for accelerating the convergence of iterative methods without spoiling the restoration. The pioneering work is due to Hanke, Nagy, Plemmons [82] and deals with a BCCB regularization preconditioner for CGLS that preconditions only in the signal subspace, without acting in the noise subspace.

As regards BTTB regularizing preconditioners for BTTB matrices, in [81] Hanke and Nagy extended the idea in [36] to ill-conditioned matrices by recurring to a preconditioner generated by a regularized inverse of the symbol, that is by inverting the symbol only in the signal subspace. This strategy was devoted to symmetric PSF, and preconditioned MR-II (a variant of the minimal residual method sometime also called conjugate residual). The preconditioner was built by the circulant padding of Toeplitz matrices mentioned in Section 1.6.

At the best of our knowledge, the only other two papers in which the structure of A is preserved in the preconditioner are [109, 47] for Reflective and Anti-Reflective BCs, respectively. As in [81], even in these two cases, the obtained results heavily depend on the symmetry properties of the PSF. In fact, both [109, 47] show that the optimal (in the sense of Frobenius norm) preconditioner in a proper algebra diagonalized by a fixed real transform is associated with the symmetrized version of the original PSF. Such preconditioning technique works well when the PSF is close to be symmetric, while has poor performance for strongly nonsymmetric PSFs.

In Section 3.3 we propose a general preconditioning technique which can be used for any type of PSF and BCs.

Remark 17. When the regularization matrix L is invertible, the generalized Tikhonov method (3.4)-(3.5) can be interpreted as a regularizing preconditioner. Indeed, starting from Tikhonov regularization (3.4)-(3.5) and imposing $y = Lx$, we obtain

$$\min_{y=Lx\in\mathbb{R}^n} \left\{ \|AL^{-1}y - b\|_2^2 + \alpha\|y\|_2^2 \right\}, \tag{3.15}$$

which is the Tikhonov method for the right preconditioned linear system $AL^{-1}y = b$. The analogous least square problem of (3.15) is

$$\min_{y=Lx\in\mathbb{R}^n} \|AL^{-1}y - b\|_2^2,$$

which corresponds to the following split preconditioning for the normal equations (3.12)

$$L^{-*}A^*AL^{-1}y = L^{-*}A^*b.$$

Previous linear system can be solved by means of iterative methods like Preconditioned Conjugate Gradient for Least Squares (PCGLS). After having computed y, x is given by $x = L^{-1}y$.

In Section 3.4 we introduce a regularization matrix for the Tikhonov method in the Fourier domain that can be interpreted as a diagonal regularization preconditioner containing the Fourier coefficients of the observed image or of an approximation of the true image aimed to provide sparse reconstructions.

3.3 Structure preserving reblurring preconditioning

In the context of the circulant regularization preconditioning, in [45], it has been proposed a variant of the normal equations, known as reblurring preconditioning, which replaces the conjugate transpose A^* of the system matrix A with a new circulant matrix. Nevertheless, as already observed in Section

1.6, BCCB preconditioners cannot provide a strong clustering of the eigenvalues of BTTB-like matrices [136], while, on the other hand, it is crucial to preserve the structure of the coefficient matrix to have an effective preconditioner [121].

In this section, we propose an extension of the reblurring preconditioning proposed in [45] by defining a preconditioner that has the same boundary conditions of the original problem and then the same structure of the system matrix. We refer to this new technique as structure preserving reblurring preconditioning. The construction of our preconditioner requires two FFTs like the BCCB preconditioner in [45]. The proposal is further extended to provide a nonstationary preconditioning. Some numerical results shows the importance to preserve the structure of the matrix both in terms of quality of the restorations and robustness of the regularization parameter.

3.3.1 Reblurring preconditioning

Independently of the imposed BCs, if we require to deal with positive (semi-)definite matrices, as already mentioned in Subsection 3.2.3, the classical strategy is to pass from (3.1) to normal equations (3.12) and then to precondition as follows

$$DA^*Ax = DA^*b \qquad \text{or} \qquad D^{1/2}A^*AD^{1/2}D^{-1/2}x = D^{1/2}A^*b,$$

where D suitably approximate the (generalized) inverse of the normal matrix A^*A [82]. With a deeper look, we can say that the classical preconditioning scenario seems to be quaint, since the preconditioner D has to speed up the slowing down produced by A^*. On these grounds, in [45] the authors proposed a new technique, which uses a single preconditioning operator directly applied to the original system (3.1). The new preconditioner, called as *reblurring matrix* Z, according to the terminology in [62], leads to the new preconditioned system

$$ZAx = Zb\,. \tag{3.16}$$

As pointed out in [45], the aim of the preconditioner Z is to allow iterative methods to become more stable (as well as usually obtained through the normal equations involving A^*) without slowing the convergence (so that no subsequent accelerating operator D is needed), especially in the signal subspace. Solving (3.16) leads to reformulate iterative methods as the Landweber method in the new preconditioning context, that is to replace the following preconditioned iterative scheme

$$x_{k+1} = x_k + \tau DA^*(b - Ax_k)$$

with

$$x_{k+1} = x_k + \tau Z(b - Ax_k), \tag{3.17}$$

where τ is a positive relaxation parameter. In the following we fix $\tau = 1$, by applying an implicit rescaling of the preconditioned system matrix ZA. Although ZA is not in general symmetric, the convergence of the modified Landweber method (3.17) can be easily assured [45].

A way to build Z is to apply a coarsening technique to the PSF of the problem (see [45]); another way is to use filtering strategies. More in detail, in the case of periodic BCs, Z is built as the BCCB-matrix whose eigenvalues v_j are obtained from the eigenvalues $\lambda_j(A)$ of A using some filter. In particular, two very popular filters are the Hanke-Nagy-Plemmons (HNP) filter [82], defined as

$$v_j = \begin{cases} \overline{\lambda}_j(A)/|\lambda_j(A)|^2, & \text{if } |\lambda_j(A)| \geq \zeta, \\ \\ \overline{\lambda}_j(A), & \text{if } |\lambda_j(A)| < \zeta, \end{cases} \qquad j = 1, \ldots, n^2, \tag{3.18}$$

and the Tikhonov filter, defined as

$$v_j = \frac{\overline{\lambda}_j(A)}{|\lambda_j(A)|^2 + \alpha}, \qquad j = 1, \ldots, n^2, \tag{3.19}$$

where α and ζ are positive regularization parameters. For any BCs, Z is built as the BCCB-matrix obtained by applying filtering to the same PSF, i.e. the same generating function f, that gives rise to the matrix A. In other words, if we call g such filtered function and we employ the symbol notation, we have that $A = A_n(f)$ and $Z = C_n(g)$ (see next subsection for details on the computation of g). Clearly, in case of periodic BCs, since we are in the circulant algebra, the classical preconditioning approach (based on D) and this new one (based on Z) are the same thing. However, for other BCs, the Z variant shows better performance and higher stability (in relation to the choice of regularization parameters, e.g. α for Tikhonov) than standard preconditioning (see [45]).

3.3.2 Structure preserving extension

From negative results in [136], BCCB preconditioner Z used in [45] will never be optimal. Hence, here we define a class of preconditioners Z endowed with the same structure of the system matrix A. We call this strategy *structure preserving reblurring preconditioners*.

First of all, we compute the eigenvalues $c_{i,j}$ of $C_n(f)$, the $n^2 \times n^2$ BCCB-matrix associated with the PSF H, by means of the two-dimensional FFT of H. By definition $c_{i,j}$ are the evaluations of f, the generating function (3.2) of the blurring matrix A, on $\Gamma_n = \{(\frac{2\pi i}{n}, \frac{2\pi j}{n}) \,|\, i,j = 0,\dots,n-1\}$

$$c_{i,j} = f\left(\frac{2\pi i}{n}, \frac{2\pi j}{n}\right), \qquad i,j = 0,\dots,n-1.$$

We can now regularize such eigenvalues. Let us denote by $v_{i,j}$ the values obtained by applying the Tikhonov filter (3.19) to $c_{i,j}$, instead of $\lambda_j(A)$. Since it usually gives very good (and often the best) numerical results, we simply consider the Tikhonov filter (3.19) but any other filter could be applied as well. On the ground of the theory of eigenvalues decomposition of BCCB matrices, the values $v_{i,j}$ can be considered as a sampling of the function

$$g(x_1,x_2) = \sum_{j_1,j_2=-\lfloor\frac{n+1}{2}\rfloor}^{\lfloor\frac{n-1}{2}\rfloor} \beta_{j_1,j_2} \mathrm{e}^{\mathbf{i}(j_1x_1+j_2x_2)} \tag{3.20}$$

at the grid points Γ_n. Namely,

$$g\left(\frac{2\pi i}{n}, \frac{2\pi j}{n}\right) := v_{i,j} = \frac{\overline{c}_{i,j}}{|c_{i,j}|^2+\alpha} = \frac{\overline{f}(\frac{2\pi i}{n}, \frac{2\pi j}{n})}{|f(\frac{2\pi i}{n}, \frac{2\pi j}{n})|^2+\alpha}. \tag{3.21}$$

Note that g is a regularized approximation of the inverse of f on Γ_n. The function g is univocally identified by the n^2 interpolation conditions (3.21), for $i,j = 0,\dots,n-1$, and the coefficients β_{j_1,j_2} can be computed by means of a two-dimensional Inverse Fast Fourier Transform (IFFT) of these values $g(\frac{2\pi i}{n}, \frac{2\pi j}{n})$. It is worth observing that up to this point the described technique is just like the one proposed in [45]. The main difference between our approach and those in [45] is that in the latter the function g is used as the symbol of a BCCB-matrix, while here we combine g with the BCs of the problem defining a matrix

$$Z := A_n(g)$$

that has the same structure of the original system matrix $A = A_n(f)$ of our blurring model (3.1).

The following algorithm summarizes how to build our structure preserving preconditioner.

Algorithm 2 Structure preserving reblurring preconditioning

Input: H, BCs

1. get $\{c_{i,j}\}_{i,j=0}^{n-1}$ by computing an FFT of H
2. get $\{v_{i,j}\}_{i,j=0}^{n-1}$ by applying Tikhonov filter (3.19) to $c_{i,j}$, i.e., $v_{i,j} = \frac{\overline{c}_{i,j}}{|c_{i,j}|^2+\alpha}$
3. get $\widetilde{H}$ by computing an IFFT of $\{v_{i,j}\}_{i,j=0}^{n-1}$
4. generate the matrix Z from the coefficient mask $\widetilde{H}$ and BCs

Output: Z

From Algorithm 2, it is clear that the only difference between the reblurring preconditioning and our structure preserving extension lies in the fact that in the last step we exploit the BCs of the problem to build a preconditioner with the same structure of the blurring matrix. Obviously, the computational cost of the reblurring preconditioning of two FFTs, is not affected by this modification.

Throughout, we refer to the circulant preconditioning technique proposed in [45] as Z_{circ}, while our structure preserving preconditioner will be denoted by Z_{struct}.

Note that when the PSF is quadrantally symmetric, i.e. $h_{i,j} = h_{\pm i,\pm j}$, then, thanks to the Eulero formula, the generating function f is a cosine polynomial. Therefore, it is worthwhile looking for g in (3.21) as a cosine polynomial, which implies that Algorithm 2 can be reformulated by replacing FFT with DCT.

Comparison with Hanke–Nagy preconditioner For zero BCs and symmetric PSF we can seek a strict link between the proposed structure preserving preconditioner Z_{struct} and the Toeplitz preconditioning proposed by Hanke and Nagy in [81] for real symmetric Toeplitz systems. For simplicity, in analyzing the analogies and the differences between these techniques, we consider the one-dimensional case. Our aim in this paragraph is to study, in a sense that will be further explained, how much the two preconditioners are close. To recognize that the PSF H is symmetric, we fix the central pixel at the center of H. Moreover, we assume n even for the sake of simplicity in the computations. Therefore, the 1D PSF is $H = [h_{\frac{n}{2}-1}, \ldots, h_0, \ldots, h_{\frac{n}{2}-1}, 0]$ and the associated generating function is

$$f(x) = h_0 + 2\sum_{j=1}^{\frac{n}{2}-1} h_j \cos(jx), \tag{3.22}$$

obtaining the $n \times n$ blurring matrix $A = T_n(f)$ in the Toeplitz case, which is related to Zero Bcs.

We briefly recall the computation of the preconditioner proposed in [81]. First of all, the matrix $T_n(f)$ is embedded in the symmetric circulant $2n \times 2n$ matrix

$$C_{2n}(f) = \begin{pmatrix} T_n(f) & R \\ R & T_n(f) \end{pmatrix} \tag{3.23}$$

whose first column is given by $[h_0, \ldots, h_{\frac{n}{2}-1}, 0, \ldots, 0, h_{\frac{n}{2}-1}, \ldots, h_1]^T$. Then, the eigenvalues of C_{2n}, computed via FFT, are inverted by means of the HNP regularization filter (3.18) in order to obtain a regularized inverse of C_{2n}. Finally the preconditioner is selected as the first $n \times n$ principal submatrix of such a $2n \times 2n$ regularized inverse of C_{2n}.

The following proposition summarizes our result.

Proposition 12. *Let $\{T_n(f)\}_n$ be a sequence of $n \times n$ image deblurring matrices with zero boundary conditions, where f denotes the generating function* (3.22) *of a real fully-symmetric PSF. For any matrix $T_n(f)$, let $Z_{\text{struct},n}$ denote the associated structure preserving preconditioner of Algorithm 2 and $Z_{\text{HN},n}$ denote the associated inverse Toeplitz preconditioner by Hanke and Nagy [81], both regularized by the same Tikhonov filter* (3.19)*. Then, the two preconditioners $Z_{\text{struct},n}$ and $Z_{\text{HN},n}$ are asymptotically equivalent, that is,*

$$\|Z_{\text{struct},n} - Z_{\text{HN},n}\| \longrightarrow 0 \qquad \text{as } n \longrightarrow +\infty\,.$$

In particular, $\|Z_{\text{struct},n} - Z_{\text{HN},n}\| = \mathrm{O}\big(\log(n)\ \mathrm{e}^{-cn}\big)$, with $c > 0$.

Proof. First, we explicitly compute the $Z_{\text{struct},n}$ preconditioner of $T_n(f)$. Algorithm 2 considers the symmetric circulant matrix $C_n(f)$ whose first column is $[h_0, \ldots, h_{\frac{n}{2}-1}, 0, h_{\frac{n}{2}-1}, \ldots, h_1]^T$. Hence, the step 1 of the algorithm computes the eigenvalues of $C_n(f)$ that are

$$c_k = f\left(\frac{2k\pi}{n}\right) = h_0 + 2\sum_{j=1}^{\frac{n}{2}-1} h_j \cos\left(\frac{2jk\pi}{n}\right), \quad k = 0, \ldots, n-1.$$

At step 2 we adopt the Tikhonov filter (3.19), so that we compute the regularized inverses v_k of the eigenvalues as

$$v_k = \frac{\overline{c}_k}{c_k^2 + \alpha}, \quad k = 0, \ldots, n-1. \tag{3.24}$$

We can write $v_k = g\left(\frac{2k\pi}{n}\right)$, $k = 0, \ldots, n-1$, where $g \in \widetilde{\Pi}_{n-1} = \mathrm{span}\{\cos(jx), j = 0, \ldots, n-1\}$ is the trigonometric interpolating polynomial on the pairs $\left(\frac{2k\pi}{n}, v_k\right)$, $k = 0, \ldots, n-1$, that is

$$g(x) = \beta_0 + 2\sum_{j=1}^{n-1} \beta_j \cos(jx). \tag{3.25}$$

Actually, it can be easily proved that g, although it depends on n interpolation conditions, has degree $\frac{n}{2}$. Indeed, by rewriting the interpolation conditions as

$$\left(\tfrac{2k\pi}{n}, v_k\right) \qquad k = 0, \ldots, \tfrac{n}{2}$$

$$\left(\tfrac{(n/2+k)\pi}{n/2}, v_{n/2+k}\right) \quad k = 1, \ldots, \tfrac{n}{2} - 1$$

and observing that $v_k = v_{n-k}$, for $k = 1, \ldots, \frac{n}{2} - 1$, from the definition (3.24) and the symmetry of g in (3.25) with respect to the interval $[0,\ 2\pi]$, we have that $g \in \widetilde{\Pi}_{\frac{n}{2}}$. For this reason we rename g as $g_{\frac{n}{2}}$ in the following. The steps 3 and 4 build the preconditioner as the matrix that preserves the Toeplitz structure and whose symbol is $g_{\frac{n}{2}}$, hence for zero BCs, the preconditioner finally is

$$Z_{\text{struct},n} = T_n(g_{\frac{n}{2}})\,. \tag{3.26}$$

Now we consider the $Z_{\text{HN},n}$ preconditioner of $T_n(f)$. The eigenvalues of the $2n \times 2n$ circulant matrix $C_{2n}(f)$ defined by (3.23) are

$$\mu_k = f\left(\frac{2k\pi}{2n}\right) = f\left(\frac{k\pi}{n}\right), \quad k = 0, \ldots, 2n-1.$$

Applying again the Tikhonov filter (3.19) (differently to the HNP filter considered in [81]), we obtain the regularized inverses u_k of the eigenvalues as follows

$$u_k = \frac{\overline{\mu}_k}{\mu_k^2 + \alpha}, \quad k = 0, \ldots, 2n-1.$$

By means of an analogous reasoning to that performed before, it can be shown that there exists a unique polynomial $h_n \in \widetilde{\Pi}_n$ that interpolates the $2n$ pairs $\left(\frac{k\pi}{n}, u_k\right)$, $k = 0, \ldots, 2n-1$. In fact, due to their symmetry, among the previous interpolation conditions only the one relating to $k = 0, \ldots, n$ are distinct. The preconditioner $Z_{\text{HN},n}$ is then selected as the first $n \times n$ principal submatrix of the $2n \times 2n$ circulant matrix $C_{2n}(h_n)$. Denoted by S the $2n \times n$ matrix whose j-th column is the j-th element of the canonical basis of $\mathbb{R}^{2n}$, the preconditioner can be expressed as

$$Z_{\text{HN},n} = T_n(h_n) = S^* C_{2n}(h_n) S\,. \tag{3.27}$$

We can now relate the two preconditioners $Z_{\text{struct},n} = T_n(g_{\frac{n}{2}})$ and $Z_{\text{HN},n} = T_n(h_n)$. Let us observe that defining

$$\psi(x) = \frac{\overline{f}(x)}{f^2(x) + \alpha},$$

we have that $g_{\frac{n}{2}}$ interpolates ψ on $\Omega_{\frac{n}{2}} = \left\{\frac{2k\pi}{n}, \quad k = 0, \ldots, \frac{n}{2}\right\}$ and h_n interpolates the same ψ on $\Omega_n = \left\{\frac{k\pi}{n}, \quad k = 0, \ldots, n\right\}$. Since $\Omega_{\frac{n}{2}} \subset \Omega_n$ and $h_n \in \widetilde{\Pi}_n$, then we can write

$$h_n = g_{\frac{n}{2}} + p_n$$

where $p_n \in \widetilde{\Pi}_n$ such that $p_n\left(\frac{2k\pi}{n}\right) = 0$, for $k = 0, \ldots, \frac{n}{2}$. On this ground we rewrite

$$p_n = E_{\frac{n}{2}} - E_n$$

where $E_{\frac{n}{2}} = \psi - g_{\frac{n}{2}}$ and $E_n = \psi - h_n$ are the classical remainder in the interpolation of ψ on $\Omega_{\frac{n}{2}}$ and Ω_n, respectively. By virtue of the linearity of T_n and by using (1.18), we have

$$\begin{aligned} \|T_n(g_{\frac{n}{2}}) - T_n(h_n)\| &= \|T_n(g_{\frac{n}{2}}) - T_n(g_{\frac{n}{2}}) - T_n(p_n)\| = \|T_n(E_{\frac{n}{2}} - E_n)\| \\ &\leq \|E_{\frac{n}{2}}\|_\infty + \|E_n\|_\infty. \end{aligned} \tag{3.28}$$

Here $\|\cdot\|_\infty = \|\cdot\|_{L^\infty(0,2\pi)}$. By construction, E_n is the remainder function of the interpolation of ψ on the $n+1$ Chebyshev-Lobatto nodes defined as $\cos(\frac{k\pi}{n})$ for $k = 0, \ldots, n$. Its Lebesgue constant is known to grow as $k_1 \log(n)$ where k_1 is a constant. Thus, $\|E_n\|_\infty$ can be bounded in the following way

$$\|E_n\|_\infty \leq k_1 \log(n) \|a_n\|_\infty,$$

where $\|a_n\|_\infty$ is the error in the best approximation of ψ in the space of polynomials of degree at most n. Applying a similar reasoning to $E_{\frac{n}{2}}$, we can bound it as $\|E_{\frac{n}{2}}\|_\infty \leq k_2 \log(\frac{n}{2}) \|a_{\frac{n}{2}}\|_\infty$. Because of the C^∞ regularity of ψ, $\|a_r\|$ is exponentially converging to zero as r tends to $+\infty$ (by Bernstein Theorem). In conclusion, from (3.28) it follows that $\|T_n(g_{\frac{n}{2}}) - T_n(h_n)\|$ is exponentially converging to 0 as n tends to ∞, that is, the two preconditioners $Z_{\text{struct},n} = T_n(g_{\frac{n}{2}})$ and $Z_{\text{HN},n} = T_n(h_n)$ are asymptotically the same. □

It is interesting to notice that the equivalence result of the previous proposition is confirmed by several numerical tests, where basically we got the same deblurring accuracy and the same convergence speed by applying the two preconditioners. On the other hand, from a computational point of view $Z_{\text{struct},n}$ of (3.26) requires two FFTs of size n instead of two FFTs of size $2n$ to obtain $Z_{\text{HN},n}$ of (3.27) .

Remark 18. In [81] the authors consider only Toeplitz deconvolution matrices, that is only zero BCs problems, but the same approach could be applied also to the other BCs discussed in Section 3.1 like reflective and anti-reflective BCs. Moreover, it could be extended also to nonsymmetric PSF.

3.3.3 Nonstationary preconditioning

When we deal with a stationary regularization method we have always to face the non-trivial task of determining a good choice for the filtering parameter α. In [55] the authors proposed a nonstationary method that can be interpreted as a nonstationary version of the reblurring preconditioning where the parameter α is dynamically estimated at every iteration instead of to be fixed a-priori. The iteration is the following

$$x_{k+1} = x_k + Z^k_{\text{circ}} r_k\,, \qquad Z^k_{\text{circ}} = C^*(CC^* + \alpha_k I)^{-1}\,, \qquad r_k = b - Ax_k, \tag{3.29}$$

where $C = C_n(f)$ is the BCCB-matrix associated with H. Note that if $\alpha_k = \alpha$ then the iteration (3.29) is exactly (3.17) with $Z = C_n(g)$, where g is defined in (3.20) imposing (3.21) as in [45]. Actually, the proposal in [55] was designed as an approximate version of the iterated Tikhonov.

Remark 19. Z^k_{circ} is the BCCB-matrix obtained by Algorithm 2 with α_k in place of α in point 2. and with periodic BCs.

In [55], the iteration-dependent regularization parameter α_k is obtained by solving the following nonlinear equation

$$\|r_k - CZ^k_{\text{circ}} r_k\|_2 = q_k \|r_k\|_2\,, \qquad 0 < q_k < 1, \tag{3.30}$$

with a few steps of the Newton iteration. Here the parameter q_k depends on the noise level and it is related to a value $0 < \rho_{\text{circ}} < 1/2$ which satisfies the assumption

$$\|(C - A)z\|_2 \le \rho_{\text{circ}} \|Az\|_2\,, \qquad \forall\, z \in \mathbb{R}^n \tag{3.31}$$

(see [55] for further details). This parameter ρ_{circ} measures how much we trust in the approximation of A with its BCCB counterpart C; the smaller ρ_{circ} is, the more we trust in that approximation (see [55] for more details). From a numerical point of view, the parameter ρ_{circ} is usually chosen among the values $\{10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}\}$. A too small ρ_{circ} can be easily recognized looking at the sequence of the regularization parameters. The iterations are stopped under a special choice of the discrepancy principle that will be discussed later. From a theoretical point of view, in [55] it is proved that under the assumption (3.30), the iteration (3.29) converge monotonically and it is a regularization method.

In this subsection we extend the nonstationary iteration (3.29) to take into account the BCs of the problem following our structure preserving strategy. More precisely, we consider the following iteration

$$x_{k+1} = x_k + Z^k_{\text{struct}} r_k\,, \qquad r_k = b - Ax_k, \tag{3.32}$$

where Z^k_{struct} is the structure preserving matrix built by means of Algorithm 2 with a special choice of the regularization parameter α_k. In practice, we would estimate α_k by solving

$$\|r_k - AZ^k_{\text{struct}} r_k\|_2 = q_k \|r_k\|_2, \tag{3.33}$$

which is not computationally practicable. Therefore, we estimate the regularization parameter α_k using an approximation of (3.33) easily computable that is the equation (3.30). A similar strategy was adopted in [42] for different regularization methods. Note that the two iterations (3.29) and (3.32) retrieve a different sequence of α_k, even if both use the equation (3.30) to estimate α_k, because the sequences of the residuals $\{r_k\}_k$ in the two iterative schemes are different. Furthermore, we have to observe that for such a modified version of (3.29) the condition (3.31) does not make sense and then we are allowed to introduce a new parameter ρ_{struct} whose value could not match with the one of ρ_{circ}.

Remark 20. Since Z^k_{struct} provides a better approximation of $A^*(AA^* + \alpha_k I)^{-1}$ with respect to Z^k_{circ}, it is expected that $\rho_{\text{struct}} < \rho_{\text{circ}}$.

On the other hand, we cannot prove convergence results as in [55], even if the numerical results in next subsection shows that our structured nonstationary preconditioning is robust and very effective.

As shown in [55], another suitable choice of the parameter α_k for iteration (3.29) is given by the geometric sequence

$$\alpha_k = \tilde{\alpha} q^k, \qquad k = 0, 1, \ldots, \tag{3.34}$$

where $\tilde{\alpha} > 0$ and $0 < q \le 1$. In the next subsection we confirm the effectiveness of the iteration (3.32) with both sequences of $\{\alpha_k\}$ obtained by (3.30) or (3.34).

3.3.4 Numerical results

In this subsection we compare the reblurring preconditioning with our structure preserving extension. In each example we impose appropriate BCs and solve the linear system $Ax = b$ using both stationary and nonstationary preconditioned iterations (3.17) and (3.32). More in detail, in the stationary case, fixed few values of the parameter α, we compare the performances of the preconditioner Z_{circ} with our Z_{struct}. In the nonstationary case our attention is devoted to the comparison between preconditioners Z^k_{circ} and Z^k_{struct}. Regarding the sequence $\{\alpha_k\}_k$ of the regularization parameter we investigate the behavior of both the geometric sequence defined in (3.34) (labeled 'geometric' in the following) and the sequence computed solving (3.30) (throughout, labeled 'DH' (Donatelli-Hanke)). For the geometric sequence we fix $\tilde{\alpha} = 0.5$ and $q = 0.7$ in (3.34), as suggested in [55]. Furthermore, we compare all the nonstationary preconditioned algorithms with the CGLS.

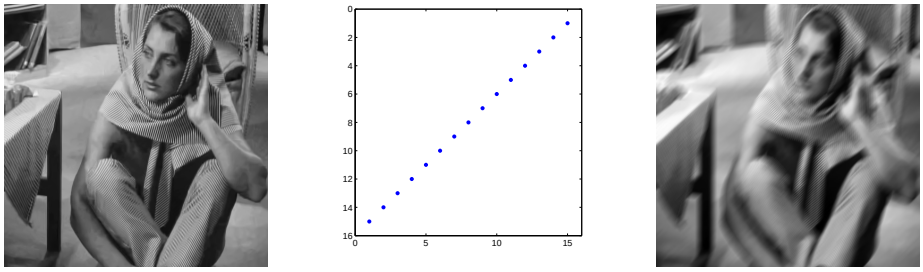
As stopping criterion, we use the discrepancy principle given in (3.14). In the following we fix $\gamma = (1+2\rho_{\text{circ}})/(1-2\rho_{\text{circ}})$ (or $\gamma = (1+2\rho_{\text{struct}})/(1-2\rho_{\text{struct}})$ dependently on the choice of the preconditioner as Z^k_{circ} or Z^k_{struct}) for the iterations (3.29) and (3.32) with the DH sequence of regularization parameter (see [55]), while we choose $\gamma = 1.01$ for the other algorithms.

The initial guess x_0 is always taken as the observed image b. Assuming to know the true image $\bar{x}$, we measure the quality of the reconstruction computing the RRE defined in (3.13). We refer to the minimum RRE and to the RRE corresponding to the discrepancy principle iteration as RRE_{min} and $\text{RRE}_{\text{discr}}$, respectively.

