UNIVERSITÀ DEGLI STUDI DELL'INSUBRIA

Facoltà di Scienze Matematiche Fisiche e Naturali
Dottorato di Ricerca in Matematica del Calcolo

Spectral Distributions of Structured Matrix-Sequences: Tools and Applications

Supervisor: Prof. Stefano Serra-Capizzano

Ph.D. thesis of:
Debora Sesana
Matr. N. 606860

Anno Accademico 2009-2010

alla mia famiglia

con affetto

Contents

Introduction

Many problems of physics, engineering and applied sciences are governed by functional equations (differential or integral) that do not admit closed-form solution and therefore require numerical discretization techniques which often involve solving large linear systems. However, the coefficient matrix of these systems usually inherits a "structure" from the continuous problem (the properties of the continuous problem moving on the discrete problem) and this information can be conveniently used for solving efficiently the discrete problem. For example in the discretization of the Navier-Stokes indefinite differential equation, we obtain systems of saddle point type also characterized by matrices with an indefinite structure. If the physical problem possesses the property of being invariant in space or time, we get operators with constant coefficients that are invariant under translation (e.g. in partial differential equations (PDEs)) and linear systems we derive from the discretization are of Toeplitz type (for the one-dimensional case); for example in the reconstruction of images where we have specific integral equations (IEs) with kernel invariant under translation.

From the discretization of continuous problems we usually obtain large linear systems where the size of the involved matrices 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 a infinite-dimensional equation (e.g. PDEs, IEs, etc.), one finds a sequence of linear systems $\{A_n x_n = b_n\}$ of increasing dimension d_n. For the resolution of these systems, the direct methods may require a high computation time and also if the coefficient matrix has a particular structure and/or a sparsity these methods generally do not exploit the information on the matrix in order to accelerate the convergence and/or optimize the storage space. Otherwise, the iterative methods, such as multigrid or preconditioned Krylov techniques, are more easily adaptable to problems with specific structural features.

The goal, however, is to choose resolution methods that are optimal.

> **Optimality**: let $\{A_n x_n = b_n\}$ be a sequence of linear systems of increasing dimension. A method is called optimal if, in terms of arithmetic operations (ops), the cost for the numerical calculation of x_n is of the same order as that of the matrix-vector product: 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.

In this sense, the analysis must refer to the sequence $\{A_n\}$, and not to a single matrix, since the objective is to quantify the difficulty of resolution of the linear system in relation to the accuracy (the better is the quality of the approximation, the larger is the size d_n) of the approximation considered.

The research of optimality has been, in some sense, the guideline that has connected the research topics covered during the Ph.D.: spectral analysis of sequences of structured matrices, preconditioning techniques, multigrid methods and saddle point problems.

The first topic is the largest part covered during the Ph.D. and is contained in the first part of the thesis: in particular we worked on the search for new mathematical tools (definitions and theorems) for analyzing the spectral distribution of sequences of matrices, mainly related to the shift-invariance property.

A key starting point concerning the spectral distribution of eigenvalues for sequences of Toeplitz matrices $\{T_n(f)\}$, where $T_n(f) = \left[\tilde{f}_{j-r}\right]_{j,r=0}^{n-1}$, $\tilde{f}_k$ being the Fourier coefficients of f, is the famous Szegö theorem which amounts to the following: if f is real-valued and essentially bounded then

$$\lim_{n\to\infty} \frac{1}{n} \sum_{\lambda \text{ eig. of } T_n(f)} F(\lambda) = \frac{1}{2\pi} \int_{[-\pi,\pi]} F(f(t))\, \mathrm{d}t, \tag{1}$$

for every continuous function F with compact support (see, for example, [48]).

The knowledge of the functional symbol f representing the limit distribution of eigenvalues has been shown to be crucial in understanding the superlinear convergence behavior [8] of Krylov methods (see, e.g., [47]), especially for the Conjugate Gradient (CG) method (see, e.g., [5]) in the Hermitian positive definite case, after resolution of the outliers: in this direction, we refer the reader to the analysis developed by Beckermann and Kuijlaars in recent years [8, 59] by using potential theory.

Several intermediate steps were done during the 20th century involving also singular values (Avram, Parter, etc.) and richer classes of symbols f (Böttcher, Silbermann, Widom, etc.).