All the numerical tests have been developed with Matlab R2011b on a PC Intel Core$^{\text{TM}}$i5 and Windows 7 operating system.

Example 1

We start with the Barbara deblurring problem of size 497×497 in Figure 3.2. The PSF is a motion of size 15×15 given by the antidiagonal matrix in Figure 3.2(b) whose nonzero entries are equal to $1/15$. In this example we impose reflective BCs and fix the noise level to 1%.



(a) (b) (c)

Figure 3.2: Example 1 - (a) true image 497×497; (b) motion PSF 15×15; (c) blurred image

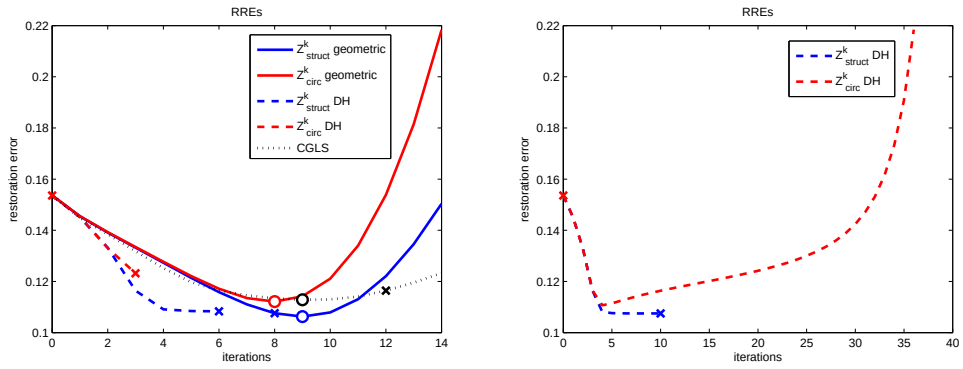
In Table 3.1 we compare Z_{struct} and Z_{circ} preconditioners for $\alpha = 0.5, 0.1, 0.05, 0.01$. We observe that both RRE_{min} and $\text{RRE}_{\text{discr}}$ provided by Z_{struct} preconditioner are smaller than the RRE_{min} obtained using the Z_{circ} one. Furthermore, the discrepancy principle does not work for Z_{circ} preconditioner when $\alpha = 0.5, 0.1, 0.05$.

	$\alpha = 0.5$		$\alpha = 0.1$		$\alpha = 0.05$		$\alpha = 0.01$	
	RRE_{min}	$\text{RRE}_{\text{discr}}$	RRE_{min}	$\text{RRE}_{\text{discr}}$	RRE_{min}	$\text{RRE}_{\text{discr}}$	RRE_{min}	$\text{RRE}_{\text{discr}}$
Z_{struct}	0.1084 (46)	0.1093 (34)	0.1072 (10)	0.1077 (8)	0.1068 (5)	0.1068 (5)	0.1070 (1)	0.1084 (2)
Z_{circ}	0.1138 (31)	-(-)	0.1125 (7)	-(-)	0.1115 (4)	-(-)	0.1096 (1)	0.1145 (2)

Table 3.1: Example 1 - RRE_{min} and $\text{RRE}_{\text{discr}}$ and corresponding iterations (in parenthesis) for Z_{circ} and Z_{struct} preconditioners.

Figures 3.3(a)(b) refer to the comparison of the nonstationary preconditioning with Z^k_{struct} for both geometric and DH sequences with the Z^k_{circ} preconditioner and with the CGLS method. When possible, together with the discrepancy iteration, we also show the iteration corresponding to the RRE_{min} (by construction for Z^k_{circ} and Z^k_{struct} with DH sequence only $\text{RRE}_{\text{discr}}$ is available). In the nonstationary Z^k_{circ} context we fix $\rho_{\text{circ}} = 10^{-1}$. Such a choice of ρ_{circ} is due to the fact that for smaller values of

this parameter, the Z^k_{circ} method does not converge (see the behavior of the RRE for Z^k_{circ} with DH sequence in Figure 3.3(b) in which $\rho_{\text{circ}} = 10^{-2}$). As regards the parameter ρ_{struct}, we use both 10^{-2} and 10^{-3} and observe that the Z^k_{struct} preconditioner works well for both choices of ρ_{struct}. Note that the previous behaviour of the parameter ρ_{circ} and ρ_{struct} agree with Remark 20. This highlights an appreciable stability of our algorithm with respect to the parameter ρ_{struct} and allows us to focus in the following numerical results and reconstructions only on one of the two values considered for ρ_{struct}. Although Z^k_{struct} is slightly more accurate for $\rho_{\text{struct}} = 10^{-3}$ ($\text{RRE}_{\text{discr}} = 0.1075$) than for $\rho_{\text{struct}} = 10^{-2}$ ($\text{RRE}_{\text{discr}} = 0.1083$) (compare also Figure 3.3(a) with Figure 3.3(b)), we decide for the last one since in this case the nonstationary structure preserving method reveals faster (6 iterations rather than 10 iterations). In this regard, we observe that the geometric and DH sequences of the regularization parameter give rise to comparable errors for Z^k_{struct} method, but the DH sequence involves less iterations.



(a) $\rho_{\text{circ}} = 10^{-1}$, $\rho_{\text{struct}} = 10^{-2}$ (b) $\rho_{\text{circ}} = 10^{-2}$, $\rho_{\text{struct}} = 10^{-3}$

Figure 3.3: Example 1 - (a) Comparisons between RREs for Z^k_{struct} geometric (solid blue line), Z^k_{circ} geometric (solid red line), Z^k_{struct} DH (dashed blue line), Z^k_{circ} DH (dashed red line) and CGLS (dotted black line) for $\rho_{\text{circ}} = 10^{-1}$ and $\rho_{\text{struct}} = 10^{-2}$; (b) Comparison between Z^k_{struct} DH (dashed blue line) and Z^k_{circ} DH (dashed red line) for $\rho_{\text{circ}} = 10^{-2}$ and $\rho_{\text{struct}} = 10^{-3}$. Key to symbols: (∘) optimal iteration, (×) discrepancy iteration.

	geometric		DH		CGLS
	Z^k_{struct}	Z^k_{circ}	Z^k_{struct}	Z^k_{circ}	
RRE_{min}	0.1063 (9)	0.1121 (8)	n/a	n/a	0.1128 (9)
$\text{RRE}_{\text{discr}}$	0.1076 (8)	-(-)	0.1083 (6)	0.1232 (3)	0.1164 (12)

Table 3.2: Example 1 - RRE_{min} and $\text{RRE}_{\text{discr}}$ and corresponding iterations (in parenthesis) for Z^k_{circ} and Z^k_{struct} preconditioners for both the geometric and DH sequences and for the CGLS method. We fix $\rho_{\text{circ}} = 10^{-1}$ and $\rho_{\text{struct}} = 10^{-2}$.

The better performances of Z^k_{struct} with respect to the Z^k_{circ} and CGLS already observed in Figure 3.3 are confirmed in Table 3.2 (within 'n/a' means that the corresponding RRE_{min} is not available). It is evident that not only the RRE_{min}, but also the $\text{RRE}_{\text{discr}}$ of both geometric and DH Z^k_{struct} algorithms are smaller than the RRE_{min} provided by the geometric Z^k_{circ} preconditioner and by the CGLS. Furthermore, note that the discrepancy principle does not work for the Z^k_{circ} preconditioner with geometric sequence. A comparison between the reconstructions in Figures 3.4(a)(c) and the ones in Figures 3.4(b)(d) highlights the effectiveness of the proposed nonstationary structure preserving preconditioners.

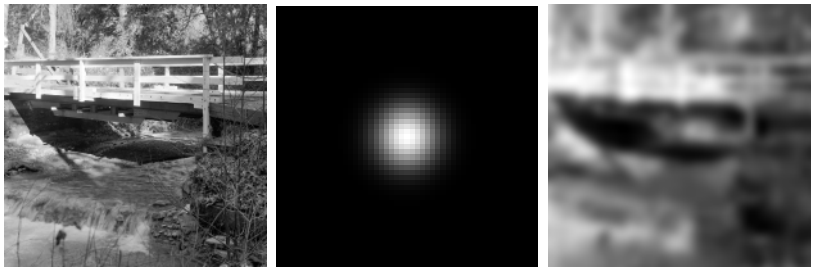
Example 2

This example refers to the bridge deblurring problem of size 205×205 in Figure 3.5. The PSF is a 51×51 pixels cropped and normalized portion of the GaussianBlur440 coming from the Restore Tools Matlab package ([106]). We choose antireflective BCs and fix the noise level to 0.2%.

Aside from the comparison between Z_{struct} and Z_{circ} (or Z^k_{struct} and Z^k_{circ} in the nonstationary case), the symmetrical nature of the PSF of this example let us to compare the performances of our preconditioner with a preconditioner obtained imposing reflexive BCs. In fact, as already observed, when the PSF is symmetric in Algorithm 2 we can use DCT instead of FFT. We refer to such a preconditioner as Z_{DCT} in the stationary case and as Z^k_{DCT} in the nonstationary one. Furthermore, we denote by

(a) Z^k_{struct} geometric (b) Z^k_{circ} geometric (c) Z^k_{struct} DH (d) Z^k_{circ} DH

Figure 3.4: Example 1 - (a) Discrepancy reconstruction with Z^k_{struct} geometric; (b) Optimal reconstruction with Z^k_{circ} geometric; (c) Discrepancy reconstruction with Z^k_{struct} DH; (d) Discrepancy reconstruction with Z^k_{circ} DH.



(a) (b) (c)

Figure 3.5: Example 2 - (a) true image 205×205; (b) GaussianBlur440 PSF 51×51; (c) blurred image

ρ_{DCT} the parameter which 'quantifies' how much the structure of the preconditioner matches with the structure of the blurring matrix.

In Table 3.3, we fix $\alpha = 0.01, 0.05, 0.001, 0.0005$ and show the RRE_{min}, $\text{RRE}_{\text{discr}}$ and corresponding iterations provided by the iteration (3.17) with Z_{circ}, Z_{struct} and Z_{DCT} preconditioners. As for Example 1, the $\text{RRE}_{\text{discr}}$ of the Z_{struct} method, for $\alpha = 0.05, 0.001, 0.0005$, is smaller than the RRE_{min} of the Z_{circ} one and in this last case the discrepancy principle does not work. Furthermore, the Z_{struct} seems to be less sensitive to the choice of α, while the RREs obtained using the Z_{circ} preconditioner increases as the value of α decreases. Concerning the Z_{DCT} preconditioner it results more accurate than Z_{circ} and Z_{struct} for $\alpha = 0.01$, but it converges very slowly and for the other three values of α it behaves almost like Z_{circ}.

In Figure 3.7(a) we compare the nonstationary Z^k_{struct} method with the Z^k_{circ} and CGLS ones. We fix $\rho_{\text{circ}} = 10^{-1}$ and $\rho_{\text{struct}} = 10^{-3}$. The choice of ρ_{circ} as 10^{-1} is in same sense forced from the fact that, as shown in Figure 3.7(b), for smaller values of that parameter (contextually $\rho_{\text{circ}} = 10^{-2}$) the DH sequence of the regularization parameter becomes to zigzag up and down and the algorithm attempts to invert noise components.

The Z^k_{struct} method is for the geometric and especially for the DH sequence more accurate and faster than the CGLS. In fact, as shown in Table 3.4 the discrepancy principle stops the CGLS iteration after 148 steps, while for the Z^k_{struct} with DH sequence only 10 iterations are needed (for this reason we omit the discrepancy and optimal iterations for the CGLS method in Figure 3.7(a)).

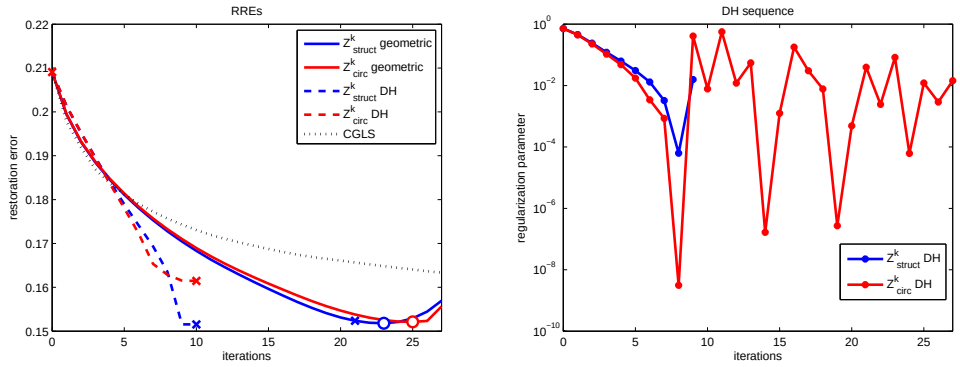
Regarding the comparison of Z^k_{struct} with Z^k_{circ}, the RRE provided by our algorithm is smaller than the ones obtained using Z^k_{circ} preconditioner for both sequences of the regularization parameter (compare Figures 3.6(a)(c) with Figures 3.6(b)(d)).

	$\alpha = 0.01$		$\alpha = 0.05$		$\alpha = 0.001$		$\alpha = 0.0005$	
	RRE_{min}	$\text{RRE}_{\text{discr}}$	RRE_{min}	$\text{RRE}_{\text{discr}}$	RRE_{min}	$\text{RRE}_{\text{discr}}$	RRE_{min}	$\text{RRE}_{\text{discr}}$
Z_{struct}	0.1519 (148)	0.1531 (58)	0.1519 (74)	0.1530 (30)	0.1519 (15)	0.1527 (7)	0.1518 (8)	0.1524 (4)
Z_{circ}	0.1522 (277)	-(-)	0.1577 (15)	-(-)	0.1596 (2)	-(-)	0.1593 (2)	-(-)
Z_{DCT}	0.1496 (279)	0.1515 (78)	0.1571 (13)	-(-)	0.1604 (2)	-(-)	0.1592 (1)	-(-)

Table 3.3: Example 2 - RRE_{min} and $\text{RRE}_{\text{discr}}$ and corresponding iterations (in parenthesis) for Z_{circ}, Z_{struct}, Z_{DCT} preconditioners.

(a) Z^k_{struct} geometric (b) Z^k_{circ} geometric (c) Z^k_{struct} DH (d) Z^k_{circ} DH

Figure 3.6: Example 2 - (a) Discrepancy reconstruction with Z^k_{struct} geometric; (b) Optimal reconstruction with Z^k_{circ} geometric; (c) Discrepancy reconstruction with Z^k_{struct} DH; (d) Discrepancy reconstruction with Z^k_{circ} DH.



(a) $\rho_{\text{circ}} = 10^{-1}$, $\rho_{\text{struct}} = 10^{-3}$ (b) $\rho_{\text{circ}} = 10^{-2}$, $\rho_{\text{struct}} = 10^{-3}$

Figure 3.7: Example 2 - (a) Comparisons between RREs for Z^k_{struct} geometric (solid blue line), Z^k_{circ} geometric (solid red line), Z^k_{struct} DH (dashed blue line), Z^k_{circ} DH (dashed red line) and CGLS (dotted black line). Key to symbols: (○) optimal iteration, (×) discrepancy iteration; (b) DH sequence for Z^k_{struct} (bullet blue line), Z^k_{circ} (bullet red line).

In Table 3.4 we show also the results obtained with Z^k_{DCT} preconditioner. In the case of the geometric sequence the Z^k_{DCT} method is more accurate than the others. Nevertheless, note that the discrepancy principle is useful only for our Z^k_{struct} preconditioner. As regards the parameter ρ_{DCT} we allow smaller values than 10^{-1} previously fixed for ρ_{circ}, since we expect that a preconditioner built with reflective BCs better preserves the structure of the problem induced by antireflective BCs than a BCCB preconditioner. On the other hand, we have numerically observed that for $\rho_{\text{DCT}} = 10^{-3}$ the approximation becomes less accurate and more iterations are involved ($\text{RRE}_{\text{discr}} = 0.1536$ after 43 iterations), so we decide for $\rho_{\text{DCT}} = 10^{-2}$. As shown in Table 3.4, although Z^k_{DCT} improves the quality of the reconstruction of Z^k_{circ} (note that in doing this more iterations are needed), it cannot perform better than our structure preserving preconditioner.

	geometric			DH			CGLS
	Z^k_{struct}	Z^k_{circ}	Z^k_{DCT}	Z^k_{struct}	Z^k_{circ}	Z^k_{DCT}	
RRE_{min}	0.1518 (23)	0.1521 (25)	0.1495(25)	n/a	n/a	n/a	0.1526 (162)
$\text{RRE}_{\text{discr}}$	0.1524 (21)	-(-)	-(-)	0.1515 (10)	0.1615 (10)	0.1529 (19)	0.1527 (148)

Table 3.4: Example 2 - RRE_{min} and $\text{RRE}_{\text{discr}}$ and corresponding iterations (in parenthesis) for Z^k_{circ}, Z^k_{struct} and Z^k_{DCT} preconditioners for both the geometric and DH sequences and for the CGLS method. We fix $\rho_{\text{circ}} = 10^{-1}$, $\rho_{\text{DCT}} = 10^{-2}$ and $\rho_{\text{struct}} = 10^{-3}$.

3.4 A Regularization preconditioning in the Fourier domain

As already observed in Subsection 3.2.1, when the convolution matrix in (3.1) can be diagonalized by two-dimensional DFT (1.22), a computationally attractive technique is given by the Tikhonov regularization. On the other hand, the provided solutions are usually oversmoothed and other regularization terms, involving e.g. total variation or the 1-norm, are often employed to preserve the edges or the sparsity of the original image. Of course, this can lead to nonlinear systems and then weighs down on the computational cost of the regularization method.

A useful strategy to reduce such nonlinear complexity is the Iteratively Reweighted Least Squares (IRLS) strategy whose aim is to approximate nonlinear regularization terms introducing a diagonal invertible weighting matrix. However, despite the simplicity of this weighting matrix, the resulting linear systems cannot be solved by FFTs.

Starting from the fact that images have sparse representations in the Fourier and wavelet domains, many deconvolution methods have been recently proposed with the aim of minimizing the 1-norm of these transformed coefficients. Such a minimization is known as synthesis approach.

In this section we combine the synthesis approach with the IRLS strategy approximating the 1-norm, in order to introduce a diagonal weighting matrix in the Fourier domain. The resulting linear system is diagonal and hence the regularization parameter can be easily estimated, for instance by the generalized cross validation. We will point out that the proposed Tikhonov regularization can be interpreted as a diagonal regularization preconditioner.

The method benefits from a proper initial approximation that can be the observed image or the Tikhonov approximation, therefore, embedding this method in an outer iteration may yield a further improvement of the solution. Finally, since some properties of the observed image, like continuity or sparsity, are obviously changed when working in the Fourier domain, we introduce a filtering factor which keeps unchanged the large singular values and preserves the jumps in the Fourier coefficients related to the low frequencies. Numerical examples are given in order to show the effectiveness of the proposed method.

3.4.1 IRLS and synthesis approach

The generalized Tikhonov regularization described in Subsection 3.2.1 usually provides oversmoothed reconstructions and it does not preserve the edges of the original image. Therefore, other penalty terms $\mathcal{R}(x)$ in (3.4) have been proposed in the literature, like the total variation regularization to preserve the edges [116], or

$$\mathcal{R}(x) = \|x\|_1, \tag{3.35}$$

to impose sparsity in the restored image [153]. Clearly, the solution of the resulting nonlinear problem is much more time consuming than Tikhonov regularization [147, 153]. *IRLS strategy* has been extensively used to reduce such nonlinear complexity [21, 152]. Recently, it has been successfully applied to deal with the 1-norm regularization term (3.35) in connection with iterative methods, like the preconditioner proposed in [93] for conjugate gradient, or the hybrid Arnoldi-Tikhonov methods proposed in [76]. IRLS method approximates the regularization term (3.35) by a term of the form (3.5) with a diagonal invertible matrix L. Despite the simplicity of L, the structure of the coefficient matrix in (3.6) is lost and such linear systems cannot longer be solved by fast transforms.

Starting from the a-priori information that every image has a sparse representation in the wavelet or Fourier coefficients (see [102]), a further class of regularization methods investigated in the last ten years considers the following regularization problem

$$\min_{\widehat{x}\in\mathbb{C}^m} \left\{ \|AW^*\widehat{x} - b\|_2^2 + \alpha\|\widehat{x}\|_1 \right\}, \tag{3.36}$$

where $x = W^*\widehat{x}$ and $W \in \mathbb{C}^{m\times n}$, with $m \geq n$, is a wavelet or tight-frame synthesis operator such that $W^*W = I$, cf. [70, 29]. The minimization problem (3.36) is usually called *synthesis approach*. If $m = n$ and it holds that $A = W^*\Lambda_A W$, then W is a unitary transform and the data fitting term in (3.36) becomes

$$\|AW^*\widehat{x} - b\|_2^2 = \|\Lambda_A\widehat{x} - Wb\|_2^2,$$

which can be easily computed only by vector operations.

In the next subsections we consider the synthesis approach (3.36) where A can be diagonalized by W and IRLS method is used to approximate the regularization term $\|\widehat{x}\|_1$ based on a diagonal matrix built by using an approximation of $\widehat{x}$ that can be computed with an easy and fast regularization strategy.

Furthermore, we include also a filter factor to avoid spoiling the information related to large singular values. Since the Fourier transformed data vector $\widehat{x}$ could exhibit some jumps, we propose a further approach designed in order to preserve discontinuity of the data. Summarizing, the presented technique exploits the properties of both blurring operator and data, and in addition it allows cheap computation of an optimal regularization parameter.

For a clear presentation of our new method, in the following, we choose the Fourier domain, i.e., W is the DFT. Hence we consider the classical image deblurring model where the matrix A is BCCB matrix which is diagonalizable by the DFT [85]. Nevertheless, the same approach can be extended to other structures, as for instance to the reflective boundary conditions when the PSF is symmetric in every direction because, as observed in Section 3.1, in this case the matrix A can be diagonalized by DCT.

3.4.2 IRLS for Tikhonov regularization

In [93] the authors introduced a data based regularization technique to improve the reconstruction of the generalized Tikhonov method by incorporating the values of the image in the penalty term by IRLS method. More precisely, given an approximation y of the true image, e.g. computed by applying the generalized Tikhonov method with $L = I$, the idea that underlies this technique is to construct the following diagonal matrix $D_y \in \mathbb{R}^{n\times n}$

$$(D_y)_{ii} = \begin{cases} |y_i| & \text{if } |y_i| > \varepsilon \\ \varepsilon & \text{otherwise} \end{cases}, \qquad i = 1, \dots, n, \tag{3.37}$$

with $0 < \varepsilon \ll 1$, in order to guarantee the invertibility of D_y, and to choose $L = D_y^{-\frac{1}{2}}$ in the generalized Tikhonov method. In detail, the problem to be solved becomes

$$\min_{x\in\mathbb{R}^n} \left\{ \|Ax - b\|_2^2 + \alpha \|D_y^{-\frac{1}{2}} x\|_2^2 \right\}. \tag{3.38}$$

Note that, for $x \in \mathbb{R}^n$, such that $|x_i| > \epsilon$, $i = 1, \dots, n$, the following equality holds

$$\|D_x^{-\frac{1}{2}} x\|_2^2 = x^* D_x^{-1} x = \|x\|_1.$$

Hence, the reason for the aforementioned choice of L lies in the fact that

$$y \approx x \;\Rightarrow\; \|D_y^{-\frac{1}{2}} x\|_2^2 \approx \|x\|_1. \tag{3.39}$$

Therefore, the regularization term in (3.38) is a good approximation of the regularization term (3.35) whenever y is a good approximation of x. Several numerical tests show that the approximation (3.39) of $\|x\|_1$ is not very sensitive to the choice of ε.

In other words, this method is a particular generalized Tikhonov method where the minimum problem (3.38) is equivalent to the linear system

$$(A^*A + \alpha D_y^{-1})x = A^*b, \tag{3.40}$$

that, as shown in [93], for sparse images provides better reconstructions than the generalized Tikhonov method with $L = I$. A drawback of this technique is that the diagonal matrix D_y^{-1} prevents the use of fast transform based algorithms for solving the linear system (3.40).

In order to improve the restoration starting from an approximation y, one of the methods known in the literature is the iterated Tikhonov regularization [68]

$$\min_{x\in\mathbb{C}^n} \left\{ \|Ax - b\|_2^2 + \alpha \|x - y\|_2^2 \right\}, \tag{3.41}$$

where the jumps are nearly removed by considering the difference between the current approximation x and the previous one y. Exploiting the same idea of the method (3.41), we can consider the following minimization problem

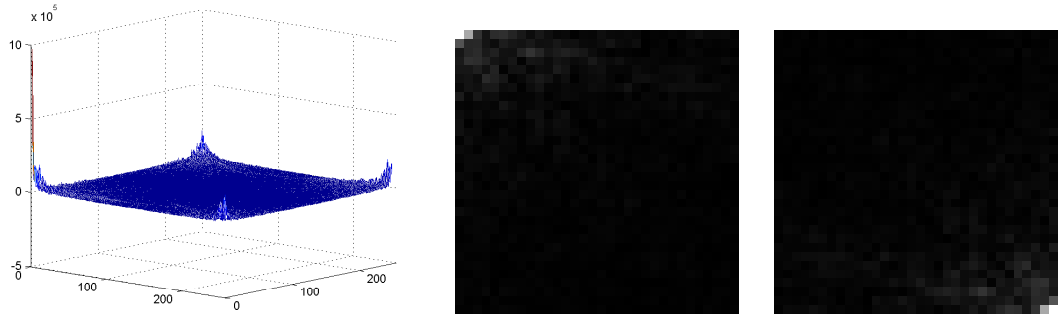
$$\min_{x\in\mathbb{C}^n} \left\{ \|Ax - b\|_2^2 + \alpha \|D_y^{-1} x\|_2^2 \right\}, \tag{3.42}$$

which can be seen as a multiplicative iterated Tikhonov regularization with the difference replaced by the quotient. Note that (3.42) maps the data vector $D_y^{-1}x$ approximately on the vector $e^T = (1, \dots, 1)^T$ which is smooth and not sparse, so the method (3.42) should be used in connection with a further (filtering) factor L that removes the regularizing effect for smooth vectors, e.g. the discretization of the derivatives or an operator dependent matrix like (3.9). Hence, (3.42) allows to treat the data as a smooth and continuous vector and to apply related regularization techniques [90, 92].

3.4.3 IRLS for Tikhonov regularization in the Fourier domain

In this subsection we show how to apply the IRLS strategy for Tikhonov regularization in the Fourier domain. The aim is to work with generic images, that are not necessarily sparse, preserving at the same time the computational efficiency of the FFT. More precisely, in spite of solving (3.38) in the space domain, we take advantage of the well-known sparsity of the Fourier coefficients of an image (see the example below and [102] for more details) and solve it in the Fourier domain.

Example (sparsity of the Fourier coefficients of an image). In Figure 3.8(a) we represent the moduli of the Fourier coefficients of the non sparse cameraman image (see Figure 3.8) as a surface plot obtaining a surface which is almost squashed on the plane $z = 0$. In other words, if that moduli are displayed as a scaled image the resulting image is, except for very few pixels in the corners, black everywhere. To highlight the presence of non zero pixels in the corners, in Figures 3.8(b) and 3.8(c) we show a zoom of the North-West and South-East corners, respectively.



(a) (b) (c)

Figure 3.8: (a) Moduli of the Fourier coefficients of the cameraman image of size 256×256 represented as a surface plot; (b) North-West corner of size 32×32 of the moduli of the Fourier coefficients of the cameraman image displayed as an image; (c) South-East corner of size 32×32 of the moduli of the Fourier coefficients of the cameraman image displayed as an image plot

Tikhonov regularization and filtering with Sparse Fourier coefficients (TSF)

Regularization. Let $\widehat{x} = Fx$, $\widehat{x} \in \mathbb{C}^n$ be the Fourier coefficients of x. Reformulating the problem (3.38) in terms of $\widehat{x}$ instead of x, it becomes

$$\min_{\widehat{x} \in \mathbb{C}^n} \left\{ \|AF^*\widehat{x} - b\|_2^2 + \alpha \|D_{\widehat{y}}^{-\frac{1}{2}} \widehat{x}\|_2^2 \right\}, \tag{3.43}$$

where y is an approximation of x and $\widehat{y} = Fy$ are the Fourier coefficients of y. To come back in the space domain it is sufficient to apply an inverse transform to the computed solution of (3.43), i.e., $x = F^*\widehat{x}$. The initial approximation y can be computed, for example, by the Tikhonov method. If y is a good approximation of x, since $\|y - x\|_2 = \|\widehat{y} - \widehat{x}\|_2$, then also $\widehat{y}$ is a good approximation of $\widehat{x}$ and hence the regularization term satisfies

$$\|D_{\widehat{y}}^{-\frac{1}{2}} \widehat{x}\|_2^2 \approx \|\widehat{x}\|_1.$$

It follows that the problem (3.43) is an approximation of the regularization problem by the synthesis approach in (3.36), where $W = F$. Moreover, recalling Remark 17, the TSF method can be seen as a regularization preconditioning technique in the Fourier domain whose preconditioner is the diagonal matrix $D_{\widehat{y}}^{-\frac{1}{2}}$.

Filtering. In the previous approach we have introduced a minimization problem in which the penalty term acts by modifying also the large singular values of the matrix A, that are those associated to the low frequencies which are less sensitive to noise. To preserve the main features of the solution, the large singular values should remain almost unchanged. In order to avoid this unwanted alteration of the low frequency components, we introduce a filtering factor constructed as follows. Looking at Figure 3.8(a) we can observe that most of the entries of $\widehat{x}$ are equal to zero, that is $\widehat{x}$ is sparse. The nonzero elements are located at the four corners, corresponding to the low frequencies (those that we want to preserve). Let us partition the set of indices $\{1, 2, \ldots, n\} = \mathcal{L} \cup \mathcal{H}$, with $\mathcal{L} \cap \mathcal{H} = \emptyset$, where $\mathcal{L}$ and $\mathcal{H}$ contain the

indices associated to the low and high frequencies, respectively. Therefore, we define as $\widehat{x}_{\rm HF}$ the vector with the Fourier coefficients of x associated to the indices in $\mathcal{H}$, and zero elsewhere (i.e. for the indices in $\mathcal{L}$). Exploiting this notation we can introduce a filtering factor Φ in (3.43) in order to obtain an approximation of $\|\widehat{x}_{\rm HF}\|_1$ instead of $\|\widehat{x}\|_1$, that is

$$\|\Phi D_{\widehat{y}}^{-\frac{1}{2}}\widehat{x}\|_2^2 \approx \|\widehat{x}_{\rm HF}\|_1.$$

This can be achieved setting Φ as a diagonal matrix such that

$$(\Phi)_{ii} \approx \begin{cases} 0\,, & i \in \mathcal{L} \\ 1\,, & i \in \mathcal{H} \end{cases}, \qquad i = 1, \ldots, n. \tag{3.44}$$

For $\widehat{x} \in \mathbb{C}^n$, such that $|\widehat{x}_i| > \epsilon$, $i = 1, \ldots, n$, then it holds

$$\|\Phi D_{\widehat{x}}^{-\frac{1}{2}}\widehat{x}\|_2^2 \approx \|\widehat{x}_{\rm HF}\|_1, \tag{3.45}$$

where, if the approximation in (3.44) is an equality, then also those in (3.45) becomes an equality. By setting $\Phi = \Lambda_{L^*L}^{\frac{1}{2}}$, where L is a matrix chosen as in Subsection 3.2.1 ($L \neq I$) for which L^*L can be diagonalized through F as in (3.8), we have precisely that the unwanted modification of the large singular values of A caused by the large coefficients of the image in the Fourier domain (those in the four corners of the 2D array like in Figure 3.8(a) associated with the low frequencies) is removed.

Similarly to what we have done in the spatial domain in which we considered (3.42) in place of (3.38), we can change the exponent in the penalty term of (3.43) and pass from $\|D_{\widehat{y}}^{-\frac{1}{2}}\widehat{x}\|_2^2$ to $\|D_{\widehat{y}}^{-1}\widehat{x}\|_2^2$. This means that we do not approximate $\|\widehat{x}\|_1$, instead we use a norm that avoids penalizing the discontinuities in the Fourier coefficients. Regarding the filtering factor Φ, also in this case the aim is to switch off the regularization for the large singular values of A and preserve the jumps (nonsmoothness) in the nonzero Fourier coefficients of the image (related to low frequencies); then we can use Φ defined as above.

Summarizing, the generalized version of the method (3.43) becomes

$$\min_{\widehat{x}\in\mathbb{C}^n} \left\{ \|AF^*\widehat{x} - b\|_2^2 + \alpha \|\Lambda_{L^*L}^{\frac{1}{2}} D_{\widehat{y}}^{-q}\widehat{x}\|_2^2 \right\}, \tag{3.46}$$

where in the following we consider both cases by using $q = \frac{1}{2}$ or $q = 1$. Again by Remark 17, (3.46) is equivalent to a regularization preconditioning strategy in the Fourier domain with preconditioner $\Lambda_{L^*L}^{\frac{1}{2}} D_{\widehat{y}}^{-q}$.

Actually, rather than (3.9) we consider

$$L^*L = I - \left(\frac{A^*A}{\rho(A)^2} \right)^p, \tag{3.47}$$

since for $0 < p < 1$ it involves a larger set of indices $\mathcal{L}$. Note that for $p = 1$ the matrix (3.47) coincides with (3.9). In Figure 3.9 we show the eigenvalues of L^*L defined as in (3.47) with A as the BCCB matrix associated to the GaussianBlur420 PSF of Example 2 in Subsection 3.4.4. As highlighted in Figure 3.9(a), when $p = 1$ the matrix L^*L behaves almost like the identity matrix and the set $\mathcal{L}$ contains few indices, while looking at Figure 3.9(b) it is clear that for $p = \frac{1}{16}$ the cardinality of $\mathcal{L}$ increases. The optimal choice of the parameter p is not within our aims, so for all the numerical tests of Subsection 3.4.4 we fix $p = \frac{1}{16}$.

When the blurring matrix A can be diagonalized by F like in (3.7), the problem (3.46) can be reformulated as

$$\min_{\widehat{x}\in\mathbb{C}^n} \left\{ \|\Lambda_A\widehat{x} - \widehat{b}\|_2^2 + \alpha \|\Lambda_{L^*L}^{\frac{1}{2}} D_{\widehat{y}}^{-q}\widehat{x}\|_2^2 \right\}, \tag{3.48}$$

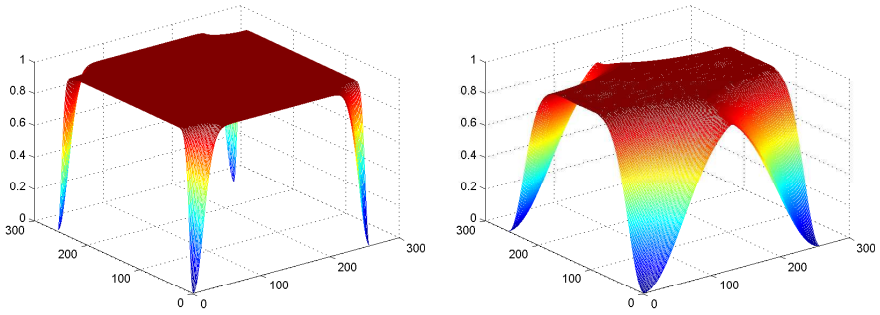
where $\widehat{b} = Fb$. Note that if L is such that $\mathcal{N}(A) \bigcap \mathcal{N}(L) = \{0\}$, like for the generalized Tikhonov method, then $\mathcal{N}(\Lambda_A) \bigcap \mathcal{N}(\Lambda_{L^*L}) = \{0\}$ and hence

$$\mathcal{N}(\Lambda_A) \bigcap \mathcal{N}\left(\Lambda_{L^*L}^{\frac{1}{2}} D_{\widehat{y}}^{-q} \right) = \{0\},$$

which guaranties the well-posedness of (3.48).

We observe that the problem (3.48) is equivalent to the linear system

$$(\Lambda_A^*\Lambda_A + \alpha \Lambda_{L^*L} D_{\widehat{y}}^{-2q})\widehat{x} = \Lambda_A^*\widehat{b},$$



(a) $p = 1$ (b) $p = \frac{1}{16}$

Figure 3.9: Λ_{L^*L} for L^*L as in (3.47)

which is diagonal and therefore very easy to solve in $O(n)$ operations. Furthermore, the GCV functional (3.10) for the problem (3.48) becomes

$$G_{\mathrm{TSF}}(\alpha) = \frac{\|(I - \Lambda_A(\Lambda_A)^{\dagger}_{\mathrm{reg}})\widehat{b}\|_2^2}{(\mathrm{tr}(I - \Lambda_A(\Lambda_A)^{\dagger}_{\mathrm{reg}}))^2},$$

with

$$(\Lambda_A)^{\dagger}_{\mathrm{reg}} = (\Lambda_A^*\Lambda_A + \alpha\Lambda_{L^*L}D_{\widehat{y}}^{-2q})^{-1}\Lambda_A^*.$$

Similarly to the form of the GCV functional in (3.11) for the generalized Tikhonov method, it holds

$$G_{\mathrm{TSF}}(\alpha) = \frac{\sum_{i=1}^{n}(\xi_i\widehat{b}_i)^2}{(\sum_{i=1}^{n}\xi_i)^2}, \tag{3.49}$$

where

$$\xi_i = \frac{\lambda_i(L^*L)/(D_{\widehat{y}})_{ii}^{2q}}{|\lambda_i(A)|^2 + \alpha\lambda_i(L^*L)/(D_{\widehat{y}})_{ii}^{2q}}.$$

Summarizing our TSF algorithm is characterized by the following steps:

Algorithm 3 $\mathbf{x} = \mathbf{TSF}(A, L, b, q, \widehat{y})$

1. Compute Λ_A and Λ_{L^*L} by two FFTs
2. $\widehat{b} = Fb$
3. Compute $D_{\widehat{y}}$ by (3.37)
4. $\alpha_{\mathrm{GCV}} = \min_{\alpha\in\mathbb{R}} G_{\mathrm{TSF}}(\alpha)$, with $G_{\mathrm{TSF}}(\alpha)$ in (3.49)
5. $\widehat{x} = (\Lambda_A^*\Lambda_A + \alpha_{\mathrm{GCV}}\Lambda_{L^*L}D_{\widehat{y}}^{-2q})^{-1}\Lambda_A^*\widehat{b}$
6. $x = F^*\widehat{x}$

Here, we can choose L as in Subsection 3.2.1 or as in (3.47).
Note, that our TSF algorithm requires four FFTs like generalized Tikhonov, the further computational cost is of $O(n)$ operations at the points 3.–5.

TSF with Outer Iterations (ROI)

Following the Regularization with Outer Iterations (ROI) proposed in [93], we can embed the TSF method (3.46) in an outer iteration. In other words, we start an iterative process of building diagonal regularization matrices based on the current reconstruction

$$\begin{cases} \widehat{x}^{(0)} = \widehat{y} \\ \widehat{x}^{(s)} = \underset{\widehat{x}\in\mathbb{C}^n}{\arg\min}\left\{\|AF^*\widehat{x} - b\|_2^2 + \alpha^{(s)}\|\Lambda^{\frac{1}{2}}_{L^*L}D^{-q}_{\widehat{x}^{(s-1)}}\widehat{x}\|_2^2\right\} \quad , \quad s = 1, \dots \ . \end{cases}$$

with $q \in \left\{\frac{1}{2}, 1\right\}$. Clearly, the first iteration of the above scheme coincides with the TSF method. As a consequence of this embedding, we observe further improvement, especially in the first iterations, while in the later steps the error saturates and the reconstruction does not change evidently. Therefore, we can stop the iterations when the relative difference between two consecutive approximations becomes smaller than a fixed tolerance. Furthermore, we add a safe guard control on the residual norm so as to avoid unexpected behaviors in the further iterations due, e.g., to an inaccurate estimation of the regularization parameter.

When the blurring matrix A can be diagonalized by F like in (3.7), the problem (3.46) can be reformulated as (3.48) and hence the ROI algorithm is characterized by the following steps:

Algorithm 4 $\mathbf{x} = \mathbf{ROI}(A, L, b, q, \widehat{x}^{(0)})$

1. Compute Λ_A and Λ_{L^*L} by two FFTs
2. $\widehat{b} = Fb$
3. $\widehat{x}^{(1)} = \mathbf{TSF}(A, L, b, q, \widehat{x}^{(0)})$
4. $r_1 = \|\Lambda_A \widehat{x}^{(1)} - \widehat{b}\|_2$
5. $s = 1$

   Repeat

   Compute $D_{\widehat{x}^{(s)}}$ by (3.37)
   $\alpha_{\mathrm{GCV}}^{(s)} = \min_{\alpha \in \mathbb{R}} G_{\mathrm{TSF}}(\alpha)$, with $G_{\mathrm{TSF}}(\alpha)$ in (3.49)
   $\widehat{x}^{(s+1)} = (\Lambda_A^* \Lambda_A + \alpha_{\mathrm{GCV}}^{(s)} \Lambda_{L^*L} D_{\widehat{x}^{(s)}}^{-2q})^{-1} \Lambda_A^* \widehat{b}$
   $r_{s+1} = \|\Lambda_A \widehat{x}^{(s+1)} - \widehat{b}\|_2$
   $s = s + 1$

   until $\frac{\|\widehat{x}^{(s)} - \widehat{x}^{(s-1)}\|_2}{\|\widehat{x}^{(s-1)}\|_2} < 10^{-2}$ or $r_{s-1} < r_s$
6. $x = F^* \widehat{x}^{(s)}$

Again, the filter L can be chosen as in Subsection 3.2.1 or as in (3.47).

3.4.4 Numerical results

In the following we provide some numerical tests for the image deblurring problem of type (3.1). To test the validity of the reconstruction provided by the TSF and ROI algorithms, we compare the computed solutions with the generalized Tikhonov method.

We consider the following different choices of the regularization matrix L:

1. the identity matrix I;
2. the 2D approximation of the first derivative

$$L_{\mathrm{der}} = L_{n_2}^{(1)} \otimes I_{n_1} + I_{n_2} \otimes L_{n_1}^{(1)}, \tag{3.50}$$

   where

$$L_k^{(1)} = \frac{1}{2} \begin{bmatrix} 1 & -1 & & \\ & \ddots & \ddots & \\ & & 1 & -1 \\ -1 & & & 1 \end{bmatrix}_{k \times k},$$

   is the scaled finite difference approximation of the first derivative with periodic boundary conditions. Note that $\Lambda_{L_{\mathrm{der}}^* L_{\mathrm{der}}}$ can be easily computed by a 2D FFT applied to the stencil

$$\frac{1}{8} \begin{bmatrix} 0 & -1 & 0 \\ -1 & 4 & -1 \\ 0 & -1 & 0 \end{bmatrix}$$

properly padded by zeros to obtain the size $n_1 \times n_2$ of the observed image.

3. the operator dependent matrix (3.47) with $p = \frac{1}{16}$.

Of course, other matrices L could be considered (see [59] and references therein) with the only constraint that L^*L has to be diagonalized by F. To fix the notation of the different methods with different regularization matrices, we refer to Table 3.5, in which the subscripts id, der, op denote the choice of L as in 1., 2., 3., respectively, while the superscript q highlights the choice of the exponent of the diagonal matrices $D_{\widehat{y}}^{-2q}$ and $D_{\widehat{x}^{(s)}}^{-2q}$ in Algorithm 3 and in Algorithm 4, respectively.

Table 3.5: Notation of the different algorithms with different regularization matrices

Method	$L = I$	$L = L_{\text{der}}$	L as in (3.47)
Tikhonov	TIK_{id}	TIK_{der}	TIK_{op}
TSF	TSF^q_{id}	$\text{TSF}^q_{\text{der}}$	TSF^q_{op}
ROI	ROI^q_{id}	$\text{ROI}^q_{\text{der}}$	ROI^q_{op}

Both TSF and ROI algorithms need an initial guess consisting of the Fourier coefficients of an approximation of the true image. In all examples we decide for the Fourier coefficients of the solution provided by TIK_{der} since, as shown in the following, it performs better than the other two Tikhonov methods.

In all considered examples we set $\varepsilon = 10^{-8}$ in (3.37). We have tested several values for ε with different problems, but the results were always comparable, showing that TSF and ROI algorithms are robust with respect to changes of ε. The regularization parameter α is estimated by minimizing the GCV functional (3.11) for Tikhonov and (3.49) for TSF (see Algorithm 3).

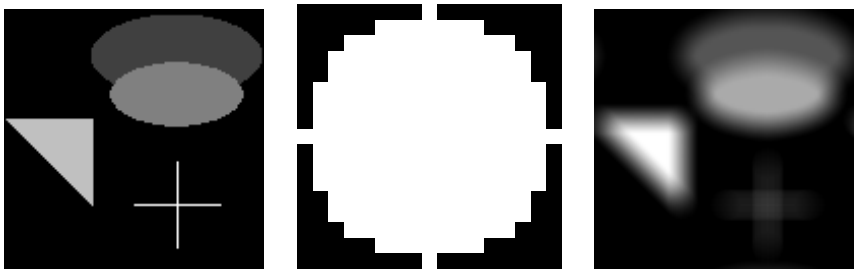
Assuming to know the true image $\bar{x}$, we measure the quality of the reconstruction by computing the Peak Signal-to-Noise Ratio (PSNR) defined as

$$\text{PSNR} := 10 \log_{10} \frac{255^2}{\text{MSE}(\bar{x}, \widetilde{x})},$$

where $\widetilde{x}$ is the current approximation of $\bar{x}$, while $\text{MSE}(\bar{x}, \widetilde{x}) = \frac{\|\bar{x} - \widetilde{x}\|_{\text{F}}}{n^2}$. The higher is the value of the PSNR, the better is the reconstruction $\widetilde{x}$. The numerical tests have been developed with Matlab R2011b on a PC Intel Core$^{\text{TM}}$i5 and Windows 7 operating system.

Example 1

We start with the test problem `blur` taken from [83] with true image of size 128×128 (see Figure 3.10), symmetric out-of-focus PSF ([85]) of size 17×17 and noise level $5 \cdot 10^{-4}$.



(a) (b) (c)

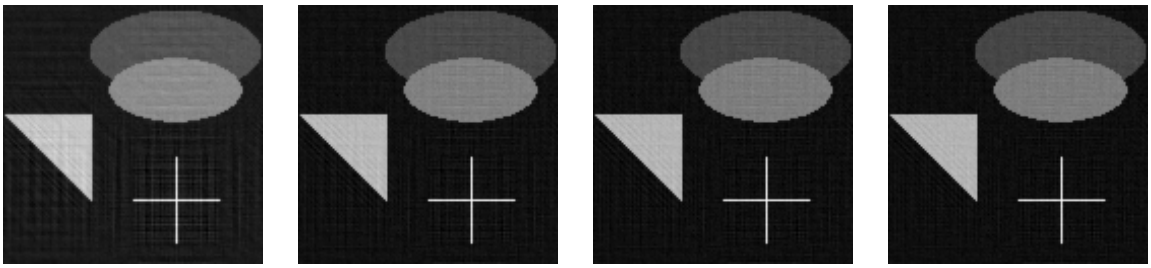
Figure 3.10: Example 1 - (a) true image 128×128; (b) out-of-focus PSF 17×17; (c) blurred and noisy image

Table 3.6: Example 1 - Comparison between PSNRs and regularization parameters for different methods

	β =id		β =der		β =op	
Method	PSNR	α_{GCV}	PSNR	α_{GCV}	PSNR	α_{GCV}
TIK_β	65.512943	6.20e-005	67.432599	6.20e-005	66.642366	6.20e-005
$\text{TSF}^{\frac{1}{2}}_\beta$	68.822621	1.49e-004	69.825451	2.36e-004	68.613248	3.43e-004
TSF^1_β	69.303950	1.60e-003	70.076095	2.36e-003	69.120838	3.79e-003

Table 3.7: Example 1 - PSNRs for the ROI

	$\mathrm{ROI}_{\mathrm{id}}^{\frac{1}{2}}$		$\mathrm{ROI}_{\mathrm{der}}^{\frac{1}{2}}$		$\mathrm{ROI}_{\mathrm{op}}^{\frac{1}{2}}$	
Iter	PSNR	α_{GCV}	PSNR	α_{GCV}	PSNR	α_{GCV}
1	68.822621	1.49e-004	69.825451	2.36e-004	68.613248	3.43e-004
2	69.000796	1.82e-004	70.377213	2.95e-004	68.686781	4.30e-004
3	-	-	70.545766	2.79e-004	-	-



(a) $\mathrm{TIK}_{\mathrm{der}}$ (b) $\mathrm{TSF}_{\mathrm{der}}^{\frac{1}{2}}$ (c) $\mathrm{TSF}_{\mathrm{der}}^{1}$ (d) $\mathrm{ROI}_{\mathrm{der}}^{\frac{1}{2}}$

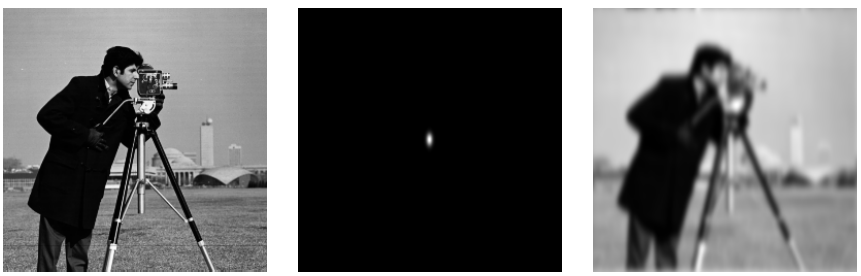
Figure 3.11: Example 1 - restored images

In Table 3.6 we compare the PSNRs for all methods described in Table 3.5. Let us observe that $\mathrm{TIK}_{\mathrm{der}}$ is more accurate than $\mathrm{TIK}_{\mathrm{id}}$ or $\mathrm{TIK}_{\mathrm{op}}$; this justifies our choice of the initial guess as the Fourier coefficients of the solution computed by that Tikhonov method. Regarding our TSF algorithm, it provides a good improvement of the Tikhonov solution for all choices of the regularization matrix (compare, e.g., Figure 3.11(a) with Figures 3.11(b)3.11(c)). Note that also when $q = 1$, that is when in the penalty term the diagonal matrix $D_{\widehat{y}}^{-\frac{1}{2}}$ is replaced by $D_{\widehat{y}}^{-1}$, the TSF method works well and the corresponding PSNR is larger than the one in Tikhonov.

In Table 3.7 we show the effect of embedding the TSF method in an outer iteration. The ROIs that stop at first iteration are not reported since they coincide with the TSF algorithm. From Table 3.7 we deduce that $\mathrm{TSF}_{\mathrm{id}}^{1}$, $\mathrm{TSF}_{\mathrm{der}}^{1}$ and $\mathrm{TSF}_{\mathrm{op}}^{1}$ do not benefit of such an embedding, while $\mathrm{ROI}_{\mathrm{id}}^{\frac{1}{2}}$, $\mathrm{ROI}_{\mathrm{der}}^{\frac{1}{2}}$, $\mathrm{ROI}_{\mathrm{op}}^{\frac{1}{2}}$ improve the quality of the corresponding TSF reconstruction performing 2/3 iterations. Let us observe that $\mathrm{TSF}_{\mathrm{der}}^{\frac{1}{2}}$ provides a PSNR which is larger than the PSNRs corresponding to the last iteration of $\mathrm{ROI}_{\mathrm{id}}^{\frac{1}{2}}$ and $\mathrm{ROI}_{\mathrm{op}}^{\frac{1}{2}}$ and that the best reconstruction is given by $\mathrm{ROI}_{\mathrm{der}}^{\frac{1}{2}}$ (compare, e.g., Figures 3.11(a)-(c) with Figure 3.11(d)). This could be ascribed to the fact that for $q = \frac{1}{2}$ the penalty term is approximating the 1-norm of the components $\widehat{x}_{\mathrm{HF}}$ of $\widehat{x}$ in the high frequencies, that is $\|\Lambda_{L_{\mathrm{der}}^* L_{\mathrm{der}}}^{\frac{1}{2}} D_{\widehat{x}^{(s)}}^{-\frac{1}{2}} \widehat{x}\|_2^2 \approx \|\widehat{x}_{\mathrm{HF}}\|_1$.

Example 2

We consider the cameraman deblurring problem of size 256×256 in Figure 3.12. The nonsymmetric GaussianBlur420 PSF of size 256×256 comes from Restore Tools Matlab package ([106]) and the noise level is 10^{-5}.



(a) (b) (c)

Figure 3.12: Example 2 - (a) true image 256×256; (b) GaussianBlur420 PSF 256×256; (c) blurred and noisy image

(a) $\mathrm{TIK}_{\mathrm{der}}$ (b) $\mathrm{TSF}^{1}_{\mathrm{op}}$ (c) $\mathrm{ROI}^{\frac{1}{2}}_{\mathrm{der}}$ (d) $\mathrm{ROI}^{1}_{\mathrm{op}}$

Figure 3.13: Example 2 - restored images

Table 3.8: Example 2 - Comparison between PSNRs and regularization parameters for different methods

	β =id		β =der		β =op	
Method	PSNR	α_{GCV}	PSNR	α_{GCV}	PSNR	α_{GCV}
TIK_{β}	24.798034	5.36e-005	25.192810	5.36e-005	24.958256	5.36e-005
$\mathrm{TSF}^{\frac{1}{2}}_{\beta}$	25.942006	5.36e-005	26.108259	5.36e-005	25.994444	5.36e-005
TSF^{1}_{β}	26.288431	5.36e-005	26.387301	5.36e-005	26.315813	5.36e-005

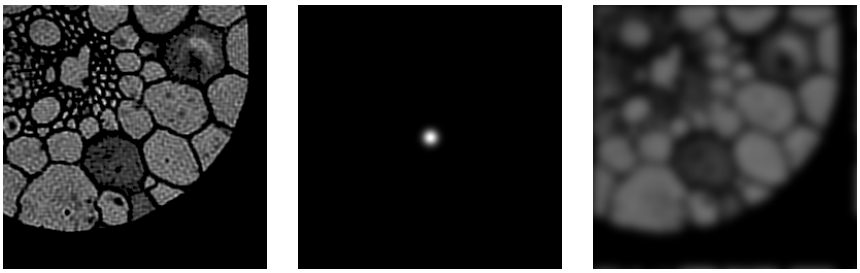
Table 3.9: Example 2 - PSNRs for the ROI

	$\mathrm{ROI}^{\frac{1}{2}}_{\mathrm{id}}$		$\mathrm{ROI}^{\frac{1}{2}}_{\mathrm{der}}$		$\mathrm{ROI}^{\frac{1}{2}}_{\mathrm{op}}$	
Iter	PSNR	α_{GCV}	PSNR	α_{GCV}	PSNR	α_{GCV}
1	25.942006	5.36e-005	26.108259	5.36e-005	25.994444	5.36e-005
2	26.187353	5.36e-005	26.385167	5.36e-005	26.244349	5.36e-005
3	26.299611	5.36e-005	26.521393	5.36e-005	26.361031	5.36e-005
4	-	-	26.591178	5.36e-005	-	-
	$\mathrm{ROI}^{1}_{\mathrm{id}}$		$\mathrm{ROI}^{1}_{\mathrm{der}}$		$\mathrm{ROI}^{1}_{\mathrm{op}}$	
Iter	PSNR	α_{GCV}	PSNR	α_{GCV}	PSNR	α_{GCV}
1	26.288431	5.36e-005	26.387301	5.36e-005	26.315813	5.36e-005
2	26.568309	5.36e-005	26.645156	5.36e-005	26.590811	5.36e-005
3	26.634729	5.36e-005	-	-	26.594446	1.03e-001

As shown in Table 3.8, also for this example the TSF method enhances the quality of reconstruction of the $\mathrm{TIK}_{\mathrm{der}}$ method independently of the choice of the regularization matrix (see, e.g., Figure 3.13(a) in comparison with Figure 3.13(b)). In particular, the $\mathrm{TSF}^{1}_{\mathrm{id}}$, $\mathrm{TSF}^{1}_{\mathrm{der}}$ and $\mathrm{TSF}^{1}_{\mathrm{op}}$ provide good reconstructions and the additional ROIs entail a further improvement (refer to Table 3.9 and compare Figure 3.13(b) with Figure 3.13(d)). Note that the noise level for this example is very low and that the PSF is a Gaussian one. Moreover, let us observe that embedding the $\mathrm{TSF}^{\frac{1}{2}}_{\mathrm{der}}$ in a ROI gives rise to 4 reconstruction-improving iterations (see Figure 3.13(c)) and the PSNR obtained at last iteration is comparable with the ones corresponding to $\mathrm{ROI}^{1}_{\mathrm{id}}$, $\mathrm{ROI}^{1}_{\mathrm{der}}$ and $\mathrm{ROI}^{1}_{\mathrm{op}}$.

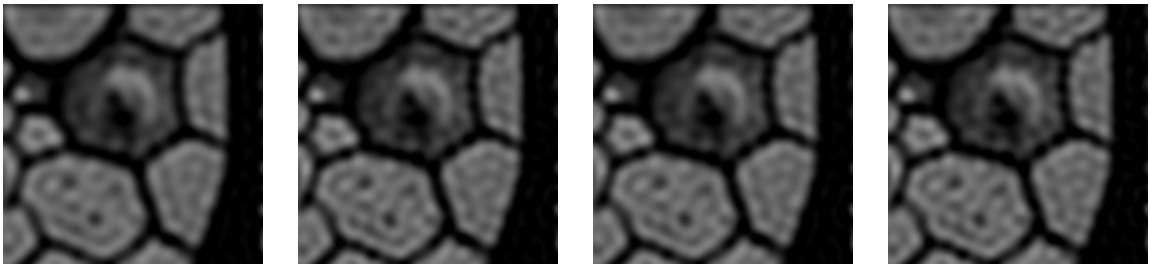
Example 3

In this example we consider the grain image of size 256×256 using a Gaussian PSF with standard deviation equal to 5 and setting the noise level to 10^{-3} (see Figure 3.14).



(a) (b) (c)

Figure 3.14: Example 3 - (a) true image 256×256; (b) Gaussian PSF 256×256; (c) blurred and noisy image



(a) $\mathrm{TIK}_{\mathrm{der}}$ (b) $\mathrm{TSF}^{\frac{1}{2}}_{\mathrm{der}}$ (c) $\mathrm{TSF}^{\frac{1}{2}}_{\mathrm{op}}$ (d) $\mathrm{ROI}^{\frac{1}{2}}_{\mathrm{op}}$

Figure 3.15: Example 3 - restored images

Table 3.10: Example 3 - Comparison between PSNRs and regularization parameters for different methods

	β =id		β =der		β =op	
Method	PSNR	α_{GCV}	PSNR	α_{GCV}	PSNR	α_{GCV}
TIK_{β}	68.939829	5.36e-005	69.660417	5.36e-005	69.137066	5.36e-005
$\mathrm{TSF}^{\frac{1}{2}}_{\beta}$	69.755546	5.36e-005	69.874618	1.94e-004	69.808882	5.36e-005
TSF^{1}_{β}	69.840102	2.91e-004	69.836100	4.44e-003	69.836237	5.24e-004

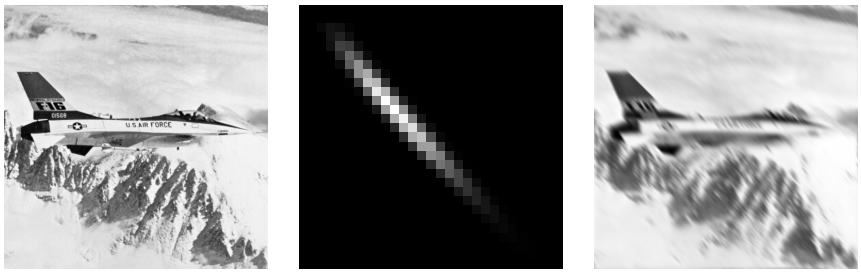
Table 3.11: Example 3 - PSNRs for the ROI

	$\mathrm{ROI}^{\frac{1}{2}}_{\mathrm{id}}$		$\mathrm{ROI}^{\frac{1}{2}}_{\mathrm{op}}$	
Iter	PSNR	α_{GCV}	PSNR	α_{GCV}
1	69.755546	5.36e-005	69.808882	5.36e-005
2	69.778181	5.36e-005	69.840942	5.36e-005
3	69.787079	5.36e-005	69.849724	5.36e-005
4	-	-	69.852618	5.36e-005

Table 3.10 gives evidence of a better accuracy of the TSF method with respect to the $\mathrm{TIK}_{\mathrm{der}}$, especially when the penalty term is given by $\|\Lambda^{\frac{1}{2}}_{L^*_{\mathrm{der}}L_{\mathrm{der}}} D^{-\frac{1}{2}}_{\widehat{y}}\widehat{x}\|^2_2$, that is in the case $\mathrm{TSF}^{\frac{1}{2}}_{\mathrm{der}}$. Furthermore, as shown in Table 3.11, $\mathrm{ROI}^{\frac{1}{2}}_{\mathrm{id}}$ and $\mathrm{ROI}^{\frac{1}{2}}_{\mathrm{op}}$ perform 3 and 4 iterations, respectively, improving the quality of the corresponding TSF reconstruction. In Figure 3.15 we focus on a detail of the restored images which highlights how the restorations provided by the TSF and the ROI are richer of details than the one obtained with $\mathrm{TIK}_{\mathrm{der}}$.

Example 4

As a final example, we consider the jetplane deblurring problem of size 512×512 in Figure 3.16 taking as PSF the nonsymmetric diagonal motion of size 29×29 and noise level 10^{-4}.



(a) (b) (c)

Figure 3.16: Example 4 - (a) true image 512×512; (b) PSF 29×29 with center in pixel (12,11); (c) blurred and noisy image

(a) $\mathrm{TIK}_{\mathrm{der}}$ (b) $\mathrm{TSF}^{1}_{\mathrm{id}}$ (c) $\mathrm{TSF}^{1}_{\mathrm{der}}$ (d) $\mathrm{ROI}^{\frac{1}{2}}_{\mathrm{der}}$

Figure 3.17: Example 4 - restored images

Table 3.12: Example 4 - Comparison between PSNRs and regularization parameters for different methods

	β =id		β =der		β =op	
Method	PSNR	α_{GCV}	PSNR	α_{GCV}	PSNR	α_{GCV}
TIK_{β}	30.123900	5.52e-005	32.523969	5.52e-005	30.882375	5.52e-005
$\mathrm{TSF}^{\frac{1}{2}}_{\beta}$	33.861210	7.78e-003	34.230730	2.76e-002	33.811442	1.49e-002
TSF^{1}_{β}	33.941120	1.47e+001	34.210584	3.81e+001	33.897847	2.87e+001

Table 3.13: Example 4 - PSNRs for the ROI

	$\mathrm{ROI}^{\frac{1}{2}}_{\mathrm{id}}$		$\mathrm{ROI}^{\frac{1}{2}}_{\mathrm{der}}$		$\mathrm{ROI}^{\frac{1}{2}}_{\mathrm{op}}$	
Iter	PSNR	α_{GCV}	PSNR	α_{GCV}	PSNR	α_{GCV}
1	33.861210	7.78e-003	34.230730	2.76e-002	33.811442	1.49e-003
2	33.823830	1.41e-002	34.353391	4.44e-002	33.755098	2.75e-002

As in Example 3, in order to point out the effectiveness of the TSF method, we focus on a detail of the blurred jetplane image. Note that the number 01568 on the tail of the jetplane is definitely more readable in Figures 3.17(b)–3.17(d) than in Figure 3.17(a). This behavior is confirmed by the PSNRs shown in Table 3.12. In particular, we deduce that, also for this example, a good choice of the penalty term is the 2-norm approximation of $\|\widehat{x}_{\mathrm{HF}}\|_1$ provided by the discretization of the first derivative (3.50). Furthermore, as shown in Table 3.13, this is the only case in which the ROI gives a slight improvement performing 2 iterations (look at Figure 3.17(d)), while in the other cases the second iteration fails in enhancing the approximation obtained with the TSF method at first iteration.

Chapter 4

Spectral analysis and structure preserving preconditioners for FDEs

In this chapter we focus on a discretization of the fractional partial order diffusion equations which leads to a linear system whose coefficient matrix has a Toeplitz-like structure (see Section 4.1). In particular, we focus our attention on the case of variable diffusion coefficients. Under appropriate conditions, in Section 4.2 we show that the sequence of the coefficient matrices belongs to the GLT class and, using the tools introduced in Chapter 1, we compute the symbol describing its asymptotic eigenvalue/singular value distribution, as the matrix size diverges. In Section 4.3 we employ the spectral information for analyzing known methods of preconditioned Krylov and multigrid type, with both positive and negative results and with a look forward to the multidimensional setting. We also propose two new tridiagonal structure preserving preconditioners to solve the resulting linear system, with Krylov methods such as CGLS and GMRES. Due to their tridiagonal structure, both preconditioners preserve the computational cost per iteration of the used Krylov method. A clustering analysis of the preconditioned matrix-sequences, even in case of nonconstant diffusion coefficients, is also provided. Finally, in Section 4.4 we give a number of numerical examples which shows that our proposal is more effective than recently used circulant preconditioners.

4.1 Problem setting: FDEs

Fractional-space diffusion equations (FDEs) are used to describe diffusion phenomena, that cannot be modeled by the second order diffusion equations. More precisely, when a fractional derivative replaces a second derivative in a diffusion model, it leads to enhanced diffusion. FDEs arise in many research topics including image processing [8] and turbulent flow [31, 138]. In [103, 104] Meerschaert and Tadjeran introduced an unconditionally stable method for approximating the FDEs. As we show in a moment, the resulting linear systems have a strong structure and indeed the related coefficient matrices can be seen as a sum of two diagonal times Toeplitz matrices (see [150]).

Throughout this chapter we denote the dimension of the involved matrices by N, in accordance with the notation used in the recent works on the FDEs (e.g. [151, 113, 100, 112]).

Let us consider the following initial-boundary value FDE problem

$$\begin{cases} \frac{\partial u(x,t)}{\partial t} = d_+(x,t)\frac{\partial^\alpha u(x,t)}{\partial_+ x^\alpha} + d_-(x,t)\frac{\partial^\alpha u(x,t)}{\partial_- x^\alpha} + f(x,t), & (x,t) \in (L,R)\times(0,T], \\ u(L,t) = u(R,t) = 0, & t \in [0,T], \\ u(x,0) = u_0(x), & x \in [L,R], \end{cases} \tag{4.1}$$

where $\alpha \in (1,2)$ is the *fractional derivative order*, $f(x,t)$ is the *source term* and the nonnegative functions $d_\pm(x,t)$ are the *diffusion coefficients*. The right-handed (–) and the left-handed (+) fractional derivatives in (4.1) are defined in Riemann-Liouville form as follows

$$\frac{\partial^\alpha u(x,t)}{\partial_+ x^\alpha} = \frac{1}{\Gamma(n-\alpha)}\frac{\partial^n}{\partial x^n}\int_L^x \frac{u(\xi,t)}{(x-\xi)^{\alpha+1-n}}d\xi,$$

$$\frac{\partial^\alpha u(x,t)}{\partial_- x^\alpha} = \frac{(-1)^n}{\Gamma(n-\alpha)}\frac{\partial^n}{\partial x^n}\int_x^R \frac{u(\xi,t)}{(\xi-x)^{\alpha+1-n}}d\xi,$$

where n is an integer such that $n-1 < \alpha \leq n$ and $\Gamma(\cdot)$ is the gamma function. If $\alpha = m$, with $m \in \mathbb{N}$, the fractional derivatives reduce to the standard integer derivatives, i.e.,

$$\frac{\partial^m u(x,t)}{\partial_+ x^m} = \frac{\partial^m u(x,t)}{\partial x^m}, \quad \frac{\partial^m u(x,t)}{\partial_- x^m} = (-1)^m \frac{\partial^m u(x,t)}{\partial x^m}.$$

Let us observe that when $\alpha = 2$ the equation in (4.1) reduces to a parabolic PDE, while when $\alpha = 1$ it becomes a hyperbolic PDE. From a numerical point of view, an interesting definition of the fractional derivatives is the shifted Grünwald definition given by

$$\begin{aligned} \frac{\partial^\alpha u(x,t)}{\partial_+ x^\alpha} &= \lim_{\Delta x \to 0^+} \frac{1}{\Delta x^\alpha} \sum_{k=0}^{\lfloor (x-L)/\Delta x \rfloor} g_k^{(\alpha)} u(x-(k-1)\Delta x, t), \\ \frac{\partial^\alpha u(x,t)}{\partial_- x^\alpha} &= \lim_{\Delta x \to 0^+} \frac{1}{\Delta x^\alpha} \sum_{k=0}^{\lfloor (R-x)/\Delta x \rfloor} g_k^{(\alpha)} u(x+(k-1)\Delta x, t), \end{aligned} \tag{4.2}$$

where $\lfloor \cdot \rfloor$ is the floor function, while $g_k^{(\alpha)}$ are the *alternating fractional binomial coefficients* defined as

$$g_k^{(\alpha)} = (-1)^k \binom{\alpha}{k} = \frac{(-1)^k}{k!} \alpha(\alpha-1)\cdots(\alpha-k+1) \quad k = 0, 1, \ldots \tag{4.3}$$

with the formal notation $\binom{\alpha}{0} = 1$. The shifted Grünwald formulas are numerically relevant since, from (4.2), we can define the following estimates of the left and right-handed fractional derivatives

$$\begin{aligned} \frac{\partial^\alpha u(x,t)}{\partial_+ x^\alpha} &= \frac{1}{\Delta x^\alpha} \sum_{k=0}^{\lfloor (x-L)/\Delta x \rfloor} g_k^{(\alpha)} u(x-(k-1)\Delta x, t) + O(\Delta x), \\ \frac{\partial^\alpha u(x,t)}{\partial_- x^\alpha} &= \frac{1}{\Delta x^\alpha} \sum_{k=0}^{\lfloor (R-x)/\Delta x \rfloor} g_k^{(\alpha)} u(x+(k-1)\Delta x, t) + O(\Delta x). \end{aligned}$$

In [103] Meerschaert and Tadjeran proved that the implicit Euler method based on the shifted Grünwald formula is consistent and unconditionally stable. Let us fix two positive integers N, M, and define the following partition of $[L, R] \times [0, T]$, i.e.,

$$\begin{gathered} x_i = L + i\Delta t, \quad \Delta x = \tfrac{(R-L)}{N+1}, \quad i = 0, \ldots, N+1, \\ t_m = m\Delta t, \quad \Delta t = \tfrac{T}{M}, \quad m = 0, \ldots, M. \end{gathered}$$

More in detail, the idea that underlies the Meerschaert-Tadjeran method is to combine a discretization in time of equation (4.1) by an implicit Euler method, with a discretization in space of the fractional derivatives by a shifted Grünwald estimate, i.e.,

$$\frac{u(x_i,t_m) - u(x_i,t_{m-1})}{\Delta t} = d_{+,i}^{(m)} \frac{\partial^\alpha u(x_i,t_m)}{\partial_+ x^\alpha} + d_{-,i}^{(m)} \frac{\partial^\alpha u(x_i,t_m)}{\partial_- x^\alpha} + f_i^{(m)} + O(\Delta t),$$

where $d_{\pm,i}^{(m)} := d_\pm(x_i,t_m)$, $f_i^{(m)} := f(x_i,t_m)$ and

$$\begin{aligned} \frac{\partial^\alpha u(x_i,t_m)}{\partial_+ x^\alpha} &= \frac{1}{\Delta x^\alpha} \sum_{k=0}^{i+1} g_k^{(\alpha)} u(x_{i-k+1}, t_m) + O(\Delta x), \\ \frac{\partial^\alpha u(x_i,t_m)}{\partial_- x^\alpha} &= \frac{1}{\Delta x^\alpha} \sum_{k=0}^{N-i+2} g_k^{(\alpha)} u(x_{i+k-1}, t_m) + O(\Delta x). \end{aligned}$$

The resulting finite difference approximation scheme is then

$$\frac{u_i^{(m)} - u_i^{(m-1)}}{\Delta t} = \frac{d_{+,i}^{(m)}}{\Delta x^\alpha} \sum_{k=0}^{i+1} g_k^{(\alpha)} u_{i-k+1}^{(m)} + \frac{d_{-,i}^{(m)}}{\Delta x^\alpha} \sum_{k=0}^{N-i+2} g_k^{(\alpha)} u_{i+k-1}^{(m)} + f_i^{(m)},$$

where by $u_i^{(m)}$ we denote a numerical approximation of $u(x_i,t_m)$. The previous approximation scheme can be written in matrix form as (see [150])

$$\left(\nu_{M,N} I + D_+^{(m)} T_{\alpha,N} + D_-^{(m)} T_{\alpha,N}^T\right) u^{(m)} = \nu_{M,N} u^{(m-1)} + \Delta x^\alpha f^{(m)}, \tag{4.4}$$

where $\nu_{M,N} = \frac{\Delta x^\alpha}{\Delta t}$, $u^{(m)} = [u_1^{(m)}, \dots, u_N^{(m)}]^T$, $f^{(m)} = [f_1^{(m)}, \dots, f_N^{(m)}]^T$, $D_\pm^{(m)} = \mathrm{diag}(d_{\pm,1}^{(m)}, \dots, d_{\pm,N}^{(m)})$, I is the identity matrix of order N and

$$T_{\alpha,N} = -\begin{bmatrix} g_1^{(\alpha)} & g_0^{(\alpha)} & 0 & \cdots & 0 & 0 \\ g_2^{(\alpha)} & g_1^{(\alpha)} & g_0^{(\alpha)} & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ g_{N-1}^{(\alpha)} & \ddots & \ddots & \ddots & g_1^{(\alpha)} & g_0^{(\alpha)} \\ g_N^{(\alpha)} & g_{N-1}^{(\alpha)} & \cdots & \cdots & g_2^{(\alpha)} & g_1^{(\alpha)} \end{bmatrix}_{N\times N}$$

is a lower Hessenberg Toeplitz matrix. The fractional binomial coefficients $g_k^{(\alpha)}$ satisfy few properties, summarized in the following proposition (see [103, 104, 150]).

Proposition 13. *Let $\alpha \in (1,2)$ and $g_k^{(\alpha)}$ be defined as in (4.3). Then we have*

$$\begin{cases} g_0^{(\alpha)} = 1, & g_1^{(\alpha)} = -\alpha, & g_0^{(\alpha)} > g_2^{(\alpha)} > g_3^{(\alpha)} > \ldots > 0, \\ \sum_{k=0}^{\infty} g_k^{(\alpha)} = 0, & \sum_{k=0}^{n} g_k^{(\alpha)} < 0, & n \geq 1. \end{cases}$$

From here onwards, we denote the coefficient matrix of the linear system (4.4) by $\mathcal{M}_{\alpha,N}^{(m)}$, that is

$$\mathcal{M}_{\alpha,N}^{(m)} = \nu_{M,N} I + D_+^{(m)} T_{\alpha,N} + D_-^{(m)} T_{\alpha,N}^T. \tag{4.5}$$

Using Proposition 13, it can be shown that $\mathcal{M}_{\alpha,N}^{(m)}$ is strictly diagonally dominant and then nonsingular (see [150]), for every choice of the parameters $m \geq 0$, $N \geq 1$, $\alpha \in (1,2)$.

The FDEs are of numerical interest, since there exist only few cases in which the analytic solution is known. As a consequence, in the past ten years, many methods have been proposed for solving numerically FDEs problems. Exploiting the structure of $\mathcal{M}_{\alpha,N}^{(m)}$, in [151] the authors employed CGLS and numerically showed that its convergence is fast when the diffusion coefficients are small, that is in this case the resulting linear system is well-conditioned. On the other hand, when the diffusion coefficients are not small, the problem becomes ill-conditioned and the convergence of the CGLS method slows down. To avoid the resulting drawback, in [113] Pang and Sun proposed a multigrid method that converges very fast, even in the ill-conditioned case. The linear convergence of such a method has been proved only in the case of constant and equal diffusion coefficients. With the same purpose, Lei and Sun used the CGLS method with a circulant preconditioner and verified that it converges superlinearly (see [100]), again in the case of constant diffusion coefficients. A further improvement of the circulant preconditioning has been proposed in [112]. Both strategies preserve the computational cost per iteration of $O(N \log N)$ operations, typical of the CGLS method when applied to Toeplitz type structures.

4.2 Spectral analysis of the coefficient matrix

Recalling the notion of symbol and of spectral distribution in the eigenvalue and singular value sense given in Chapter 1, in this section we provide a spectral analysis of the coefficient matrix-sequence $\left\{\mathcal{M}_{\alpha,N}^{(m)}\right\}_N$. In the constant coefficient case, as already observed in papers like [100], the coefficient matrix-sequence is a Toeplitz sequence: then using well-known spectral tools for Toeplitz sequences we determine its symbol and study its spectral distribution. In the nonconstant coefficients case, under appropriate conditions, we show that, $\left\{\mathcal{M}_{\alpha,N}^{(m)}\right\}_N$ belongs to the GLT class and use the GLT machinery to analyze its singular value/eigenvalue distribution. The resulting spectral information is then used in Section 4.3 for the analysis and the design of numerical solvers to be applied to the considered linear systems.

4.2.1 Constant diffusion coefficients case

Let us assume that both diffusion coefficients are constant. Under this condition, $\left\{\mathcal{M}_{\alpha,N}^{(m)}\right\}_N$ is a sequence of Toeplitz matrices. For convenience, we rewrite Definition 7 in the case $k = s = 1$, obtaining the definition of unilevel Toeplitz sequence and of corresponding symbol.

Definition 17. Let $f \in L^1(I_1)$ and let $\{f_j\}_{j\in\mathbb{Z}}$ be the sequence of its Fourier coefficients defined as

$$f_j = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(\theta)\mathrm{e}^{-\mathrm{i}j\theta}d\theta, \quad j \in \mathbb{Z}.$$

Then the Toeplitz sequence $\{T_N\}_N$ with $T_N = [f_{i-j}]_{i,j=1}^N$ is called the family of Toeplitz matrices generated by f, which in turn is called the symbol of $\{T_N\}_N$ and T_N is denoted by $T_N(f)$.

Remark 21. Let $\{T_N\}_N$ be a Toeplitz sequence, with $T_N = [f_{i-j}]_{i,j=1}^N$. If $\{f_j\}_{j\in\mathbb{Z}}$ is such that $\sum_{k=-\infty}^{\infty}|f_k| < \infty$, then the series $\sum_{k=-\infty}^{\infty} f_k\mathrm{e}^{\mathrm{i}k\theta}$ uniformly converges to a continuous and 2π–periodic function f which belongs to the Wiener class (see Definition 13) and is the symbol of $\{T_N\}_N$, i.e., $T_N = T_N(f)$, $\forall N \in \mathbb{N}$.

We determine the sequence of symbols associated to $\left\{\mathcal{M}_{\alpha,N}^{(m)}\right\}_N$ as a corollary of the following proposition.

Proposition 14. *Let $\alpha \in (1,2)$. The symbol associated to the matrix-sequence $\{T_{\alpha,N}\}_N$ belongs to the Wiener class and its formal expression is given by*

$$f_\alpha(\theta) = -\sum_{k=-1}^{\infty} g_{k+1}^{(\alpha)}\mathrm{e}^{\mathbf{i}k\theta} = -\mathrm{e}^{-\mathbf{i}\theta}\left(1+\mathrm{e}^{\mathbf{i}(\theta+\pi)}\right)^{\alpha}. \tag{4.6}$$

Proof. Let us observe that $T_{\alpha,N} = [-g_{i-j+1}^{(\alpha)}]_{i,j=1}^N$ with $g_k^{(\alpha)} = 0$ for $k < 0$ and let us define the function $f_\alpha(\theta) = -\sum_{k=-1}^{\infty} g_{k+1}^{(\alpha)}\mathrm{e}^{\mathbf{i}k\theta}$. When $\alpha \in (1,2)$, it is easy to see that $f_\alpha(\theta)$ lies in the Wiener class. In detail, from Proposition 13 we know that $g_1^{(\alpha)} = -\alpha < 0$, $g_k^{(\alpha)} > 0$ for $k \geq 0$ and $k \neq 1$, and $g_k^{(\alpha)} = 0$ for $k < 0$. Then

$$\sum_{k=-1}^{\infty}|g_{k+1}^{(\alpha)}| = \sum_{\substack{k=-1\\k\neq 0}}^{\infty} g_{k+1}^{(\alpha)} + \alpha.$$

Again from Proposition 13 we deduce

$$\sum_{k=0}^{\infty} g_k^{(\alpha)} = 0 \iff \sum_{\substack{k=-1\\k\neq 0}}^{\infty} g_{k+1}^{(\alpha)} = -g_1^{(\alpha)} = \alpha,$$

that is $\sum_{k=-1}^{\infty}|g_{k+1}^{(\alpha)}| = 2\alpha$, which means that $f_\alpha(\theta)$ belongs to the Wiener class for $\alpha \in (1,2)$. To obtain an explicit formula for the symbol $f_\alpha(\theta)$, let us recall the definition of $g_k^{(\alpha)}$ given in (4.3) and let us rewrite $f_\alpha(\theta)$ as follows

$$\begin{aligned} f_\alpha(\theta) &= -\sum_{k=0}^{\infty} g_k^{(\alpha)}\mathrm{e}^{\mathbf{i}(k-1)\theta} = -\sum_{k=0}^{\infty}(-1)^k\binom{\alpha}{k}\mathrm{e}^{\mathbf{i}(k-1)\theta} \\ &= -\sum_{k=0}^{\infty}\binom{\alpha}{k}\mathrm{e}^{\mathbf{i}(k-1)\theta}\mathrm{e}^{\mathbf{i}k\pi} = -\mathrm{e}^{-\mathbf{i}\theta}\sum_{k=0}^{\infty}\binom{\alpha}{k}\mathrm{e}^{\mathbf{i}k(\theta+\pi)}. \end{aligned}$$

Applying the well-known binomial series

$$(1+z)^\alpha = \sum_{k=0}^{\infty}\binom{\alpha}{k}z^k, \quad z \in \mathbb{C}, \quad |z| \leq 1, \quad \alpha > 0,$$

with $z = \mathrm{e}^{\mathbf{i}(\theta+\pi)}$ we obtain

$$f_\alpha(\theta) = -\mathrm{e}^{-\mathbf{i}\theta}\left(1+\mathrm{e}^{\mathbf{i}(\theta+\pi)}\right)^{\alpha}.$$

□

Corollary 1. *Let us assume that $d_+(x,t) = d_+ > 0$, $d_-(x,t) = d_- > 0$. The matrix $\mathcal{M}_{\alpha,N}^{(m)}$ defined as in (4.5) is the Toeplitz matrix $\mathcal{M}_{\alpha,N}^{(m)} = [\varphi_{i-j}]_{i,j=1}^N$ with*

$$\varphi_j = \frac{1}{2\pi}\int_{-\pi}^{\pi}\varphi_{\alpha,N}(\theta)\mathrm{e}^{-\mathrm{i}j\theta}d\theta, \quad j \in \mathbb{Z},$$

where

$$\varphi_{\alpha,N}(\theta) = \nu_{M,N} + d_+ f_\alpha(\theta) + d_- f_\alpha(-\theta).$$

Now we focus our attention on the spectral distribution of $\left\{\mathcal{M}_{\alpha,N}^{(m)}\right\}_N$, under the further assumption that the diffusion coefficients are equal. By this hypothesis, $\left\{\mathcal{M}_{\alpha,N}^{(m)}\right\}_N$ is a sequence of symmetric Toeplitz matrices. Let us recall Definition 3 with $s = 1$, $d_n = n$ and $n = N$. Note that, under these conditions, 'tr' in the right-hand side of (1.11) and (1.12) disappears. The following proposition concerns the eigenvalue distribution of the coefficient matrix-sequence $\left\{\mathcal{M}_{\alpha,N}^{(m)}\right\}_N$, when diffusion coefficients are constant and equal.

Proposition 15. *Let us assume that $d_\pm(x,t) = d > 0$ and that $\nu_{M,N} = o(1)$. Given the matrix-sequence $\left\{\mathcal{M}_{\alpha,N}^{(m)}\right\}_N$ with $\mathcal{M}_{\alpha,N}^{(m)}$ defined as in (4.5), we have*

$$\left\{\mathcal{M}_{\alpha,N}^{(m)}\right\}_N \sim_\lambda (d \cdot p_\alpha(\theta), I_1),$$

where

$$p_\alpha(\theta) = f_\alpha(\theta) + f_\alpha(-\theta) = f_\alpha(\theta) + \overline{f_\alpha(\theta)} \tag{4.7}$$

is a real-valued continuous function.

Proof. Since the diffusion coefficients $d_\pm(x,t)$ are constant and equal to a real positive number d, the matrices of the sequence $\left\{dT_{\alpha,N} + dT_{\alpha,N}^T\right\}_N$ are symmetric. The function $p_\alpha(\theta) = f_\alpha(\theta) + f_\alpha(-\theta) = f_\alpha(\theta) + \overline{f_\alpha(\theta)}$ belongs to the Wiener algebra since $f_\alpha(\theta)$ itself is in the same algebra (see Proposition 14). Furthermore, from its expression it also follows that $p_\alpha(\theta)$ is real-valued and globally continuous.

From Theorem 5, it follows that $\left\{dT_{\alpha,N} + dT_{\alpha,N}^T\right\}_N \sim_\lambda (d \cdot p_\alpha, I_1)$. Furthermore, using (1.18) ($k = s = 1$), we have that $\|dT_{\alpha,N} + dT_{\alpha,N}^T\| \le d\|p_\alpha\|_{L^\infty} = d2^{\alpha+1}$, while under the hypothesis that $\nu_{M,N} = o(1)$, the remaining term $\nu_{M,N} I$ is such that $\|\nu_{M,N} I\|_1 = o(N)$ and $\|\nu_{M,N} I\| = \nu_{N,M} < C$ for some constant C independent of N. By Theorem 2, we conclude that the distribution of $\left\{\mathcal{M}_{\alpha,N}^{(m)}\right\}_N$ is decided only by $d \cdot p_\alpha(\theta)$. □

Combining (4.7) with (4.6), we can explicitly rewrite $p_\alpha(\theta)$ as follows

$$p_\alpha(\theta) = f_\alpha(\theta) + f_\alpha(-\theta) = -\mathrm{e}^{-\mathbf{i}\theta}\left(1 - \mathrm{e}^{\mathbf{i}\theta}\right)^\alpha - \mathrm{e}^{\mathbf{i}\theta}\left(1 - \mathrm{e}^{-\mathbf{i}\theta}\right)^\alpha.$$

It is obvious that $p_\alpha(0) = 0$. We want to show that such a zero is of order α, with $\alpha \in (1,2)$, according to the following definition.

Definition 18. Let $f : [a,b] \subset \mathbb{R} \to \mathbb{R}$ be a continuous nonnegative function. We say that f has a zero of *order* $\beta > 0$ at $\theta_0 \in [a,b]$ if there exist two real constants $C_1, C_2 > 0$ such that

$$\liminf_{\theta\to\theta_0} \frac{f(\theta)}{|\theta - \theta_0|^\beta} = C_1, \qquad \limsup_{\theta\to\theta_0} \frac{f(\theta)}{|\theta - \theta_0|^\beta} = C_2.$$

Recalling the definition of $f_\alpha(\theta)$ in equation (4.6), it is easy to see that $p_\alpha(\theta)$ is nonnegative; in fact making use of the Proposition 13 we obtain

$$\begin{aligned} p_\alpha(\theta) &= -\sum_{k=-1}^{\infty} g_{k+1}^{(\alpha)}(\mathrm{e}^{\mathbf{i}k\theta} + \mathrm{e}^{-\mathbf{i}k\theta}) \\ &= -\left[2g_1^{(\alpha)} + (g_0^{(\alpha)} + g_2^{(\alpha)})(\mathrm{e}^{\mathbf{i}\theta} + \mathrm{e}^{-\mathbf{i}\theta}) + \sum_{k=2}^{\infty} g_{k+1}^{(\alpha)}(\mathrm{e}^{\mathbf{i}k\theta} + \mathrm{e}^{-\mathbf{i}k\theta})\right] \\ &= -\left[2g_1^{(\alpha)} + 2(g_0^{(\alpha)} + g_2^{(\alpha)})\cos\theta + 2\sum_{k=2}^{\infty} g_{k+1}^{(\alpha)}\cos(k\theta)\right] \ge -2\sum_{k=-1}^{\infty} g_{k+1}^{(\alpha)} = 0. \end{aligned}$$

Proposition 16. *Let $\alpha \in (1,2)$, then the function $p_\alpha(\theta)$ defined in (4.7) has a zero of order α at 0.*

Proof. Let us rewrite $1-\mathrm{e}^{\mathbf{i}\theta}$ and $1-\mathrm{e}^{-\mathbf{i}\theta}$ in polar form

$$1-\mathrm{e}^{\mathbf{i}\theta} = \sqrt{2-2\cos\theta}\ \mathrm{e}^{\mathbf{i}\phi},$$
$$1-\mathrm{e}^{-\mathbf{i}\theta} = \sqrt{2-2\cos\theta}\ \mathrm{e}^{\mathbf{i}\psi},$$

where

$$\phi = \begin{cases} \arctan\left(\frac{-\sin\theta}{1-\cos\theta}\right), & \theta \neq 0 \\ \lim_{\theta\to 0^+} \arctan\left(\frac{-\sin\theta}{1-\cos\theta}\right) = -\frac{\pi}{2}, & \theta = 0 \end{cases}$$

and $\psi = -\phi$. We can then express $p_\alpha(\theta)$ as follows

$$\begin{aligned} p_\alpha(\theta) &= -\mathrm{e}^{-\mathbf{i}\theta}\left(\sqrt{2-2\cos\theta}\mathrm{e}^{\mathbf{i}\phi}\right)^\alpha - \mathrm{e}^{\mathbf{i}\theta}\left(\sqrt{2-2\cos\theta}\mathrm{e}^{-\mathbf{i}\phi}\right)^\alpha \\ &= -\sqrt{(2-2\cos\theta)^\alpha}\mathrm{e}^{\mathbf{i}(\alpha\phi-\theta)} - \sqrt{(2-2\cos\theta)^\alpha}\mathrm{e}^{-\mathbf{i}(\alpha\phi-\theta)} \\ &= -2\sqrt{(2-2\cos\theta)^\alpha}r_\alpha(\theta), \end{aligned}$$

where $r_\alpha(\theta) = \cos(\alpha\phi-\theta)$. Let us observe that $\lim_{\theta\to 0^-} r_\alpha(\theta) = \lim_{\theta\to 0^+} r_\alpha(\theta) = \cos\left(\alpha\frac{\pi}{2}\right)$. Now it is easy to see that

$$\lim_{\theta\to 0}\frac{p_\alpha(\theta)}{|\theta|^\alpha} = -2\lim_{\theta\to 0}\frac{(2-2\cos\theta)^{\frac{\alpha}{2}}}{|\theta|^\alpha}r_\alpha(\theta) = -2\cos\left(\alpha\frac{\pi}{2}\right) \in (0,2),$$

which proves that p_α has a zero of order α at 0, according to Definition 18. □

Remark 22. In Proposition 16 we assumed that $\alpha \in (1,2)$. Let us observe that when $\alpha = 1$ the order of the zero at 0 of $p_\alpha(\theta)$ is 2 since

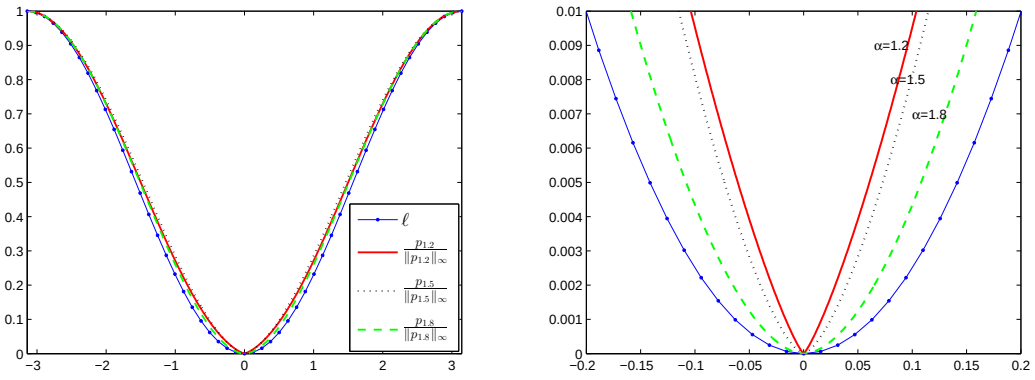
$$p_1(\theta) = -\mathrm{e}^{-\mathbf{i}\theta}\left(1-\mathrm{e}^{\mathbf{i}\theta}\right) - \mathrm{e}^{\mathbf{i}\theta}\left(1-\mathrm{e}^{-\mathbf{i}\theta}\right) = 2-2\cos(\theta).$$

Hence the statement in Proposition 16 is not true for $\alpha = 1$, while it remains true for $\alpha = 2$: indeed the polynomial

$$p_2(\theta) = -\mathrm{e}^{-\mathbf{i}\theta}\left(1+\mathrm{e}^{\mathbf{i}2\theta}-2\mathrm{e}^{\mathbf{i}\theta}\right) - \mathrm{e}^{\mathbf{i}\theta}\left(1+\mathrm{e}^{-\mathbf{i}2\theta}-2\mathrm{e}^{-\mathbf{i}\theta}\right) = 4-4\cos(\theta)$$

has a zero of order $\alpha = 2$ at 0, as expected.

Figure 4.1(a) compares the symbol $p_\alpha(\theta)$ normalized by $\|p_\alpha\|_{L^\infty}$ with the normalized symbol of the Laplacian operator given by $\ell(\theta) = 1-1\cos(\theta)$ for $\alpha = 1.2, 1.5, 1.8$ and varying θ in I_1. Figure 4.1(b) is a zoom of Figure 4.1(a) in a neighborhood of 0. Recalling that $\ell(\theta)$ has a zero of order 2 at 0, we observe that $\frac{p_\alpha(\theta)}{\|p_\alpha\|_{L^\infty}}$ approaches $\ell(\theta)$ and the order of its zero in 0 increases up to 2 as α tends to 2.



(a) (b)

Figure 4.1: (a) Comparison between the normalized symbol of the Laplacian operator $\ell(\theta)$ (blue bullet line) with $\frac{p_\alpha(\theta)}{\|p_\alpha\|_{L^\infty}}$ for $\alpha = 1.2$ (red solid line), $\alpha = 1.5$ (black dotted line) and $\alpha = 1.8$ (green dashed line) varying θ in I_1; (b) zoom of Figure 4.1(a) in a neighborhood of 0.

4.2.2 Nonconstant diffusion coefficients case

Now we focus on the symbol associated to $\left\{\mathcal{M}_{\alpha,N}^{(m)}\right\}_N$ and on its spectral distribution, when both $d_+(x,t)$ and $d_-(x,t)$ are nonconstant. For this purpose we need the notion of GLT sequences and the related theory (for $k=1$), introduced in Section 1.5.

Proposition 17. *Let us assume that $\nu_{M,N} = o(1)$ and that, fixed the instant of time t_m, $d_+(x) := d_+(x,t_m)$ and $d_-(x) := d_-(x,t_m)$ are both Riemann integrable over $[L,R]$. For the matrix $\mathcal{M}_{\alpha,N}^{(m)}$ defined as in (4.5), it holds*

$$\left\{\mathcal{M}_{\alpha,N}^{(m)}\right\}_N \sim_{\text{GLT}} \hat{h}_\alpha(\hat{x},\theta)$$

with

$$\hat{h}_\alpha(\hat{x},\theta) = h_\alpha(L+(R-L)\hat{x},\theta), \qquad h_\alpha(x,\theta) = d_+(x)f_\alpha(\theta) + d_-(x)f_\alpha(-\theta), \tag{4.8}$$

where $(\hat{x},\theta) \in [0,1]\times I_1$, $(x,\theta)\in[L,R]\times I_1$. Furthermore,

$$\left\{\mathcal{M}_{\alpha,N}^{(m)}\right\}_N \sim_\sigma (h_\alpha(x,\theta),[L,R]\times I_1),$$

and whenever $h_\alpha(x,\theta)$ is real-valued, i.e., if and only if $d_+(x)=d_-(x)$, we also have

$$\left\{\mathcal{M}_{\alpha,N}^{(m)}\right\}_N \sim_\lambda (h_\alpha(x,\theta),[L,R]\times I_1),$$

and indeed all the matrices $\mathcal{M}_{\alpha,N}^{(m)}$ have only real eigenvalues.

Proof. Let us observe that, fixed the instant of time t_m, the diagonal elements of the matrices $D_\pm^{(m)}$ are a uniform sampling of the functions $d_\pm(x)$, $x\in[L,R]$, and then $\left\{D_\pm^{(m)}\right\}_N \sim_{\text{GLT}} \hat{d}_\pm(\hat{x}) = d_\pm(L+(R-L)\hat{x})$, $\hat{x}\in[0,1]$ (see item **[GLT3]**). Since the GLT class is stable under linear combinations and products, as reported in item **[GLT2]**, and since Toeplitz sequences with L^1 symbols lie in the GLT class (see item **[GLT3]**), it is immediate to see that the matrix-sequence $\left\{D_+^{(m)}T_{\alpha,N} + D_-^{(m)}T_{\alpha,N}^T\right\}_N$ is still a member of the GLT class. The symbol of $\left\{D_+^{(m)}T_{\alpha,N} + D_-^{(m)}T_{\alpha,N}^T\right\}_N$ is $\hat{h}_\alpha(\hat{x},\theta) = \hat{d}_+(\hat{x})f_\alpha(\theta) + \hat{d}_-(\hat{x})f_\alpha(-\theta)$, $(\hat{x},\theta)\in[0,1]\times I_1$, again by item **[GLT2]**. Under the hypothesis that $\nu_{M,N}=o(1)$, the sequence $\{\nu_{M,N}I\}_N$ is a GLT sequence with zero symbol, as in item **[GLT4]**. This implies that $\left\{\mathcal{M}_{\alpha,N}^{(m)}\right\}_N \sim_{\text{GLT}} \hat{h}_\alpha(\hat{x},\theta)$, according to item **[GLT2]**.

Exploiting the Riemann integrability of $d_\pm(x)$ over $[L,R]$ and by item **[GLT1]**, we can conclude $\left\{\mathcal{M}_{\alpha,N}^{(m)}\right\}_N \sim_\sigma (\hat{h}_\alpha(\hat{x},\theta),[0,1]\times I_1)$ and hence $\left\{\mathcal{M}_{\alpha,N}^{(m)}\right\}_N \sim_\sigma (h_\alpha(x,\theta),[L,R]\times I_1)$, after an affine change of variable (refer to the integral expression in Definition 3).

Now, by exploiting Proposition 14 and Proposition 15, since $p_\alpha(\theta)$ is real-valued, it is clear that $h_\alpha(x,\theta)$ is real-valued if and only if $d_+(x)=d_-(x)$. Furthermore, under the condition that $d_+(x)=d_-(x)$ we deduce that $D_+^{(m)}=D_-^{(m)}$ which is a positive definite diagonal matrix, whence, choosing D as the positive definite square root of $D_+^{(m)}$, we find that $D^{-1}\,\mathcal{M}_{\alpha,N}^{(m)}\,D$ is similar to $\mathcal{M}_{\alpha,N}^{(m)}$ and real symmetric. Therefore all the eigenvalues of $\mathcal{M}_{\alpha,N}^{(m)}$ are real and we plainly have $\left\{\mathcal{M}_{\alpha,N}^{(m)}\right\}_N \sim_\lambda (h_\alpha(x,\theta),[L,R]\times I_1)$, by exploiting again the GLT machinery, as done before but in the Hermitian setting. □

Here we show in Proposition 18 that, if both diffusion coefficients are bounded and positive, the symbol $h_\alpha(x,\theta)$ (and hence $\hat{h}_\alpha(\hat{x},\theta)$), for the set of interest $\alpha\in(1,2)$, has always a zero at $\theta=0$ of order α (see Proposition 16 for the constant and equal coefficients case). This property is true independently of the constant or nonconstant character of the diffusion coefficients.

Proposition 18. *Given $p_\alpha(\theta)$ as in (4.7) and $h_\alpha(x,\theta)$ as in (4.8), the following two limit relations hold*

$$\lim_{\theta\to0^+}\frac{h_\alpha(x,\theta)}{p_\alpha(\theta)} = \frac{d_+(x)+d_-(x)}{2} - \mathbf{i}\tan\left(\alpha\frac{\pi}{2}\right)\frac{d_+(x)-d_-(x)}{2},$$
$$\lim_{\theta\to0^-}\frac{h_\alpha(x,\theta)}{p_\alpha(\theta)} = \frac{d_+(x)+d_-(x)}{2} + \mathbf{i}\tan\left(\alpha\frac{\pi}{2}\right)\frac{d_+(x)-d_-(x)}{2}.$$

Proof. As in the proof of Proposition 16 we exploit the polar form of $1-\mathrm{e}^{\mathbf{i}\theta}$ and $1-\mathrm{e}^{-\mathbf{i}\theta}$ and rewrite the quotient $\frac{h_\alpha(x,\theta)}{p_\alpha(\theta)}$ as follows

$$\begin{aligned}\frac{h_\alpha(x,\theta)}{p_\alpha(\theta)} &= \frac{-d_+(x)\sqrt{(2-2\cos\theta)^\alpha}\mathrm{e}^{\mathbf{i}(\alpha\phi-\theta)} - d_-(x)\sqrt{(2-2\cos\theta)^\alpha}\mathrm{e}^{-\mathbf{i}(\alpha\phi-\theta)}}{-2\sqrt{(2-2\cos\theta)^\alpha}\cos(\alpha\phi-\theta)}\\ &= \frac{d_+(x)\mathrm{e}^{\mathbf{i}(\alpha\phi-\theta)} + d_-(x)\mathrm{e}^{-\mathbf{i}(\alpha\phi-\theta)}}{2\cos(\alpha\phi-\theta)}\\ &= \frac{d_+(x)(\cos(\alpha\phi-\theta)+\mathbf{i}\sin(\alpha\phi-\theta))}{2\cos(\alpha\phi-\theta)} + \frac{d_-(x)(\cos(\alpha\phi-\theta)-\mathbf{i}\sin(\alpha\phi-\theta))}{2\cos(\alpha\phi-\theta)}\\ &= \frac{d_+(x)+d_-(x)}{2} + \mathbf{i}\tan(\alpha\phi-\theta)\,\frac{d_+(x)-d_-(x)}{2},\end{aligned}$$

where

$$\phi = \begin{cases} \arctan\left(\frac{-\sin\theta}{1-\cos\theta}\right), & \theta\neq 0,\\ \lim_{\theta\to 0^+}\arctan\left(\frac{-\sin\theta}{1-\cos\theta}\right) = -\frac{\pi}{2}, & \theta = 0.\end{cases}$$

It is easy to see that for $\alpha\in(1,2)$

$$\begin{aligned}\lim_{\theta\to 0^+}\tan(\alpha\phi-\theta) &= -\tan\left(\alpha\frac{\pi}{2}\right) > 0,\\ \lim_{\theta\to 0^-}\tan(\alpha\phi-\theta) &= \tan\left(\alpha\frac{\pi}{2}\right) < 0,\end{aligned}$$

and the thesis is proved. □

The previous Proposition 18 shows also the importance of the diffusion coefficients functions d_+ and d_-, that should be properly taken into account when defining a good preconditioner.

4.3 Analysis and design of numerical methods, via the spectral information

In this section we use the spectral information discussed in Section 4.2 to analyze in more detail the convergence of some recently proposed techniques ([100, 113]) and to design some structure preserving preconditioners for Krylov methods. It is divided in three parts. In Subsection 4.3.1, we observe that the superlinear convergence obtained in the constant coefficient case for the CGLS with a circulant preconditioner discussed in [100] cannot be ensured for any Krylov method when the diffusion coefficients are nonconstant or in the multidimensional setting even when the diffusion coefficients are constant. In Subsection 4.3.2 structure preserving preconditioners are studied and a preconditioning proposal with minimal bandwidth (and so with efficient computational cost) is proposed. Finally, in Subsection 4.3.3, with reference to the method indicated in [113], we briefly give a compact proof of the two-grid convergence (already proved in [113]), simply based on the properties of the symbol $p_\alpha(\theta)$, according to the results in [71, 33, 128]. Moreover, we give a theoretical motivation of the constant convergence rate of the V-cycle multigrid experimentally observed in [113] using the results in [4].

4.3.1 Negative results for the circulant preconditioner

We show here that the circulant preconditioning, which ensures a clustering at the unity in the case of constant coefficients (see Theorem 1 in [100]), cannot be extended in the variable coefficient setting.

Since circulant structures are special instances of Toeplitz structures, if a sequence of circulant matrices $\{C_N\}_N$ has a symbol $f(\theta)$, then its Toeplitz counterpart $\{T_N\}_N$ is such that $\{T_N-C_N\}_N\sim_\sigma(0,I_1)$. Hence, by invoking items **[GLT1-4]**, we deduce that the sequence $\{T_N-C_N\}_N$ is a GLT sequence with zero symbol and that both $\{C_N\}_N$, $\{T_N\}_N$ are also GLT sequences with symbol $f(\theta)$.

As a consequence, again using item **[GLT2]**, we infer that $\left\{C_N^{-1}\mathcal{M}_{\alpha,N}^{(m)}\right\}_N$ is a GLT sequence such that

$$\left\{C_N^{-1}\mathcal{M}_{\alpha,N}^{(m)}\right\}_N \sim_\sigma \left(\frac{\hat{h}_\alpha(\hat{x},\theta)}{f(\theta)}, [0,1]\times I_1\right)$$

when $\nu_{M,N} = o(1)$. Now if we look carefully at the expression of the function $\hat{h}_\alpha(\hat{x}, \theta)$ as reported in (4.8), we plainly see that the preconditioned sequence cannot be clustered at one, since the function $\hat{h}_\alpha(\hat{x}, \theta)/f(\theta)$ is a nontrivial function depending on the variable $\hat{x}$, whenever the diffusion coefficients are nonconstant functions. Therefore the superlinear behavior of any preconditioned Krylov method is lost, as long as we employ circulant preconditioners, in contrast with what happens in the constant coefficient case.

The second negative result concerns the possible application of the circulant preconditioner to multidimensional problems also in the constant coefficient setting. Indeed, we observe that in the constant coefficient case the matrix structures arising in the approximation of a FDE in multidimensional domain are essentially of multilevel Toeplitz type: we refer the reader to [148, 149] for the study of the related matrices in a variable coefficients setting in two and three dimensional spaces. As a consequence, the multilevel circulant preconditioning cannot ensure a superlinear convergence character, due to the negative results in [136] already mentioned in Section 1.6. More precisely, in [136], when considering k-level Toeplitz matrices and any type of circulant preconditioner, it is shown that at least $O(N^{\frac{k-1}{k}})$ outliers show up, where N is the size of the matrix: as a consequence, the superlinear behavior can be observed only for unilevel Toeplitz structures, i.e., for $k = 1$ and this agrees with the numerical results reported in the literature. However, in some specific cases, the resulting Krylov methods may be still very fast, especially if the conditioning is moderate and there are not outliers tending to zero.

4.3.2 Structure preserving preconditioners

The importance of preserving the same structure of the original matrix when designing a preconditioner is crucial to overcome the negative results in the multidimensional case and to have a preconditioned matrix with a well-conditioned matrix of the eigenvectors, which is relevant for the convergence of GMRES (see Tables 4.1–4.2).

To define a preconditioner with the same structure of the matrix $\mathcal{M}^{(m)}_{\alpha,N}$, and keeping at the same time a low computational cost, a small bandwidth matrix should be considered. On the other hand, the symbol of a bandwidth Toeplitz matrix is a trigonometric polynomial and hence the zero of the symbol cannot be of fractional order. We now introduce two preconditioners with minimal bandwidth and whose structure is the same of $\mathcal{M}^{(m)}_{\alpha,N}$.

The first preconditioner is defined as

$$P^{(m)}_{1,N} = \nu_{M,N} I + D^{(m)}_+ B_N + D^{(m)}_- B^T_N, \tag{4.9}$$

where B_N is the following approximation of the first derivative operator

$$B_N = \begin{bmatrix} 1 & -1 & 0 & \cdots & 0 \\ 0 & 1 & -1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & 0 & 1 & -1 \\ 0 & \cdots & \cdots & 0 & 1 \end{bmatrix}_{N\times N}.$$

The second preconditioner is given by

$$P^{(m)}_{2,N} = \nu_{M,N} I + D^{(m)}_+ L_N + D^{(m)}_- L^T_N, \tag{4.10}$$

where L_N is the Laplacian matrix

$$L_N = \begin{bmatrix} 2 & -1 & 0 & \cdots & 0 \\ -1 & 2 & -1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & -1 & 2 & -1 \\ 0 & \cdots & \cdots & -1 & 2 \end{bmatrix}_{N\times N}.$$

Both $P^{(m)}_{1,N}$ and $P^{(m)}_{2,N}$ are tridiagonal matrices, and hence the associated linear system can be solved optimally in $O(N)$ operations, by the standard Gaussian Elimination (known also as Thomas algorithm in the case of banded matrices). Therefore, the preconditioned Krylov method (CGLS, GMRES,

etc.) leads to a minimal computational cost per iteration of $O(N \log N)$ operations, typical of the un-preconditioned method with the considered matrices.

Let us assume that $\nu_{M,N} = o(1)$. The spectral distribution of sequences of the two preconditioners $P_{1,N}^{(m)}$ and $P_{2,N}^{(m)}$ can be derived using the tools in Subsection 4.2.2 like in Proposition 17. In particular, we have that

$$\left\{P_{1,N}^{(m)}\right\}_N \sim_\sigma (p_1^{(m)}(x,\theta), [L,R]\times I_1), \qquad p_1^{(m)}(x,\theta) = d_+(x,t_m)(1-\mathrm{e}^{-\mathbf{i}\theta}) + d_-(x,t_m)(1-\mathrm{e}^{\mathbf{i}\theta}),$$

and

$$\left\{P_{2,N}^{(m)}\right\}_N \sim_\lambda (p_2^{(m)}(x,\theta), [L,R]\times I_1), \qquad p_2^{(m)}(x,\theta) = (d_+(x,t_m) + d_-(x,t_m))(2-2\cos(\theta)).$$

If we further assume that $d_\pm(x,t) = d > 0$, i.e., $h_\alpha(x,\theta) = d \cdot p_\alpha(\theta)$ from Remark 22 and from Proposition 16, it holds that

$$\lim_{\theta\to 0} \frac{h_\alpha(x,\theta)}{p_k^{(m)}(x,\theta)} = \infty, \qquad k \in \{1,2\}, \tag{4.11}$$

hence both $P_{1,N}^{(m)}$ and $P_{2,N}^{(m)}$ cannot provide a clustering of the singular values or of the eigenvalues. On the other hand, as shown in the next numerical section, these two preconditioners show to be very effective. To justify such a behaviour we need the following theorem (Theorem 3.1 in [122]).

Theorem 19. *Let $f \in L^1(I_1)$ having in θ_0 the unique zero of order ρ and let $2l$ be the even number which minimizes the distance from ρ. If $g \in L^1(I_1)$ is a trigonometric polynomial which has a unique zero in θ_0 of order $2l$, then condition number of the preconditioned matrix $T_N^{-1}(g)T_N(f)$ is asymptotical to $N^{|2l-\rho|}$.*

The symbol $p_\alpha(\theta)$ is a real-valued nonnegative and continuous function with a zero of order α in 0. Consequently, using Theorem 19 with $f = p_\alpha$, $\theta_0 = 0$, $\rho = \alpha$, $l = 1$, $g = p_2$, we can conclude that the condition number of the preconditioned matrix $P_{2,N}^{(m)}\mathcal{M}_{\alpha,N}^{(m)}$, is asymptotical to $N^{|\alpha-2|}$, with $|\alpha-2| < 1$. In other words, thanks to the Axelsson-Lindskog bounds (see [6]), the number of iterations of a conjugate gradient type method grows as $O(N^{\frac{|\alpha-2|}{2}})$, which justifies the effectiveness of the preconditioner $P_{2,N}^{(m)}$ when α is close to 2. An analogous reasoning can be done for $P_{1,N}^{(m)}$.

Actually, we expect that similar arguments can be used to motivate the efficiency of both $P_{1,N}^{(m)}$ and $P_{2,N}^{(m)}$ independently of the constant or nonconstant character of the diffusion coefficients, despite a clustering of the eigenvalues or of the singular values of the preconditioned matrix-sequence cannot be ensured in any case. Indeed, using again Remark 22 and recalling Proposition 18, it is easy to see that the limit relation (4.11) is still true when the diffusion coefficients are nonconstant and equal functions, say $d_\pm(x,t) = d(x)$, with $d(x)$ positive and bounded, while when $d_\pm(x,t) = d_\pm > 0$ and $d_+ \neq d_-$, or when they are nonconstant and different from each other

$$\lim_{\theta\to 0} \frac{h_\alpha(x,\theta)}{p_k^{(m)}(x,\theta)} = \begin{cases} 0 & k=1, \\ \infty & k=2. \end{cases} \tag{4.12}$$

From relation (4.12) we deduce that $\frac{h_\alpha(x,\theta)}{p_k^{(m)}(x,\theta)}$ asymptotically behaves as $|\theta|^{\alpha-k}$, $\forall\alpha \in (1,2)$. As a consequence, we expect the condition number of the preconditioned matrix to be asymptotical to $N^{|\alpha-k|}$. In the light of this, we state that $P_{1,N}^{(m)}$ is a good preconditioner for α close to 1, while $P_{2,N}^{(m)}$ is a good preconditioner for α close to 2, (we can say $\alpha \geq 1.5$); cf. Tables 4.1–4.2 for the variable diffusion coefficients case.

4.3.3 Linear convergence of multigrid methods

Multigrid methods have shown to be a valid alternative to preconditioned Krylov methods also for FDEs [113]. Using the Ruge–Stuben theory [117], Theorem 4 in [113] shows that, in the constant coefficient case, i.e., $d_\pm(x,t) = d > 0$, the two-grid method converges with a constant convergence rate independent of N and m. Since in this case the matrix $\mathcal{M}_{\alpha,N}^{(m)}$ is a Toeplitz matrix, the classical multigrid theory for Toeplitz matrices developed in [71, 33, 128, 4] can be directly applied when the symbol is known. Under the assumptions that $d_\pm(x,t) = d > 0$ and $\nu_{M,N} = o(1)$, according to our previous analysis in Subsection 4.2.1, the symbol of the Toeplitz sequence $\left\{\mathcal{M}_{\alpha,N}^{(m)}\right\}_N$ is $d \cdot p_\alpha(\theta)$ (cf. Proposition 15).

When the grid transfer operator is the classical linear interpolation like in [113], the associated symbol is $2 + 2\cos(\theta)$. Therefore, according to Proposition 6, given a sequence of Toeplitz matrices $\{T_N(f)\}_N$ with a nonnegative symbol f, if

$$\limsup_{\theta\to 0} \frac{(2 + 2\cos(\theta + \pi))^2}{f(\theta)} = c < \infty, \tag{4.13}$$

then the two-grid method has a constant convergence rate. For $f(\theta) = d \cdot p_\alpha(\theta)$, the condition (4.13) is trivially satisfied with $c = 0$.

The varying coefficient case can be addressed thanks to the extension of the previous results given in [128]. Let d_+ and d_- be two uniformly bounded and positive functions. Then the linear convergence rate of the two-grid method is preserved combining Proposition 18 with Lemma 6.2 in [128].

The convergence analysis of the V-cycle, according to condition (1.48), it has to hold

$$\limsup_{\theta\to 0} \frac{2 + 2\cos(\theta + \pi)}{f(\theta)} = c < \infty. \tag{4.14}$$

Note that $f(\theta) = d \cdot p_\alpha(\theta)$ satisfies also the condition (4.14) with $c = 0$. This gives a theoretical justification of the linear convergence of the V-cycle experimentally observed in [113]. Actually, the Ruge-Stuben theory used to derive the condition (4.14) requires the Galerkin approach, while for computational convenience in [113] a rediscretization strategy is adopted. On the other hand, $c = 0$ suggests that the grid transfer operator is powerful enough, to work also under some perturbations.

In conclusion, taking into account that the order of the zero at 0 of $h_\alpha(x, \theta)$ in (4.8) remains bounded by 2, multigrid methods with linear interpolation, like that proposed in [113], represent a good solver or at least a robust preconditioner for Krylov methods. Moreover, despite to what happens for the circulant preconditioning (see Subsection 4.3.1), thanks to the theoretical results in [128, 1] we expect the multigrid to be optimal also in the multidimensional setting for variable and different diffusion coefficients, provided that they are uniformly bounded and positive.

Finally, we note that the knowledge of the symbol is crucial to define both the symbol of the preconditioner and the grid transfer operator of a multigrid method. The advantage of multigrid methods is that for the grid transfer operator it is enough that the associated symbol possesses a proper zero with an order larger than the order of the zero of $h_\alpha(x, \theta)$. Conversely, the preconditioner symbol has to match exactly the order of the zero of $h_\alpha(x, \theta)$. For this reason, the linear interpolation provides a multigrid with a constant convergence rate, while we cannot prove the eigenvalues clustering for the preconditioner $P^{(m)}_{2,N}$ in (4.10).

4.4 Numerical results

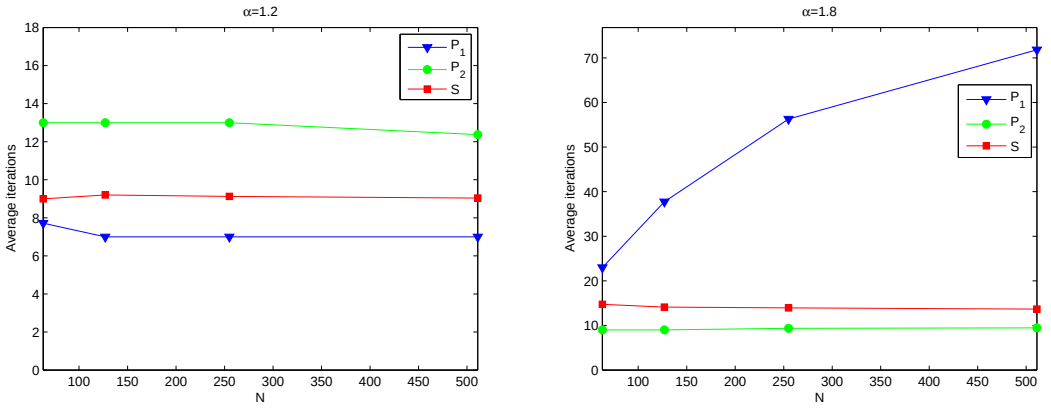
In this section we compare the new preconditioners $P^{(m)}_{1,N}$ and $P^{(m)}_{2,N}$ defined in (4.9) and (4.10), respectively, with the circulant preconditioner proposed in [100] defined as

$$S^{(m)}_N = \nu_{M,N} I + \bar{d}^{(m)}_+ s(T_{\alpha,N}) + \bar{d}^{(m)}_- s(T_{\alpha,N})^T,$$

where $\bar{d}^{(m)}_\pm = \frac{1}{N}\sum_{i=1}^N d^{(m)}_{\pm,i}$ and $s(T_{\alpha,N})$ is the Strang circulant matrix for $T_{\alpha,N}$. For notational simplicity, in the following, we remove the subscript N to each considered preconditioner. In all examples, we make also comparisons with a slightly modified version of $P^{(m)}_1$ and $P^{(m)}_2$, obtained by replacing the matrices $D^{(m)}_\pm$ with the averages $\bar{d}^{(m)}_\pm$, in their definition. We refer to these Toeplitz preconditioners as $P^{(m),\mathrm{av}}_1$, $P^{(m),\mathrm{av}}_2$, respectively. All the considered preconditioners are used to solve the FDE system (4.4), with the preconditioned CGLS and with the preconditioned GMRES methods. Regarding the stopping criterion for the CGLS, we use $\|r^k\|/\|r^0\| < 10^{-7}$, where r^k is the residual vector after k iterations. The GMRES method is computationally performed using the built-in `gmres` Matlab function with tolerance 10^{-7}. The initial guess at each time step is chosen for both methods as the zero vector.

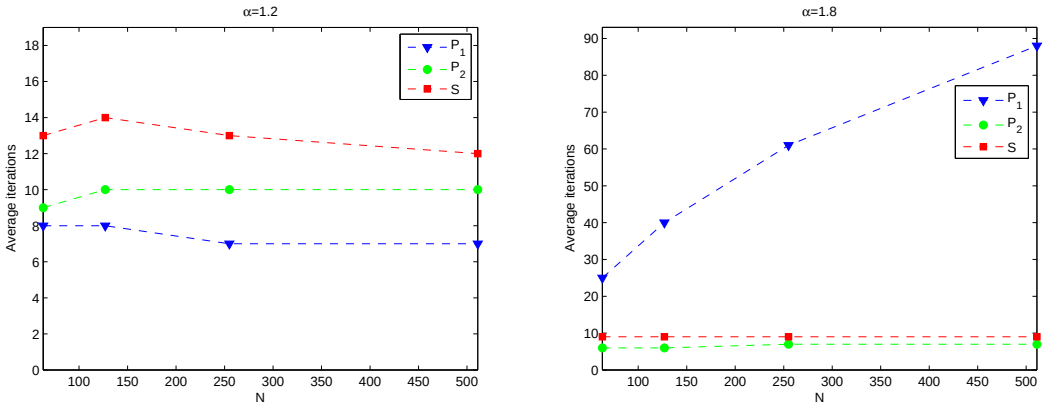
We do not test CGLS without preconditioning, since, as already shown in [100] for Example 1 and in [113] for Example 2 (when $\alpha = 1.5$), in these examples much more iterations are needed. Similar results can be obtained using GMRES without preconditioning.

The linear system with coefficient matrix $S^{(m)}$ is solved within $O(N\log(N))$ arithmetic operations by two FFTs, while the tridiagonal Toeplitz preconditioners can be implemented in $O(N)$ arithmetic operations by the Thomas algorithm. In the light of this, a comparison of the two preconditioning



(a) (b)

Figure 4.2: Example 1 - CGLS: (a) Average number of iterations varying N for $\alpha = 1.2$; (b) Average number of iterations varying N for $\alpha = 1.8$.



(a) (b)

Figure 4.3: Example 1 - GMRES: (a) Average number of iterations varying N for $\alpha = 1.2$; (b) Average number of iterations varying N for $\alpha = 1.8$.

strategies in terms of number of iterations is a reliable test which does not penalize the circulant preconditioner, rather gives it an edge.

We point out that apart form the preconditioner $S^{(m)}$, there is in literature a sophisticated preconditioning method (see [112]) which involves a sum of circulant matrices. The main idea of this method is to start from the following preconditioner

$$Q^{-1} = \sum_{i=1}^{N} e_i e_i^{\mathrm{T}} C_i^{-1},$$

where e_i denotes the i-th column of the identity matrix, while $C_i = \nu_{M,N} I + d_{+,i}^{(m)} s(T_{\alpha,N}) + d_{-,i}^{(m)} s(T_{\alpha,N})^T$, and then to apply a piecewise linear interpolation on a small subset $\{\tilde{x}_j\}_{j=1}^{\ell}$ of $\{x_i\}_{i=1}^{N}$, with $\ell \ll N$ which covers most of the interval $[L, R]$ to approximate C_i^{-1} as

$$C_i^{-1} \approx \sum_{j=1}^{\ell} \phi_j(x_i) \tilde{C}_j^{-1}, \quad i = 1, \ldots, N,$$

where $\{\phi_j\}_{j=1}^{\ell}$ is a basis for the space of the piecewise linear polynomials, while $\tilde{C}_j = \nu_{M,N} I + \tilde{d}_{+,j}^{(m)} s(T_{\alpha,N}) + \tilde{d}_{-,j}^{(m)} s(T_{\alpha,N})^T$, with $\tilde{d}_{\pm,j}^{(m)} = d_{\pm}(\tilde{x}_j, t_m)$, $j = 1, \ldots, \ell$. As shown in [112], this preconditioner reveals faster than $S^{(m)}$. On the other hand, it still involves the use of FFTs, more precisely, $O(\ell)$ FFTs per iteration are required which means that the product $Q^{-1}y$ for any vector y is computed in $O(\ell N \log N)$ operations. Moreover, in the multidimensional setting such a circulant

preconditioner suffers of the drawbacks already discussed in Subsection 4.3.1. Therefore, for the sake of simplicity, in the next examples we present a comparison of our two preconditioners only with $S^{(m)}$.

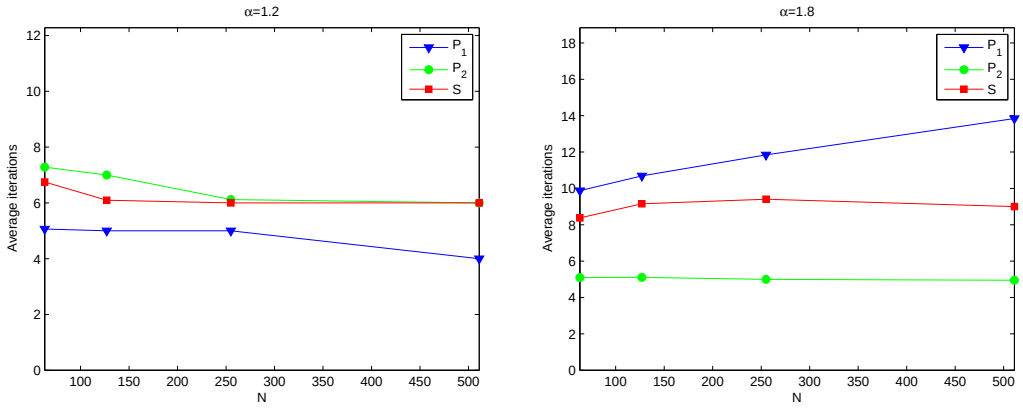
In the following tables, we display the average number of iterations computed as follows

$$\frac{1}{M}\sum_{m=1}^{M}\text{Iter}(m),$$

where $\text{Iter}(m)$ is the number of iterations required for solving (4.4) at time t_m.

α	$N+1$	P_1		P_2		S		P_1^{av}		P_2^{av}	
		CGLS	GMRES	CGLS	GMRES	CGLS	GMRES	CGLS	GMRES	CGLS	GMRES
1.2	2^6	7.7	8.0	13.0	9.0	9.0	13.0	10.0	12.0	12.0	10.0
	2^7	7.0	8.0	13.0	10.0	9.2	14.0	11.8	13.0	12.4	10.0
	2^8	7.0	7.0	13.0	10.0	9.1	13.0	12.8	14.0	12.2	10.0
	2^9	7.0	7.0	12.4	10.0	9.0	12.0	13.1	14.0	12.0	10.0
1.5	2^6	15.3	16.0	12.6	8.0	10.9	12.0	11.0	11.0	10.3	9.0
	2^7	18.3	20.0	13.2	9.0	10.7	12.0	13.2	13.0	11.8	10.0
	2^8	20.3	24.0	13.9	9.0	11.0	12.0	16.4	15.0	13.0	10.0
	2^9	22.4	26.0	14.3	10.0	10.6	12.0	18.6	16.0	13.9	11.0
1.8	2^6	23.0	25.0	9.0	6.0	14.8	9.0	13.0	10.0	9.4	8.0
	2^7	37.8	40.0	9.0	6.0	14.1	9.0	14.0	11.0	9.0	8.0
	2^8	56.3	61.0	9.4	7.0	14.0	9.0	15.7	12.0	9.0	8.0
	2^9	71.8	88.0	9.5	7.0	13.7	9.0	17.6	13.0	9.4	9.0

Table 4.1: Example 1 - Comparison of iterations in the CGLS and GMRES methods with preconditioners P_1, P_2, S, P_1^{av} and P_2^{av} for $\alpha = 1.2, 1.5, 1.8$ and $M = \frac{N+1}{2}$.



(a) (b)

Figure 4.4: Example 2 - CGLS: (a) Average number of iterations varying N for $\alpha = 1.2$; (b) Average number of iterations varying N for $\alpha = 1.8$.

Example 1. In this example we consider a FDE problem of type (4.1) with nonconstant diffusion coefficients

$$d_+(x,t) = \Gamma(3-\alpha)x^\alpha, \quad d_-(x,t) = \Gamma(3-\alpha)(2-x)^\alpha.$$

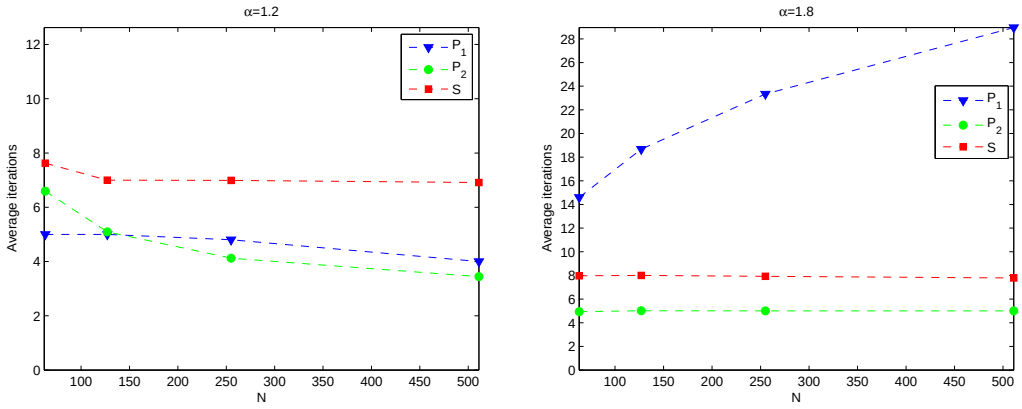
The spatial domain is $[L,R] = [0,2]$, while the time interval is $[0,T] = [0,1]$. The source term and the initial condition are given by

$$f(x,t) = -32\mathrm{e}^{-t}\left(x^2 + \frac{1}{8}(2-x)^2(8+x^2) - \frac{3}{3-\alpha}[x^3+(2-x)^3] + \frac{3}{(4-\alpha)(3-\alpha)}[x^4+(2-x)^4]\right),$$

$$u(x,0) = 4x^2(2-x)^2.$$

The exact solution of this problem is known and is given by $u(x,t) = 4\mathrm{e}^{-t}x^2(2-x)^2$. Since the diffusion coefficients do not depend on t, the coefficient matrix and all preconditioners for this example are independent of the time step. For this reason we omit the superscript (m). For this example we choose $\Delta x = \Delta t$. In this case,

$$\nu_{M,N} = \frac{\Delta x^\alpha}{\Delta t} = \Delta x^{\alpha-1}$$



(a) (b)

Figure 4.5: Example 2 - GMRES: (a) Average number of iterations varying N for $\alpha = 1.2$; (b) Average number of iterations varying N for $\alpha = 1.8$.

α	$N+1$	P_1		P_2		S		P_1^{av}		P_2^{av}	
		CGLS	GMRES	CGLS	GMRES	CGLS	GMRES	CGLS	GMRES	CGLS	GMRES
1.2	2^6	5.1	5.0	7.3	6.6	6.8	7.6	6.8	6.3	6.8	6.9
	2^7	5.0	5.0	7.0	5.1	6.1	7.0	6.7	5.3	6.0	5.3
	2^8	5.0	4.8	6.1	4.1	6.0	7.0	6.3	5.1	5.2	4.2
	2^9	4.0	4.0	6.0	3.4	6.0	6.9	6.0	4.4	5.0	3.5
1.4	2^6	6.3	7.5	7.1	5.8	7.1	8	8.1	7.3	6.4	6.4
	2^7	6.1	7.3	7.0	5.1	7.0	8.6	8.2	7.2	6.0	5.3
	2^8	5.9	7.1	7.0	5.0	7.0	8.6	8.1	7.0	5.8	5.0
	2^9	5.2	7.0	6.4	4.7	7.0	8.0	8.0	7.0	5.5	5.0
1.5	2^6	7.1	8.8	7.0	5.6	7.2	8.4	8.7	7.6	6.3	6.0
	2^7	6.8	9.2	7.0	5.1	7.1	8.8	8.8	8.0	6.0	5.5
	2^8	6.2	9.2	7.0	5.0	7.0	8.8	8.6	8.0	5.7	5.3
	2^9	6.0	9.4	6.5	5.0	7.0	8.7	8.3	8.0	5.7	5.1
1.6	2^6	7.8	10.6	6.9	5.3	7.6	8.0	9.3	7.9	6.2	5.7
	2^7	7.5	11.4	7.0	5.1	7.7	8.7	9.3	8.2	5.6	5.5
	2^8	7.4	12.1	6.6	5.0	7.3	8.7	8.8	8.1	5.4	6.0
	2^9	7.6	12.8	6.3	5.0	7.0	8.6	8.3	8.0	5.3	6.0
1.8	2^6	9.9	14.6	5.1	4.9	8.4	8.0	9.3	7.8	4.5	5.3
	2^7	10.7	18.7	5.1	5.0	9.2	8.0	8.8	8.3	5.0	5.2
	2^8	11.8	23.3	5.0	5.0	9.4	7.9	8.4	8.1	5.0	5.2
	2^9	13.8	29.0	4.9	5.0	9.0	7.8	7.7	8.0	4.8	5.1

Table 4.2: Example 2 - Number of iterations in the CGLS and GMRES methods with preconditioners $P_1^{(m)}$, $P_2^{(m)}$, $S^{(m)}$, $P_1^{(m),\text{av}}$ and $P_2^{(m),\text{av}}$ for $\alpha = 1.2, 1.4, 1.5, 1.6, 1.8$ and $M = \frac{N+1}{2}$.

which, being $0 < \alpha - 1 < 1$, tends to zero as N tends to ∞. Such a choice implies that the number of time steps M is given by $M = \frac{(N+1)T}{R-L} = \frac{N+1}{2}$. In Table 4.1 we compare the iterations provided by the CGLS and the GMRES methods with preconditioners P_1, P_2, S, P_1^{av} and P_2^{av} for $\alpha = 1.2, 1.5, 1.8$. We observe that preconditioner P_1 is suitable for α close to 1 (see Figures 4.2(a) and 4.3(a)). When α is close to 2 both P_2 (see Figures 4.2(b) and 4.3(b)) and P_2^{av} are good preconditioners for CGLS and GMRES methods.

Example 2. The following example consists in an anomalous diffusive process of a Gaussian pulse. Let us define

$$d_+(x,t) = 0.1(1 + x^2 + t^2), \quad d_-(x,t) = 0.1(1 + (2-x)^2 + t^2)$$

and set $[L, R] = [0, 2]$ and $[0, T] = [0, 1]$. The initial condition is given by

$$u(x, 0) = \mathrm{e}^{-\frac{(x-x_c)^2}{2\sigma^2}},$$

with $x_c = 1.2$ and $\sigma = 0.08$, and the source term is $f(x,t) = 0$. As in the previous example, we set $\Delta x = \Delta t$. In Table 4.2 we compare the number of iterations provided by the CGLS and GMRES methods with preconditioners $P_1^{(m)}$, $P_2^{(m)}$, $P_1^{(m),\text{av}}$, $P_2^{(m),\text{av}}$ and $S^{(m)}$ for $\alpha = 1.2, 1.4, 1.5, 1.6, 1.8$. As

in Example 1, we observe that $P_1^{(m)}$ is the best preconditioner for both CGLS and GMRES methods when α is close to 1 (see Figures 4.4(a) and 4.5(a)). For α close to 2, CGLS and GMRES methods perform better with preconditioners $P_2^{(m)}$ (see Figures 4.4(b) and 4.5(b)) and $P_2^{(m),\mathrm{av}}$. To be precise, for $\alpha = 1.4, 1.5, 1.6, 1.8$ these numerical results suggest using preconditioner $P_2^{(m)}$ with the GMRES method and preconditioner $P_2^{(m),\mathrm{av}}$ with the CGLS method.

Chapter 5

A block multigrid strategy for two-dimensional coupled PDEs

Finite element approximation of coupled differential boundary value problems gives rise to a sequence of large scale structured two-by-two block matrices. We are interested in the efficient iterative solution of the so arising linear systems, with the aim of constructing optimal preconditioning methods that are robust with respect to the relevant parameters of the problem. As recalled in Section 5.1, a classical approach is based on a exact factorization of the coefficient matrix, which leads to the requirement of fast solvers for the head-block linear system and for the Schur complement linear system. In this chapter we focus on the former issue and use the spectral tools introduced in Chapter 1 to perform a spectral analysis of the head-block matrix (Section 5.2). Moreover, by exploiting the spectral information, in Section 5.3 we design a multigrid method with an ad hoc grid transfer operator. Choosing damped Jacobi or Gauss-Seidel as smoothers and using the resulting solver as preconditioner for Krylov methods we obtain a more competitive strategy than the aggregation based algebraic multigrid, widely employed in the relevant literature (Section 5.4).

5.1 Problem setting: coupled PDEs

We are interested in solving large linear systems arising from the finite element approximation of a coupled system of PDEs. As an example we consider the linear elasticity problem in saddle point form. Such a problem can be viewed as a subproblem of a more general coupled system of PDEs arising from the so-called Glacial Isostatic Adjustment (GIA) model, used in Geophysics to describe the response of the Earth to redistribution of mass due to alternating glaciation and deglaciation periods, cf. e.g., [154, 155, 101]. When the Earth is modeled as a flat homogeneous incompressible material body and only its elastic response is considered, we obtain a two-dimensional coupled system of PDEs, which in its simplest form reads as follows

$$\begin{aligned} -2\mu\,\Delta u_1 + \mu\,\frac{\partial}{\partial x_2}\left(\frac{\partial u_2}{\partial x_1} - \frac{\partial u_1}{\partial x_2}\right) - c_1\frac{\partial u_1}{\partial x_1} + \mu\frac{\partial p}{\partial x_1} &= f_1, \\ -2\mu\,\Delta u_2 + \mu\,\frac{\partial}{\partial x_1}\left(\frac{\partial u_1}{\partial x_2} - \frac{\partial u_2}{\partial x_1}\right) - c_2\frac{\partial u_2}{\partial x_2} + \mu\frac{\partial p}{\partial x_2} &= f_2 \\ \mu\nabla\cdot u - \rho\, p &= 0, \end{aligned} \tag{5.1}$$

Here $u_1(x_1, x_2)$ and $u_2(x_1, x_2)$ are the displacements in x_1 and x_2 directions, respectively, $(x_1, x_2) \in \Omega \subset \mathbb{R}^2$, λ and μ are the so-called Lamé coefficients, for simplicity assumed not to vary in space, $c = [c_1, c_2]^T$ is an advection vector, $\rho = \mu^2/\lambda$ and $\rho = 0$ in the case of compressible materials. The two unknowns are the displacements $u = [u_1, u_2]^T$ and the pressure p. Discretizing (5.1) by the stable Finite Element Method (FEM) pair Q1isoQ1 (cf. [24]), we obtain a linear system with a two-by-two block matrix of saddle point form,

$$\mathcal{A} = \begin{bmatrix} K & B^T \\ B & -\rho M \end{bmatrix}. \tag{5.2}$$

Here, M is the mass matrix, ρ is a positive integer, and B and B^T correspond to discrete divergent and gradient operators, respectively. The head-block K itself is a two-by-two block matrix where the

structure is due to imposing the so-called separate displacement ordering on the components of the vector u (see [67] for details). Furthermore, discretizing the problem of interest for a sequence of discretization parameters we obtain a sequence of matrices of size that grows to infinity as the approximation error tends to zero. In other words, the more accurate the approximation is, the larger the related system size becomes. This rules out the direct methods as demanding too much computer resources, and other methods as preconditioned Krylov or multigrid methods have to be applied. Most of the preconditioning strategies for two-by-two block systems as in (5.2) are based on the following factorization of $\mathcal{A}$

$$\mathcal{A} = \begin{bmatrix} K & B^T \\ B & -\rho M \end{bmatrix} = \begin{bmatrix} K & 0 \\ B & -S \end{bmatrix} \begin{bmatrix} I & K^{-1}B^T \\ 0 & \tilde{I} \end{bmatrix},$$

where S is the negative *Schur complement* of $\mathcal{A}$ defined as $S = \rho M + BK^{-1}B^T$ and I and $\tilde{I}$ are identity matrices of proper order. Two examples of preconditioners are

$$\mathcal{B}_1 = \begin{bmatrix} K & 0 \\ B & -\widehat{S} \end{bmatrix}, \qquad \mathcal{B}_2 = \begin{bmatrix} K & 0 \\ 0 & -\widehat{S} \end{bmatrix},$$

where $\widehat{S}$ is an approximation of the exact Schur complement S.

It is well-known that a necessary condition for the above preconditioners to be efficient is that $\widehat{S}$ is an high quality approximation of S. Some approaches can be found in [66, 108, 14]. Another necessary condition is to solve the linear system with K accurately enough, thus, we need efficient inner solvers. Since the considered problem is elliptic, the AMG method is a suitable choice. As observed in [66], however, solving systems with K is the most time consuming part when applying the preconditioner. In addition, in three dimensions the memory demands become rather prohibitive.

In this chapter we exploit the fact that, up to low-rank perturbations, the matrix block K is 2-level block Toeplitz matrix. We provide a spectral analysis of it and its 2×2 matrix-valued symbol to design a 2D block multigrid with an ad hoc grid transfer operator and formulate our multigrid for a more general k-level Toeplitz matrix, associated to a $s \times s$ matrix-valued symbol. Choosing damped Jacobi or Gauss-Seidel methods as smoothers, the resulting method reveals to be more efficient than some of the AMG methods in use.

5.2 Structure and symbol of the coefficient matrix

Our aim is to efficiently solve a linear system

$$Ku = b, \quad u, b \in \mathbb{C}^N,$$

where $K \in \mathcal{M}_N$ is, for instance, the head-block of $\mathcal{A}$ defined in (5.2), taking advantage of its structure and especially of its spectral features. To do that, we use the spectral tools introduced in Chapter 1, namely the notion of a multilevel block Toeplitz matrix associated with a matrix-valued symbol, of spectral distribution, and of GLT. Indeed, a multigrid strategy for Toeplitz matrices requires information about the symbol of the coefficient matrix to define both the projector and the matrix at the coarse level. Such spectral information compactly contained in the symbol is crucial, not only in the Toeplitz setting, but even when the coefficient matrix is Toeplitz up to low-rank perturbations and can be interpreted as a GLT.

Using Definition 7, we can explicitly express the symbol of the matrix K. Denote $n = (n_1, n_2)$ and $\widehat{n} = n_1 n_2$. From (5.1) we have that K is a two-by-two block matrix of size $N = 2\widehat{n}$, that is

$$K = \begin{bmatrix} K_{11} & K_{12} \\ K_{21} & K_{22} \end{bmatrix}, \qquad K_{ij} \in \mathcal{M}_{\widehat{n}}, \qquad i, j = 1, 2. \tag{5.3}$$

As is seen from (5.1), K can be symmetric positive definite ($c = 0$) or nonsymmetric. Consider the symmetric case. If for simplicity, we discretize the original problem by a Finite Difference Method (FDM), we obtain that $K_{ij} = T_n(f_{ij})$, $i, j = 1, 2$, where $T_n(f_{ij})$ is a 2-level Toeplitz matrix generated by $f_{ij} : I_2 \to \mathbb{C}$, $i, j = 1, 2$, with (see [67])

$$\begin{aligned} f_{11}(\theta_1, \theta_2) &= 4 - 2\cos\theta_1(1 + \cos\theta_2), \\ f_{12}(\theta_1, \theta_2) &= f_{21}(\theta_1, \theta_2) = \sin\theta_1 \sin\theta_2, \\ f_{22}(\theta_1, \theta_2) &= 4 - 2\cos\theta_2(1 + \cos\theta_1). \end{aligned} \tag{5.4}$$

From Theorem 5, it holds that $\{T_n(f_{ij})\}_{n\in\mathbb{N}^2} \sim_\lambda (f_{ij}, \mathcal{I}_1)$, $i,j = 1,2$.

It is easy to see that through a proper permutation matrix Π of size $2\widehat{n}$ we can write

$$\Pi K \Pi^T = \Pi \begin{bmatrix} T_n(f_{11}) & T_n(f_{12}) \\ T_n(f_{21}) & T_n(f_{22}) \end{bmatrix} \Pi^T = T_n(f), \tag{5.5}$$

where $f : I_2 \to \mathcal{M}_2$ is defined as follows

$$f(\theta_1,\theta_2) = \begin{bmatrix} f_{11}(\theta_1,\theta_2) & f_{12}(\theta_1,\theta_2) \\ f_{21}(\theta_1,\theta_2) & f_{22}(\theta_1,\theta_2) \end{bmatrix}.$$

Therefore, K is a 2-level block Toeplitz matrix associated to the $\mathcal{M}_2$-valued function f. Note that this permutation of the matrix K imposes the structure that arises if, when discretizing (5.1), the displacements are ordered per mesh point, i.e., the separate displacement ordering is not imposed. Since f is symmetric, from Theorem 8 it holds that $\{T_n(f)\}_{n\in\mathbb{N}^2} \sim_\lambda (f, I_k)$. To write explicitly the permutation matrix Π, let us define by e_j, $j = 1,\dots,2\widehat{n}$ the j-th column of the identity matrix of size $2\widehat{n}$ and by π_j, $j = 1,\dots,2\widehat{n}$ the j-th column of Π. Then,

$$\pi_j = \begin{cases} e_{2j-1} & j = 1,\dots,\widehat{n} \\ e_{2(j-\widehat{n})} & j = \widehat{n}+1,\dots,2\widehat{n} \end{cases}. \tag{5.6}$$

In other words, Π is the $2\widehat{n} \times 2\widehat{n}$ matrix whose first $\widehat{n}$ columns are the odd columns of $I_{2\widehat{n}}$, while the remaining ones are the even columns of the same matrix.

If instead of finite differences we use the Q1isoQ1 FEM scheme for discretizing the original problem, then we obtain $K_{ij} = T_n(f_{ij}) + E_n^{(ij)}$, $i,j = 1,2$, where $E_n^{(ij)}$ is a low-rank perturbation whose rank grows at most proportionally to $\sqrt{\widehat{n}}$. This means that, $\{E_n^{(ij)}\}_{n\in\mathbb{N}^2} \sim_\sigma 0$ and so, by the **[GLT4]**, the sequence $\{E_n^{(ij)}\}_{n\in\mathbb{N}^2}$ is a GLT sequence with symbol identically zero. Using **[GLT3]**, also $\{T_n(f_{ij})\}_{n\in\mathbb{N}^2}$ is a GLT sequence with symbol f_{ij} and then, by **[GLT2]**, the sequence $\{T_n(f_{ij}) + E_n^{(ij)}\}_{n\in\mathbb{N}^2}$ is a GLT sequence with the same symbol. As a consequence, it is clear that the symbol does not depends on the scheme (FDM or Q1isoQ1 FEM) used to discretize the problem. In the light of this, in the next sections, we do not consider the low-rank perturbation and discuss only the construction of multigrid methods for Toeplitz matrices.

Remark 23. Note that discretizing the original problem using the Q1isoQ1 finite elements means that the matrix K corresponds to a Q1 (bilinear basis functions on a quadrilateral mesh) discretization of the part of the first two equations in (5.1), that contains only derivatives of the displacements.

5.3 AMG for Toeplitz matrices with matrix-valued symbol

In Section 1.8, we recalled the multigrid idea and its application to multilevel Toeplitz sequences with scalar-valued symbol. As regards, multigrid method for Toeplitz matrices with block symbol, seminal results have been proposed in [96], even though, up to our knowledge, when the block symbol is not diagonal no convergence analysis has been performed yet. In this section we describe a AMG for multilevel Toeplitz matrices with matrix-valued symbol formally extending the idea in [96]. As observed in that paper, for the 1-level case, when the generating function is an $s \times s$ diagonal matrix-valued function, a multigrid method on the whole matrix can be seen as s independent multigrid methods for 1-level Toeplitz matrices with scalar-valued symbols. This approach can be applied also to the multilevel case. Let $f : I_k \to \mathcal{M}_s$ defined as

$$f(\theta) = \begin{bmatrix} f_{11}(\theta) & 0 & \cdots & 0 \\ 0 & f_{22}(\theta) & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & f_{ss}(\theta) \end{bmatrix}.$$

If we assume that $f_{jj} : I_k \to \mathbb{C}$, $j = 1,\dots,s$ has only a single isolated zero in I_k of order (at most) $2q$, we can define s AMG methods as discussed in Subsection 1.8.2 (one for each f_{jj}) choosing polynomials like in (1.50) as symbol for the projectors.

In [96], the authors consider the more general case when the generating matrix is not diagonal, but HPD with a constant basis of eigenvectors. In brief, the main idea is to diagonalize the generating function and to choose the projector in view of the location of the zeroes of its eigenvalues. We formally

extend this idea also in the multidimensional setting and to nonconstant basis of eigenvectors. Let $f : I_k \to \mathcal{M}_s$ be HPD and let us diagonalize $f(\theta)$ as follows

$$f(\theta) = Q(\theta)\Lambda(\theta)Q(\theta)^*, \tag{5.7}$$

where

$$\Lambda(\theta) = \begin{bmatrix} \lambda_1(\theta) & 0 & \cdots & 0 \\ 0 & \lambda_2(\theta) & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & \lambda_s(\theta) \end{bmatrix},$$

with $\lambda_j : I_k \to \mathbb{C}$, $j = 1, \ldots, s$ is a nonnegative function.

As Proposition 3 can be straightforwardly extended to any HPD matrix, the smoothing properties are ensured.

Proposition 19. *Let $A = T_n(f)$ with $f : I_k \to \mathcal{M}_s$ HPD and defined according to equation* (5.7)*. Let $S = I - \omega A$ and $\tilde{S} = I - \tilde{\omega}A$. If*

$$0 \le \omega, \tilde{\omega} \le \frac{2}{\max_{j=1,\ldots,s} \|\lambda_j\|_{L^\infty}},$$

then there exist $\alpha, \beta > 0$ such that the smoothing properties (1.43) *and* (1.44) *hold with $\nu, \theta \in \mathbb{N}$.*

Proof. Since f is HPD, then A is HPD and so $A = Q\Lambda(A)Q^*$. Note that the notation $A \le B$ for any given two Hermitian matrices, means that the matrix $B - A$ is HPSD. Note that (1.43) is equivalent to require that

$$S^\nu A S^\nu \le A - \alpha S^\nu A^2 S^\nu,$$

which reduces to

$$\Lambda(A)(I - \omega\Lambda(A))^{2\nu} \le \Lambda(A) - \alpha\Lambda(A)^2(I - \omega\Lambda(A))^{2\nu} \tag{5.8}$$

thanks to the expression of S. The matrix inequality (5.8) is implied by the function inequalities

$$\lambda_j(I - \omega\lambda_j)^{2\nu} \le \lambda_j - \alpha\lambda_j^2(I - \omega\lambda_j)^{2\nu}, \qquad j = 1, \ldots, s.$$

Performing the same function study presented in [1, Proposition 3], we deduce that the smoothing property (1.43) follows whenever the parameter ω satisfies the inequalities $0 \le \omega \le 2/\max_{j=1,\ldots,s} \|\lambda_j\|_{L^\infty}$.

Similarly, we can prove the smoothing property (1.44) when $0 \le \tilde{\omega} \le 2/\max_{j=1,\ldots,s} \|\lambda_j\|_{L^\infty}$. □

To define the projector at level i, $i = 0, \ldots, m-1$ for a fixed $N_i = (2^{t-i} - 1)e$, ($e = (1, \ldots, 1) \in \mathbb{N}^k$, i, t positive integers with $i < t$) we use the following cutting matrix

$$K_{N_i}^{[s]} = K_{(N_i)_1} \otimes \cdots \otimes K_{(N_i)_k} \otimes I_s,$$

where I_s is the $s \times s$ identity matrix and $K_{(N_i)_\ell} \in \mathbb{R}^{(N_{i+1})_\ell \times (N_i)_\ell}$, $\ell = 1, \ldots, k$ is either defined as in (1.45) for $N_i = (2^{t-i} - 1)e$ or in the case of $N_i = (2^{t-i} + 1)e$ it is chosen as

$$K_{(N_i)_\ell} = \begin{bmatrix} 1 & 0 & & & & \\ & 0 & 1 & 0 & & \\ & & & & \ddots & \\ & & & & 0 & 1 \end{bmatrix}.$$

Assuming that for the matrix A_i the associated symbol f_i has all eigenvalues functions $\lambda_j^{(i)}$ with only a single isolated zero at the same point $\theta_i^0 \in I_k$ of order (at most) $2q$ for every $j = 1, \ldots, s$, we define the projector as

$$P_{N_i}^{[s]} = T_{N_i}(p_i)(K_{N_i}^{[s]})^T, \qquad p_i(\theta) = c \cdot \prod_{j=1}^{k} [1 + \cos(\theta_j - (\theta_i^0)_j)]^q \cdot I_s. \tag{5.9}$$

The matrix at the coarse grid is obtained by the Galerkin approach, that is as $A_{i+1} = (P_{N_i}^{[s]})^T A_i P_{N_i}^{[s]}$.

5.3.1 AMG for 2-level block matrix discretized by Q1 FEM

Let us recall that, according to the notation in Definition 7, in our test problem (5.5) we have $k = 2$, $s = 2$, and $f : I_2 \to \mathcal{M}_2$ is the following symmetric positive defined matrix function

$$f(\theta_1,\theta_2) = \begin{bmatrix} f_{11}(\theta_1,\theta_2) & f_{12}(\theta_1,\theta_2) \\ f_{12}(\theta_1,\theta_2) & f_{22}(\theta_1,\theta_2) \end{bmatrix} = \begin{bmatrix} 4 - 2\cos\theta_1(1+\cos\theta_2) & \sin\theta_1\sin\theta_2 \\ \sin\theta_1\sin\theta_2 & 4 - 2\cos\theta_2(1+\cos\theta_1) \end{bmatrix} \tag{5.10}$$

according to (5.4). Note that functions f_{11} and f_{22} have a zero in $(0,0)$ of order 2.

By direct computation of the zeros of the characteristic polynomial of $f(\theta_1,\theta_2)$, we obtain the following two eigenvalue functions

$$\lambda_1(\theta_1,\theta_2) = \frac{f_{11}(\theta_1,\theta_2) + f_{22}(\theta_1,\theta_2) + \sqrt{(f_{11}(\theta_1,\theta_2) - f_{22}(\theta_1,\theta_2))^2 + 4f_{12}^2(\theta_1,\theta_2)}}{2},$$
$$\lambda_2(\theta_1,\theta_2) = \frac{f_{11}(\theta_1,\theta_2) + f_{22}(\theta_1,\theta_2) - \sqrt{(f_{11}(\theta_1,\theta_2) - f_{22}(\theta_1,\theta_2))^2 + 4f_{12}^2(\theta_1,\theta_2)}}{2}.$$

By using the definition of f in (5.10) and the identity $2 - 2\cos\theta = 4\sin^2\frac{\theta}{2}$ we can write λ_1, λ_2 explicitly as follows

$$\lambda_1(\theta_1,\theta_2) = 4 - (\cos\theta_1 + \cos\theta_2) - 2\cos\theta_1\cos\theta_2 + \sqrt{(\cos\theta_1 - \cos\theta_2)^2 + \frac{1}{4}(1-\cos 2\theta_1)(1-\cos 2\theta_2)},$$
$$\lambda_2(\theta_1,\theta_2) = 4 - (\cos\theta_1 + \cos\theta_2) - 2\cos\theta_1\cos\theta_2 - \sqrt{(\cos\theta_1 - \cos\theta_2)^2 + \frac{1}{4}(1-\cos 2\theta_1)(1-\cos 2\theta_2)}.$$

Since f is symmetric, there exists a unitary matrix $Q \in \mathcal{M}_2$ such that $f = Q\Lambda Q^*$, where

$$\Lambda = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} \qquad \text{and} \qquad Q = \begin{pmatrix} q_1 & -q_2 \\ q_2 & q_1 \end{pmatrix}.$$

Computing the eigenvectors corresponding to λ_1, λ_2, we have

$$f_{11}q_1 + f_{12}q_2 = \lambda_1 q_1 \iff q_2 = \frac{\lambda_1 - f_{11}}{f_{12}} q_1.$$

Using the expressions for λ_1 and f_{11} we obtain

$$\lambda_1 - f_{11} = \cos\theta_1 - \cos\theta_2 + \sqrt{(\cos\theta_1 - \cos\theta_2)^2 + \sin^2\theta_1\sin^2\theta_2}.$$

Note that the factor $(\lambda_1 - f_{11})/f_{12} \to 0$ when θ_1 and θ_2 approach zero. Therefore, for the ill-conditioned subspace associated to $\theta_1, \theta_2 \to 0$ the eigenvector $(q_1(\theta_1,\theta_2), q_2(\theta_1,\theta_2))^T \to (1,0)^T$ and hence $Q \to I_2$. It follows that when $\theta_1, \theta_2 \to 0$ the symbol (5.10) is almost diagonal and the construction of the projector (5.9) requires only the computation of the zeros, with their order, of the eigenvalue functions λ_1 and λ_2.

Both eigenvalues λ_1 and λ_2 have a zero of order 2 in $(0,0)$ (see Figures 5.1(a) and 5.1(b)). In fact, if we fix $\theta_2 = 0$, then

$$\lim_{\theta_1\to 0} \frac{\lambda_1(\theta_1,0)}{\theta_1^2} = \lim_{\theta_1\to 0} \frac{4 - (\cos\theta_1 + 1) - 2\cos\theta_1 + \sqrt{(\cos\theta_1 - 1)^2}}{\theta_1^2} = \lim_{\theta_1\to 0} \frac{4 - 4\cos\theta_1}{\theta_1^2} = 2,$$
$$\lim_{\theta_1\to 0} \frac{\lambda_2(\theta_1,0)}{\theta_1^2} = \lim_{\theta_1\to 0} \frac{4 - (\cos\theta_1 + 1) - 2\cos\theta_1 - \sqrt{(\cos\theta_1 - 1)^2}}{\theta_1^2} = \lim_{\theta_1\to 0} \frac{2 - 2\cos\theta_1}{\theta_1^2} = 1. \tag{5.11}$$

Similarly, for a fixed $\theta_1 = 0$,

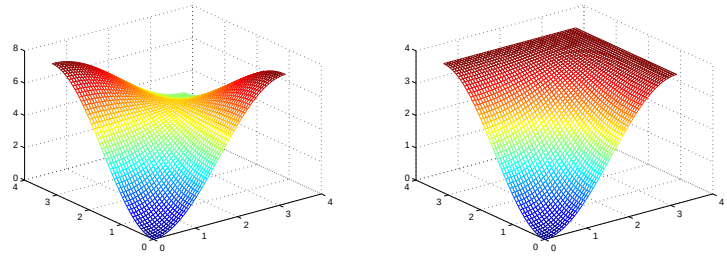
$$\lim_{\theta_2\to 0} \frac{\lambda_1(0,\theta_2)}{\theta_2^2} = \lim_{\theta_2\to 0} \frac{4 - (\cos\theta_2 + 1) - 2\cos\theta_2 + \sqrt{(\cos\theta_2 - 1)^2}}{\theta_2^2} = \lim_{\theta_2\to 0} \frac{4 - 4\cos\theta_2}{\theta_2^2} = 2,$$
$$\lim_{\theta_2\to 0} \frac{\lambda_2(0,\theta_2)}{\theta_2^2} = \lim_{\theta_2\to 0} \frac{4 - (\cos\theta_2 + 1) - 2\cos\theta_2 - \sqrt{(\cos\theta_2 - 1)^2}}{\theta_2^2} = \lim_{\theta_2\to 0} \frac{2 - 2\cos\theta_2}{\theta_2^2} = 1. \tag{5.12}$$

Due to (5.11) and (5.12), we expect the eigenvalues of the generating function at the coarse level to have a zero of order 2 at the origin and therefore we define the projectors as

$$P_{N_i}^{[2]} = T_{N_i}(p)(K_{N_i}^{[2]})^T, \quad K_{N_i}^{[2]} = K_{(N_i)_1} \otimes K_{(N_i)_2} \otimes I_2,$$

where

$$p(\theta_1,\theta_2) = 4\prod_{j=1}^{2}[1+\cos(\theta_j)] \cdot I_2 = \begin{bmatrix} (2+2\cos\theta_1)(2+2\cos\theta_2) & 0 \\ 0 & (2+2\cos\theta_1)(2+2\cos\theta_2) \end{bmatrix}.$$



(a) $\lambda_1(\theta_1,\theta_2)$, $(\theta_1,\theta_2) \in [0,\pi)^2$ (b) $\lambda_2(\theta_1,\theta_2)$, $(\theta_1,\theta_2) \in [0,\pi)^2$

Note that the restriction of a vector from a fine to a coarser grid is obtained by product with the matrix $(P_{N_i}^{[2]})^T = K_{N_i}^{[2]} T_{N_i}(p)$, where $T_{N_i}(p)$ has at most nine nonzero entries in every row and $K_{N_i}^{[2]}$ simply performs a proper down sampling of two entries every four, due to the tensor with I_2. In detail, let z be defined on the fine grid and let y be its projection into the coarser grid obtained by applying $(P_{N_i}^{[2]})^T$, then the entries of y arranged as a 2D array can be computed as

$$y_{i,j} = z_{2i-2,2j+2} + z_{2i-2,2j-2} + 2z_{2i,2j-2} + 2z_{2i-2,2j} + 4z_{2i,2j} + 2z_{2i+2,2j} + 2z_{2i,2j+2} + z_{2i+2,2j+2} + z_{2i+2,2j-2}.$$

Similarly, the prolongation of a vector from a coarse to the finer grid is obtained by product with the matrix $P_{N_i}^{[2]} = T_{N_i}(p)(K_{N_i}^{[2]})^T$, where $(K_{N_i}^{[2]})^T$ adds the zeros corresponding to the new grid points and $T_{N_i}(p)$ performs the average of nine "near" points: again there is a jump of size 2 because of the tensor by I_2 in the formula of $K_{N_i}^{[2]}$. From a geometrical point of view, the grid transfer operator is the standard bilinear interpolation associated to the linear tensor B-spline [49], again tensored with I_2. The coarser matrices are computed by the Galerkin approach as $A_{i+1} = (P_{N_i}^{[2]})^T A_i P_{N_i}^{[2]}$, for $i = 0,\dots,m-1$ in a setup phase. According to the results in [4, 1], the matrix A_{i+1} inherits the same 2-level block Toeplitz structure with 2×2 blocks of A_i. In particular, thanks to (1.47), the generating function of A_{i+1} is again $f(\theta_1,\theta_2)$ defined in (5.4) up to a factor 16. Hence, A_{i+1} could be simply computed by rediscretization with double step size scaling properly the projector.

Regarding the smoother, because of its simplicity, just as in the scalar-case, we opt for the weighted Jacobi method. Note that since the constant coefficient of f_{11} and f_{22} is the same and equal to 4, then weighted Jacobi iteration is simply the weighted Richardson iteration considered in Proposition 19. Hence, according to Proposition 19 and scaling the weight by a factor 4, we have that $0 \le \omega, \tilde{\omega} \le 8/\max_{j=1,2}\|\lambda_j\|_{L^\infty}$. Computing

$$\max_{j=1,2}\|\lambda_j\|_{L^\infty} = \|\lambda_1\|_{L^\infty} = \lambda_1(0,\pi) = 8,$$

we have that $0 \le \omega, \tilde{\omega} \le 1$. Therefore, following the multi-iterative idea (see [120]), we use both pre-smoother and post-smoother with parameters $\omega = 1$ and $\tilde{\omega} = 2/3$, respectively. Moreover, as shown in the next section, we test also the performances of our method choosing Gauss-Seidel method as smoother.

5.4 Numerical results

In this section we use the block multigrid defined in Section 5.3 (denoted as GLT-MG) as a preconditioner for Krylov methods in order to solve the linear system $Ku = b$, with K as in (5.3) and $n_1 = n_2 = 2^t + 1$.

To be precise, we solve the $2(2^t+1)^2 \times 2(2^t+1)^2$ linear system associated to the matrix $\Pi K \Pi^T$, with Π defined as in (5.6). In the following we refer to this permuted matrix again as K.

Although our analysis has been focused on symmetric positive definite matrices, in the following we test GLT-MG also on the nonsymmetric matrices coming from the discretization of the original problem without neglecting the advection term. We choose the Preconditioned Conjugate Gradient (PCG) method for symmetric matrices and the GMRES method for the case where the matrices are nonsymmetric. Both methods are applied using the built-in Matlab functions `pcg` and `gmres`. Moreover, for all involved methods (GLT-MG, PCG and GMRES), we choose as initial guess the zero vector and use a relative stopping criterion $tol \in \{10^{-3}, 10^{-6}\}$.

tol	$n_1 = n_2$	GLT-MG$_{\text{J}}$		GLT-MG$_{\text{GS}}$		AGMG	
		Iter	Error	Iter	Error	Iter	Error
10^{-3}	2^5+1	4	4.27e-004	3	2.09e-004	6	1.66e-003
	2^6+1	4	4.24e-004	3	2.17e-004	6	3.19e-003
	2^7+1	4	4.56e-004	3	2.25e-004	6	2.51e-003
10^{-6}	2^5+1	8	8.09e-007	5	5.95e-007	11	8.87e-007
	2^6+1	8	4.47e-007	5	5.73e-007	12	1.88e-006
	2^7+1	8	4.60e-007	5	5.77e-007	12	3.73e-006

Table 5.1: Symmetric case - Comparison between GLT-MG$_{\text{J,GS}}$ (used as preconditioner for PCG) and AGMG method both in terms of iterations and approximation error fixed $tol = 10^{-3}, 10^{-6}$.

tol	$n_1 = n_2$	GLT-MG$_{\text{J}}$		GLT-MG$_{\text{GS}}$		AGMG	
		Iter	Error	Iter	Error	Iter	Error
10^{-3}	2^5+1	4	4.41e-004	3	9.76e-005	7	4.15e-003
	2^6+1	4	3.23e-004	3	8.55e-005	7	3.74e-003
	2^7+1	4	2.92e-004	3	7.77e-005	6	2.78e-003
	2^8+1	4	2.87e-004	3	7.23e-005	6	2.51e-003
10^{-6}	2^5+1	8	2.45e-007	5	3.28e-007	16	3.71e-006
	2^6+1	8	2.08e-007	5	2.74e-007	17	3.84e-006
	2^7+1	8	1.60e-007	5	2.61e-007	16	7.04e-006
	2^8+1	8	1.53e-007	5	2.47e-007	15	7.75e-006

Table 5.2: Nonsymmetric case - Comparison between GLT-MG$_{\text{J,GS}}$ (used as preconditioner for GMRES) and AGMG method both in terms of iterations and approximation error fixed $tol = 10^{-3}, 10^{-6}$.

Recall that the GLT-MG uses weighted Jacobi (in this case we label our method as 'GLT-MG$_{\text{J}}$') or Gauss-Seidel (in this case we label our method as 'GLT-MG$_{\text{GS}}$') as pre-smoothers and post-smoothers. The parameters of Jacobi method are $\omega = 1$ and $\tilde{\omega} = 2/3$. In our test we make only one pre- and post-smoothing steps and perform only one V-cycle.

Assuming to know the true solution $\widetilde{u}$, we compute the relative error of the approximated solution u as the measure of accuracy.

In [66] the authors solve $Ku = b$ with an aggregation-based algebraic multigrid (AGMG), see [110]. AGMG performs one forward and one backward Gauss-Seidel sweep for pre- and post-smoothing, respectively, and performs also a K-cycle, i.e., two Krylov accelerated iterations at each intermediate level. The main iterative solver in AGMG is the Generalized Conjugate Residual method.

We check the validity of the strategy proposed in this chapter and compare it with AGMG both in terms of iterations and of approximation error.

We choose $\widetilde{u}$ as an equispaced sampling of the function

$$\varphi(x_1, x_2) = \sin(3x_1) + \sin(3x_2), \qquad (x_1, x_2) \in \Omega$$

and as right-hand side $b = K\widetilde{u}$. The numerical tests with GLT-MG are performed in Matlab and AGMG is used via its Matlab interface.

Table 5.1 shows a comparison between of GLT-MG$_{\text{J}}$ and GLT-MG$_{\text{GS}}$ (used as preconditioners for PCG) and AGMG in terms of iterations and approximation error, in the case when K is symmetric. We

observe that for both stopping criteria the GLT-MG method requires less iterations than the AGMG method, especially when the Gauss-Seidel smoother is used. Moreover, the accuracy of the computed iterative solution, achieved when using GLT-MG is better than that provided by AGMG. Note that for both methods the number of iterations does not increase when increasing the size of the problem, which means that both GLT-MG and AGMG are of optimal order. In Table 5.2 we show the number of iterations and accuracy achieved by GLT-MG$_{\mathrm{J}}$, GLT-MG$_{\mathrm{GS}}$ (used as preconditioners for GMRES) and AGMG when K is a nonsymmetric matrix. Even in this case GLT-MG converges with a fewer number of iterations compared to AGMG and the resulting iteratively computed solution for the considered test problem is more accurate.

Conclusions

This thesis can be seen as a byproduct of the combined use of powerful tools like symbol, spectral distribution, and GLT, when dealing with the numerical solution of structured linear systems. We approached such an issue both from a theoretical and practical viewpoint. Chapter 2 is the 'theoretical core' of this work and contains new distribution results for combinations of some algebraic operations on non-Hermitian multilevel block Toeplitz matrices. The remaining Chapters 3-5 have a more 'applicative taste', being focused on structured linear systems arising from the following three applications: image deblurring, FDEs, coupled PDEs. In the following we summarize our main results with a look to some open problems.

The new tools introduced in Chapter 2 have been used for defining a band Toeplitz preconditioner for Krylov methods in Section 2.1 and for building up a fast PHSS method for Toeplitz matrices in Section 2.2. As regards future developments, the encouraging numerical results provided by the band Toeplitz preconditioning show that there is room both for relaxing the hypotheses of Theorem 13 and for providing a more complete picture on the spectral localization and the number of outliers. Moreover, due to the satisfactory performances of PHSS method when applied to the convection diffusion equation with constant coefficients defined on the unitary square (see Case 11), it would be interesting to further investigate this type of equations by considering nonconstant coefficients, general domains and nonuniform gridding. Once again, in such a setting the most promising tool is the theory of GLT sequences.

Chapter 3 was devoted to image deblurring. In Section 3.1 we have considered iterative methods for image restoration and we have proposed a regularization preconditioning technique, which preserves the structure of the blurring matrix. The presented preconditioning strategy represents a generalization and an improvement both with respect to circulant [45, 82] and structured preconditioning available in the literature [47, 81, 109]. About future research lines, we mention that, in order to preserve edges or to enforce sparsity in a certain basis (usually wavelet) on the restored image, regularization terms that lead to non-linear problems are usually employed. Nevertheless, the resulting numerical methods usually require the solution of a regularized least square method, e.g. the linearized Bregman splitting [29, 86]. Therefore, improvements in classical least square methods can be useful also for these more computationally challenging models as shown in [30] where, among other strategies, the reblurring preconditioner and its non-stationary variant have been adapted to be included in the linearized Bregman splitting for the synthesis approach proposed in [29]. The proposed structure preserving preconditioners could be similarly considered to improve the performances of such numerical methods.

In Section 3.2 exploiting the sparsity of the Fourier coefficients, in order to reduce the oversmoothing effects of the Tikhonov regularization, we introduced a diagonal weighting matrix in the Fourier domain. The arising diagonal linear system allows very fast computations and the regularization parameter can be efficiently estimated by the GCV. A future issue could be to investigate the treatment of the boundary artifacts by imposing appropriate boundary conditions, like reflective, antireflective, or other boundary conditions which lead to a matrix that can be diagonalized by fast transforms [109, 132, 2, 50]. Another strategy to deal with the discrete Fourier transform with reduced boundary artifacts is based on the enlargement of the domain [115, 9].

In Chapter 4 we performed a spectral analysis of the FDE problem in the case of variable coefficients showing that the coefficient matrix-sequence belongs to the GLT class. We used the spectral information for analyzing known methods of preconditioned Krylov and multigrid type and for identifying two numerical effective tridiagonal structure preserving preconditioners. Several issues could be the subjects of future investigation. On the one hand, we aim to furnish a detailed analysis of the problem in the multidimensional setting, taking in mind that, according to the preliminary comments in Section 4.3, the only promising technique even in the case of nonconstant and different diffusion coefficients (provided that they are uniformly bounded and positive) seems to be the use of appropriate multigrid strate-

gies. On the other hand, we will focus on the spectral analysis of the FDE problem when alternative discretizations like radial basis functions methods [114] are used.

In Chapter 5 we derived a very efficient multigrid preconditioner for matrices originating from FEM discretization of a coupled system of PDEs and we illustrated the technique on a 2D linear elasticity problem, discretized using Q1 FEM elements. We aim to make detailed analysis of the proposed multigrid in terms of convergence and optimality. Furthermore, we plan to extend this strategy to linear systems arising from other finite element discretizations, from the isogeometric analysis, and with special attention to the three-dimensional case due to its importance in real-world applications such as the Glacial Isostatic Adjustment model (see [101] and references therein).

Bibliography

[1] A. Aricò and M. Donatelli. AV-cycle multigrid for multilevel matrix algebras: proof of optimality. *Numerische Mathematik*, 105(4):511–547, 2007.

[2] A. Aricò, M. Donatelli, J. Nagy, and S. Serra-Capizzano. The Anti-Reflective Transform and Regularization by Filtering. In *Numerical Linear Algebra in Signals, Systems and Control*, volume 80, pages 1–21. Springer, 2011.

[3] A. Aricò, M. Donatelli, and S. Serra-Capizzano. Spectral analysis of the anti-reflective algebra. *Linear Algebra and its Applications*, 428(2):657–675, 2008.

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

[5] O. Axelsson and V. A. Barker. *Finite element solution of boundary value problems: theory and computation*, volume 35. SIAM, 2001.

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

[7] O. Axelsson and M. Neytcheva. The algebraic multilevel iteration methods: theory and applications. In D. D. Bainov and V. Covachev, editors, *Proceedings of the 2nd International Colloquium on Numerical Analysis, Plovdiv, Bulgaria (August 1993)*, pages 13–23. Katholieke Universiteit Nijmegen. Mathematisch Instituut, 1993.

[8] J. Bai and X.-C. Feng. Fractional-order anisotropic diffusion for image denoising. *Image Processing, IEEE Transactions on*, 16(10):2492–2502, 2007.

[9] Z.-J. Bai, M. Donatelli, and S. Serra-Capizzano. Fast Preconditioners for Total Variation Deblurring with Antireflective Boundary Conditions. *SIAM Journal on Matrix Analysis and Applications*, 32(3):785–805, 2011.

[10] Z.-Z. Bai and G. H. Golub. Accelerated Hermitian and skew-Hermitian splitting iteration methods for saddle-point problems. *IMA Journal of Numerical Analysis*, 27(1):1–23, 2007.

[11] Z.-Z. Bai, G. H. Golub, and M. K. Ng. Hermitian and skew-Hermitian splitting methods for non-Hermitian positive definite linear systems. *SIAM Journal on Matrix Analysis and Applications*, 24(3):603–626, 2003.

[12] Z.-Z. Bai, G. H. Golub, and J.-Y. Pan. Preconditioned Hermitian and skew-Hermitian splitting methods for non-Hermitian positive semidefinite linear systems. *Numerische Mathematik*, 98(1):1–32, 2004.

[13] C. Baker, L. Fox, D. Mayers, and K. Wright. Numerical solution of Fredholm integral equations of first kind. *The Computer Journal*, 7(2):141–148, 1964.

[14] E. Bängtsson and M. Neytcheva. Numerical simulations of glacial rebound using preconditioned iterative solution methods. *Applications of Mathematics*, 50(3):183–201, 2005.

[15] B. Beckermann and S. Serra-Capizzano. On the asymptotic spectrum of Finite Element matrix sequences. *SIAM Journal on Numerical Analysis*, 45(2):746–769, 2007.

[16] M. Benzi. A generalization of the Hermitian and skew-Hermitian splitting iteration. *SIAM Journal on Matrix Analysis and Applications*, 31(2):360–374, 2009.

[17] D. Bertaccini, G. H. Golub, S. Serra-Capizzano, and C. Tablino Possio. Preconditioned HSS methods for the solution of non-Hermitian positive definite linear systems and applications to the discrete convection-diffusion equation. *Numerische Mathematik*, 99(3):441–484, 2005.

[18] M. Bertero and P. Boccacci. *Introduction to inverse problems in imaging*. CRC press, 1998.

[19] R. Bhatia. *Matrix Analysis*, volume 169. Springer-Verlag, New York, 1997.

[20] D. A. Bini and B. Meini. Effective methods for solving banded Toeplitz systems. *SIAM Journal on Matrix Analysis and Applications*, 20(3):700–719, 1999.

[21] A. Björck. *Numerical Methods for Least Squares Problems*. SIAM, Philadephia, 1996.

[22] M. Bolten, M. Donatelli, and T. Huckle. Analysis of smoothed aggregation multigrid methods based on Toeplitz matrices. *Electronic Transactions on Numerical Analysis*, 44:25–52, 2015.

[23] A. Böttcher and S. M. Grudsky. *Spectral properties of banded Toeplitz matrices*. SIAM, 2005.

[24] D. Braess. *Finite elements: Theory, Fast Solvers, and applications in Solid Mechanics (2nd edn)*. Cambridge University Press, Cambridge, 2001.

[25] A. Brandt. Rigorous quantitative analysis of multigrid, I: Constant coefficients two-level cycle with l_2-norm. *SIAM Journal on Numerical Analysis*, 31(6):1695–1730, 1994.

[26] T. Breiten, V. Simoncini, and M. Stoll. Low-rank solvers for fractional differential equations, 2015. Manuscript, http://www.mpi-magdeburg.mpg.de/2956732/BSS15.pdf.

[27] H. Brezis. *Functional analysis, Sobolev spaces and partial differential equations*. Springer, New York, 2011.

[28] J.-F. Cai, R. H. Chan, and Z. Shen. A framelet-based image inpainting algorithm. *Applied and Computational Harmonic Analysis*, 24(2):131–149, 2008.

[29] J.-F. Cai, S. Osher, and Z. Shen. Linearized Bregman iterations for frame-based image deblurring. *SIAM Journal on Imaging Sciences*, 2(1):226–252, 2009.

[30] Y. Cai, M. Donatelli, D. Bianchi, and T.-Z. Huang. Regularization preconditioners for frame-based image deblurring with reduced boundary artifacts. *SIAM Journal on Scientific Computing*. In press.

[31] B. Carreras, V. Lynch, and G. Zaslavsky. Anomalous diffusion and exit time distribution of particle tracers in plasma turbulence model. *Physics of Plasmas*, 8(12):5096–5103, 2001.

[32] R. H. Chan. Toeplitz preconditioners for Toeplitz systems with nonnegative generating functions. *IMA Journal of Numerical Analysis*, 11(3):333–345, 1991.

[33] R. H. Chan, Q.-S. Chang, and H.-W. Sun. Multigrid method for ill-conditioned symmetric Toeplitz systems. *SIAM Journal on Scientific Computing*, 19(2):516–529, 1998.

[34] R. H. Chan and W.-K. Ching. Toeplitz-circulant preconditioners for Toeplitz systems and their applications to queueing networks with batch arrivals. *SIAM Journal on Scientific Computing*, 17(3):762–772, 1996.

[35] R. H. Chan and X.-Q. Jin. A family of block preconditioners for block systems. *SIAM Journal on Scientific Computing*, 13(5):1218–1235, 1992.

[36] R. H. Chan and K.-P. Ng. Toeplitz preconditioners for Hermitian Toeplitz systems. *Linear Algebra and its Applications*, 190:181–208, 1993.

[37] R. H. Chan and M. K. Ng. Conjugate gradient methods for Toeplitz systems. *SIAM Review*, 38(3):427–482, 1996.

[38] R. H. Chan and G. Strang. Toeplitz equations by conjugate gradients with circulant preconditioner. *SIAM Journal on Scientific Computing*, 10(1):104–119, 1989.

[39] R. H. Chan and P. T. P. Tang. Fast band-Toeplitz preconditioners for Hermitian Toeplitz systems. *SIAM Journal on Scientific Computing*, 15(1):164–171, 1994.

[40] T. F. Chan. An optimal circulant preconditioner for Toeplitz systems. *SIAM Journal on Scientific Computing*, 9(4):766–771, 1988.

[41] T. F. Chan and J. A. Olkin. Circulant preconditioners for Toeplitz-block matrices. *Numerical Algorithms*, 6(1):89–101, 1994.

[42] J. M. Chung, M. E. Kilmer, and D. P. O'Leary. A Framework for Regularization via Operator Approximation. *SIAM Journal on Scientific Computing*, 37(2):B332–B359, 2015.

[43] E. Cinlar. *Introduction to stochastic processes*. Prentice-Hall, New Jersey, US, 1975.

[44] V. Del Prete, F. Di Benedetto, M. Donatelli, and S. Serra-Capizzano. Symbol approach in a signal-restoration problem involving block Toeplitz matrices. *Journal of Computational and Applied Mathematics*, 272:399–416, 2014.

[45] P. Dell'Acqua, M. Donatelli, and C. Estatico. Preconditioners for image restoration by reblurring techniques. *Journal of Computational and Applied Mathematics*, 272:313–333, 2014.

[46] P. Dell'Acqua, M. Donatelli, C. Estatico, and M. Mazza. Structured preserving reblurring preconditioners for image deblurring. 2015. Submitted.

[47] P. Dell'Acqua, M. Donatelli, S. Serra-Capizzano, D. Sesana, and C. Tablino-Possio. Optimal preconditioning for image deblurring with anti-reflective boundary conditions. *Linear Algebra and its Applications*, 2015. In press.

[48] F. Di Benedetto, G. Fiorentino, and S. Serra. C.G. Preconditioning for Toeplitz matrices. *Computers & Mathematics with Applications*, 25(6):35–45, 1993.

[49] M. Donatelli. An algebraic generalization of local Fourier analysis for grid transfer operators in multigrid based on Toeplitz matrices. *Numerical Linear Algebra with Applications*, 17(2-3):179–197, 2010.

[50] M. Donatelli. Fast transforms for high order boundary conditions in deconvolution problems. *BIT Numerical Mathematics*, 50(3):559–576, 2010.

[51] M. Donatelli, A. Dorostkar, M. Mazza, M. Neytcheva, and S. Serra-Capizzano. A block multigrid strategy for two-dimensional coupled PDEs. Technical Report 2016-001, Department of Information Technology, Uppsala University, 2016.

[52] M. Donatelli, C. Garoni, C. Manni, S. Serra-Capizzano, and H. Speleers. Spectral analysis and spectral symbol of matrices in isogeometric collocation methods. *Mathematics of Computation*, 2015. DOI: http://dx.doi.org/10.1090/mcom/3027.

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

[54] M. Donatelli, C. Garoni, M. Mazza, S. Serra-Capizzano, and D. Sesana. Preconditioned HSS method for large multilevel block Toeplitz linear systems via the notion of matrix-valued symbol. *Numerical Linear Algebra with Applications*, 23(1):83–119, 2016. DOI: 10.1002/nla.2007.

[55] M. Donatelli and M. Hanke. Fast nonstationary preconditioned iterative methods for ill-posed problems, with application to image deblurring. *Inverse Problems*, 29(9):095008, 2013.

[56] M. Donatelli, T. Huckle, M. Mazza, and D. Sesana. Image deblurring by sparsity constraint on the Fourier coefficients. *Numerical Algorithms*, 2015. DOI: 10.1007/s11075-015-0047-x.

[57] M. Donatelli, M. Mazza, and S. Serra-Capizzano. Spectral analysis and structure preserving preconditioners for fractional diffusion equations. *Journal of Computational Physics*, 307:262–279, 2016. DOI: 10.1016/j.jcp.2015.11.061.

[58] M. Donatelli, M. Molteni, V. Pennati, and S. Serra-Capizzano. Multigrid methods for cubic spline solution of two point (and 2D) boundary value problems. *Applied Numerical Mathematics*, 2014. DOI: 10.1016/j.apnum.2014.04.004.

[59] M. Donatelli, A. Neuman, and L. Reichel. Square regularization matrices for large linear discrete ill-posed problems. *Numerical Linear Algebra with Applications*, 19(6):896–913, 2012.

[60] M. Donatelli, M. Neytcheva, and S. Serra-Capizzano. Canonical eigenvalue distribution of multilevel block Toeplitz sequences with non-Hermitian symbols. *Operator Theory: Advances and Applications*, 221:269–291, 2012.

[61] M. Donatelli and L. Reichel. Square smoothing regularization matrices with accurate boundary conditions. *Journal of Computational and Applied Mathematics*, 272:334–349, 2014.

[62] M. Donatelli and S. Serra-Capizzano. Anti-reflective boundary conditions and re-blurring. *Inverse Problems*, 21(1):169–182, 2005.

[63] M. Donatelli and S. Serra-Capizzano. On the regularizing power of multigrid-type algorithms. *SIAM Journal on Scientific Computing*, 27(6):2053–2076, 2006.

[64] M. Donatelli and S. Serra-Capizzano. Antireflective Boundary Conditions for Deblurring Problems. *Journal of Electrical and Computer Engineering*, 2010:2, 2010.

[65] M. Donatelli, S. Serra-Capizzano, and D. Sesana. Multigrid methods for Toeplitz linear systems with different size reduction. *BIT Numerical Mathematics*, 52(2):305–327, 2012.

[66] A. Dorostkar, M. Neytcheva, and B. Lund. Numerical and computational aspects of some block-preconditioners for saddle point systems. *Parallel Computing*, 49:164–178, 2015.

[67] A. Dorostkar, M. Neytcheva, and S. Serra-Capizzano. Spectral analysis of coupled PDEs and of their Schur complements via the notion of Generalized Locally Toeplitz sequences. Technical report, Technical Report 2015-008, Department of Information Technology, Uppsala University, 2015.

[68] H. W. Engl, M. Hanke, and A. Neubauer. *Regularization of Inverse Problems*, volume 375. Kluwer Academic Publishers, Dordrecht, 1996.

[69] C. Estatico. A classification scheme for regularizing preconditioners, with application to Toeplitz systems. *Linear Algebra and its Applications*, 397:107–131, 2005.

[70] M. A. T. Figueiredo and R. D. Nowak. An EM algorithm for wavelet-based image restoration. *Image Processing, IEEE Transactions on*, 12(8):906–916, 2003.

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

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

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

[74] C. Garoni and S. Serra-Capizzano. Generalized Locally Toeplitz sequences: a review and an extension. *Springer*. Under revision. Technical report/Department of Information Technology, Uppsala University, ISSN 1404-3203; 2015-016.

[75] C. Garoni, S. Serra-Capizzano, and D. Sesana. Spectral analysis and spectral symbol of d-variate $\mathbb{Q}_p$ Lagrangian FEM stiffness matrices. *SIAM Journal on Matrix Analysis and Applications*, 36(3):1100–1128, 2015.

[76] S. Gazzola and J. Nagy. Generalized Arnoldi–Tikhonov method for sparse reconstruction. *SIAM Journal on Scientific Computing*, 36(2):B225–B247, 2014.

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

[78] R. M. Gray. Toeplitz and circulant matrices: A review. *Foundations and Trends in Communications and Information Theory*, 2(3):155–239, 2006.

[79] U. Grenander and G. Szegö. *Toeplitz Forms and Their Applications*, volume 321. Second Edition, Chelsea, New York, 1984.

[80] M. Hanke. *Conjugate gradient type methods for ill-posed problems*, volume 327. CRC Press, 1995.

[81] M. Hanke and J. Nagy. Inverse Toeplitz preconditioners for ill-posed problems. *Linear Algebra and its Applications*, 284(1):137–156, 1998.

[82] M. Hanke, J. Nagy, and R. Plemmons. Preconditioned iterative regularization for ill-posed problems. *Numerical Linear Algebra*, pages 141–163, 1993. Proceedings of the Conference in Numerical Linear Algebra and Scientific Computation, Kent, Ohio, March 13-14 1992.

[83] P. C. Hansen. Regularization Tools: A Matlab package for analysis and solution of discrete ill-posed problems. *Numerical Algorithms*, 6(1):1–35, 1994.

[84] P. C. Hansen. *Rank-deficient and discrete ill-posed problems: numerical aspects of linear inversion*, volume 4. SIAM, Philadelphia, 1998.

[85] P. C. Hansen, J. Nagy, and D. P. O'Leary. *Deblurring Images: Matrices, Spectra, and Filtering*, volume 3. SIAM, Philadelphia, 2006.

[86] J. Huang, M. Donatelli, and R. H. Chan. Nonstationary Iterated Thresholding Algorithms for Image Deblurring. *Inverse Problems and Imaging*, 7(3):717–736, 2013.

[87] T. Huckle. Compact Fourier analysis for designing multigrid methods. *SIAM Journal on Scientific Computing*, 31(1):644–666, 2008.

[88] T. Huckle and C. Kravvaritis. Compact Fourier analysis for multigrid methods based on block symbols. *SIAM Journal on Matrix Analysis and Applications*, 33(1):73–96, 2012.

[89] T. Huckle and D. Noutsos. Preconditioning block Toeplitz matrices. *Electronic Transactions on Numerical Analysis*, 29:31–45, 2007/08.

[90] T. Huckle and M. Sedlacek. Smoothing and Regularization with Modified Sparse Approximate Inverses. *Journal of Electrical and Computer Engineering*, 2010:16 pages, 2010.

[91] T. Huckle and M. Sedlacek. Data based regularization matrices for the Tikhonov-Phillips regularization. *Proceedings in Applied Mathematics and Mechanics*, 12(1):643–644, 2012.

[92] T. Huckle and M. Sedlacek. Tikhonov–Phillips regularization with operator dependent seminorms. *Numerical Algorithms*, 60(2):339–353, 2012.

[93] T. Huckle and M. Sedlacek. Data based regularization for discrete deconvolution problems. *BIT Numerical Mathematics*, 53(2):459–473, 2013.

[94] T. Huckle, S. Serra-Capizzano, and C. Tablino-Possio. Preconditioning strategies for non-Hermitian Toeplitz linear systems. *Numerical Linear Algebra with Applications*, 12(2-3):211–220, 2005.

[95] T. Huckle and J. Staudacher. Multigrid preconditioning and Toeplitz matrices. *Electronic Transactions on Numerical Analysis*, 13:81–105, 2002.

[96] T. Huckle and J. Staudacher. Multigrid methods for block Toeplitz matrices with small size blocks. *BIT Numerical Mathematics*, 46(1):61–83, 2006.

[97] T. Kailath and A. H. Sayed. Displacement structure: theory and applications. *SIAM Review*, 37(3):297–386, 1995.

[98] P. P. Korovkin. *Linear operators and approximation theory.* Hindustan Publishing Corp, 1960.

[99] T. Ku and C.-C. J. Kuo. On the spectrum of a family of preconditioned block Toeplitz matrices. *SIAM Journal on Scientific Computing*, 13(4):948–966, 1992.

[100] S.-L. Lei and H.-W. Sun. A circulant preconditioner for fractional diffusion equations. *Journal of Computational Physics*, 242:715–725, 2013.

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

[102] S. Mallat. *A wavelet tour of signal processing*. Academic Press, California, 1999.

[103] M. M. Meerschaert and C. Tadjeran. Finite difference approximations for fractional advection–dispersion flow equations. *Journal of Computational and Applied Mathematics*, 172(1):65–77, 2004.

[104] M. M. Meerschaert and C. Tadjeran. Finite difference approximations for two-sided space-fractional partial differential equations. *Applied Numerical Mathematics*, 56(1):80–90, 2006.

[105] N. M. Nachtigal, S. C. Reddy, and L. N. Trefethen. How fast are nonsymmetric matrix iterations? *SIAM Journal on Matrix Analysis and Applications*, 13(3):778–795, 1992.

[106] J. Nagy, K. Palmer, and L. Perrone. RestoreTools: an object oriented MATLAB package for image restoration. 2002. http://www. mathcs. emory. edu/˜ nagy.

[107] M. F. Neuts. *Structured stochastic matrices of M/G/1 type and their applications*. Dekker, New York, 1989.

[108] M. Neytcheva. On element-by-element Schur complement approximations. *Linear Algebra and Its Applications, special Issue: Devoted to the 2nd NASC 08 Conference in Nanjing (NSC)*, 434(11):2308–2324, 2011.

[109] M. K. Ng, R. H. Chan, and W.-C. Tang. A fast algorithm for deblurring models with Neumann boundary conditions. *SIAM Journal on Scientific Computing*, 21(3):851–866, 1999.

[110] Y. Notay. AGMG, http://homepages.ulb.ac.be/ ynotay/AGMG/.

[111] D. Noutsos, S. Serra-Capizzano, and P. Vassalos. Matrix algebra preconditioners for multilevel Toeplitz systems do not insure optimal convergence rate. *Theoretical Computer Science*, 315(2):557–579, 2004.

[112] J. Pan, R. Ke, M. K. Ng, and H.-W. Sun. Preconditioning Techniques for Diagonal-Times-Toeplitz Matrices in Fractional Diffusion Equations. *SIAM Journal on Scientific Computing*, 36(6):A2698–A2719, 2014.

[113] H.-K. Pang and H.-W. Sun. Multigrid method for fractional diffusion equations. *Journal of Computational Physics*, 231(2):693–703, 2012.

[114] C. Piret and E. Hanert. A radial basis functions method for fractional diffusion equations. *Journal of Computational Physics*, 238:71 – 81, 2013.

[115] S. J. Reeves. Fast image restoration without boundary artifacts. *Image Processing, IEEE Transactions on*, 14(10):1448–1453, 2005.

[116] L. I. Rudin, S. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, 60(1):259–268, 1992.

[117] J. Ruge and K. Stüben. *Algebraic multigrid*. in Frontiers in Applied Mathematics: Multigrid Methods, S. McCormick Ed. SIAM, Philadelphia, 1987.

[118] Y. Saad. *Iterative Methods for Sparse Linear Systems*. PWS Publishing, Boston, MA, 2003.

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

[120] S. Serra-Capizzano. Multi-iterative methods. *Computers & Mathematics with Applications*, 26(4):65–87, 1993.

[121] S. Serra-Capizzano. Preconditioning strategies for asymptotically ill-conditioned block Toeplitz systems. *BIT Numerical Mathematics*, 34(4):579–594, 1994.

[122] S. Serra-Capizzano. New PCG based algorithms for the solution of Hermitian Toeplitz systems. *CALCOLO*, 32(3-4):153–176, 1995.

[123] S. Serra-Capizzano. Optimal, quasi-optimal and superlinear band-Toeplitz preconditioners for asymptotically ill-conditioned positive definite Toeplitz systems. *Mathematics of Computation*, 66(218):651–665, 1997.

[124] S. Serra-Capizzano. Asymptotic results on the spectra of block Toeplitz preconditioned matrices. *SIAM Journal on Matrix Analysis and Applications*, 20(1):31–44, 1998.

[125] S. Serra-Capizzano. A Korovkin-based approximation of multilevel Toeplitz matrices (with rectangular unstructured blocks) via multilevel trigonometric matrix spaces. *SIAM Journal on Numerical Analysis*, 36(6):1831–1857, 1999.

[126] S. Serra-Capizzano. Spectral and computational analysis of block Toeplitz matrices having nonnegative definite matrix-valued generating functions. *BIT Numerical Mathematics*, 39(1):152–175, 1999.

[127] S. Serra-Capizzano. Spectral behavior of matrix sequences and discretized boundary value problems. *Linear Algebra and its Applications*, 337(1):37–78, 2001.

[128] S. Serra-Capizzano. Convergence analysis of Two-Grid methods for elliptic Toeplitz and PDEs matrix-sequences. *Numerische mathematik*, 92(3):433–465, 2002.

[129] S. Serra-Capizzano. Matrix algebra preconditioners for multilevel Toeplitz matrices are not superlinear. *Linear Algebra and its Applications*, 343:303–319, 2002.

[130] S. Serra-Capizzano. More inequalities and asymptotics for matrix valued linear positive operators: the noncommutative case. *Operator Theory: Advances and Applications*, 135:293–315, 2002.

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

[132] S. Serra-Capizzano. A note on antireflective boundary conditions and fast deblurring models. *SIAM Journal on Scientific Computing*, 25(4):1307–1325, 2004.

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

[134] S. Serra-Capizzano, D. Sesana, and E. Strouse. The eigenvalue distribution of products of Toeplitz matrices–clustering and attraction. *Linear Algebra and its Applications*, 432(10):2658–2678, 2010.

[135] S. Serra-Capizzano and P. Tilli. Extreme singular values and eigenvalues of non-Hermitian block Toeplitz matrices. *Journal of Computational and Applied Mathematics*, 108(1):113–130, 1999.

[136] S. Serra Capizzano and E. Tyrtyshnikov. Any circulant-like preconditioner for multilevel Toeplitz matrices is not superlinear. *SIAM Journal on Matrix Analysis and Applications*, 21(2):431–439, 2000.

[137] S. Serra-Capizzano and E. Tyrtyshnikov. How to prove that a preconditioner cannot be superlinear. *Mathematics of Computation*, 72(243):1305–1316, 2003.

[138] M. Shlesinger, B. West, and J. Klafter. Lévy dynamics of enhanced diffusion: Application to turbulence. *Physical Review Letters*, 58(11):1100–1103, 1987.

[139] G. Strang. A Proposal for Toeplitz Matrix Calculations. *Studies in Applied Mathematics*, 74(2):171–176, 1986.

[140] H.-W. Sun, X.-Q. Jin, and Q.-S. Chang. Convergence of the multigrid method of ill-conditioned block Toeplitz systems. *BIT Numerical Mathematics*, 41(1):179–190, 2001.

[141] P. Tilli. Locally Toeplitz sequences: spectral properties and applications. *Linear Algebra and its Applications*, 278(1):91–120, 1998.

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

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

[144] E. E. Tyrtyshnikov. Circulant preconditioners with unbounded inverses. *Linear Algebra and its Applications*, 216:1–23, 1995.

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

[146] E. E. Tyrtyshnikov and N. L. Zamarashkin. Spectra of multilevel Toeplitz matrices: advanced theory via simple matrix relationships. *Linear Algebra and its Applications*, 270(1):15–27, 1998.

[147] C. R. Vogel. *Computational Methods for Inverse Problems*, volume 23. SIAM, Philadelphia, 2002.

[148] H. Wang and T. S. Basu. A fast finite difference method for two-dimensional space-fractional diffusion equations. *SIAM Journal on Scientific Computing*, 34(5):A2444–A2458, 2012.

[149] H. Wang and N. Du. A fast finite difference method for three-dimensional time-dependent space-fractional diffusion equations and its efficient implementation. *Journal of Computational Physics*, 253:50–63, 2013.

[150] H. Wang, K. Wang, and T. Sircar. A direct $O(N \log^2 N)$ finite difference method for fractional diffusion equations. *Journal of Computational Physics*, 229(21):8095–8104, 2010.

[151] K. Wang and H. Wang. A fast characteristic finite difference method for fractional advection–diffusion equations. *Advances in Water Resources*, 34(7):810–816, 2011.

[152] B. Wohlberg and P. Rodriguez. An iteratively reweighted norm algorithm for minimization of total variation functionals. *Signal Processing Letters, IEEE*, 14(12):948–951, 2007.

[153] S. J. Wright, R. D. Nowak, and M. A. T. Figueiredo. Sparse Reconstruction by Separable Approximation. *Signal Processing, IEEE Transactions on*, 57(7):2479–2493, 2009.

[154] P. Wu. Deformation of an incompressible viscoelastic flat earth with Power Law Creep: a Finite Element approach. *Geophysical Journal International*, 108(1):35–51, 1992.

[155] P. Wu. Using commercial finite element packages for the study of earth deformations, sea levels and the state of stress. *Geophysical Journal International*, 158(2):401–408, 2004.

[156] A. Zygmund. *Trigonometric series*, volume 1. Cambridge University Press, 2002.

Acknowledgments

I would like to thank my supervisors Marco Donatelli and Stefano Serra-Capizzano for teaching me an awful lot, without ceasing to encourage me during these three years.

Moreover, a special thank to Carlo Garoni and Debora Sesana for all their help and support.

Many thanks to Pietro Dell'Acqua, Ali Dorostkar, Claudio Estatico, Thomas Huckle, and Maya Neytcheva, for their friendship and collaboration.

I would also like to thank the referees for their remarks, which have improved the quality and the readability of this Ph.D. thesis.

Finally, I wish to thank all the people I met in Como for welcoming me and for making me feel so at home.