In the 1990's, independently, Tilli and Tyrtyshnikov/Zamarashkin showed [109, 118] that equation (1) holds for any integrable real-valued function f. The corresponding result for a complex-valued function f and the sequence of sets of its singular values (with $|f|$ in the place of f) was first obtained by Parter (continuous times uni-modular symbols [72]), Avram (essentially bounded symbols [4]), and by Tilli and Tyrtyshnikov/Zamarashkin [109, 118], independently, for any integrable symbol f. (The book [20] gives a synopsis of all these results in Chapters 5 and 6). The relation (1) was established for a more general class of test functions F in [109, 87, 22] and the case of functions f of several variables (the *multi-level* case) and matrix-valued functions was studied in [109] and in [82] in the context of preconditioning (other related results were established by Linnik, Widom, Doktorski, see [20, Section 6.9]). Szegö-like results using different choices of families of test functions can be found in [19, 87, 109] (see [20, Chapters 5 and 6] for a synthesis of all these results).

However we have to be careful: a simple yet striking example where the eigenvalue result does *not* hold is given by the Toeplitz sequence related to the function $e^{-\mathrm{i}t}$, $\mathrm{i}^2 = -1$, which has only zero eigenvalues so that the condition (1) means that $F(0) = \frac{1}{2\pi}\int_{[-\pi,\pi]} F\left(e^{-\mathrm{i}t}\right) \mathrm{d}t$ which is far from being satisfied for all continuous functions with compact support, even though it *is* satisfied for every harmonic function (in cases like the latter it is better to consider the pseudospectrum, see [20]). In fact, Tilli was able to show that, if f is any complex-valued integrable function, then the condition (1) holds for all harmonic test functions F [110] and that it is even satisfied by all continuous functions with compact support as long as the symbol f satisfies a certain geometric conditions. More specifically, the function f must be essentially bounded and such that its (essential) range does not disconnect the complex plane and has empty interior, see [112]. We call this set of functions the *Tilli class*. In other contexts such a property is informally called "thin spectrum". It is clear that the set of all real-valued L^∞ functions is properly included in the Tilli class.

Further extensions of the Szegö result can be considered. An important direction of research is represented by the case of variable Toeplitz sequences or generalized locally Toeplitz sequences (see the pioneering work by Kac, Murdoch and Szegö [56] and by Parter [71], and, more recently, papers [111, 89, 90, 103, 17]). Another important direction is represented by the algebra generated by Toeplitz sequences and this is the main subject of the Chapter 4 of this thesis, with special attention to the case of eigenvalues, by using and extending tools from matrix theory (see Chapter 3) and finite-dimensional linear algebra; the case of singular values being already known (see [20, 90] and the references therein).

In our study we are motivated by the variety of fields where such matrices can be encountered such as, e.g., multigrid methods [50, 34], wavelet analysis [29], and subdivision

algorithms or, equivalently, in the associated refinement equations; see [36] and the references therein. Furthermore, it is interesting to remind that Strang [107] has shown rich connections between dilation equations in the wavelets context and multigrid algorithms [50, 114], when constructing the restriction/prolongation operators [43, 3] with various boundary conditions. It is worth noting that the use of different boundary conditions is quite natural when dealing with signal/image restoration problems or differential equations; see [88, 85].

The second part of the thesis is more focused on computational problems in which the previous spectral analysis comes into the play. More specifically, by employing the analysis from g-circulant/g-Toeplitz structures (see Chapter 5), we designed optimal regularizing preconditioners and multigrid methods. A further research topic concerns the saddle point problem and related preconditioning in which again a careful spectral analysis is essential. We can give a brief account of the findings in these three directions.

Regarding the first topic, the preconditioning of Toeplitz matrices via circulant matrices is widely studied in the literature (see, e.g., [24, 26] for the one-level case, [80] for the multi-level case, and [81] for the multi-level block case), in particular we know that, given a sequence of Toeplitz matrices $\{T_n(f)\}$ is possible to construct a sequence of circulant matrices $\{C_n(f)\}$ such that the sequence $\{C_n^{-1}(f)\,T_n(f)\}$ is clustered at 1 and this speed up the convergence of any Krylov like technique; but if we consider the more general definition of g-Toeplitz and g-circulant matrix (where for $g = 1$ we obtain the "classical" circulant and Toeplitz matrix) this is not true, then the case $g = 1$ is exceptional. However, in Chapter 6 we will see that, under suitable assumptions on the generating function f, there exist choices of g-circulant sequences which are *regularizing preconditioning sequence* for the corresponding g-Toeplitz structures.

The multigrid methods, in the last 20 years, have gained a remarkable reputation as fast solvers for structured matrices associated to shift-invariant operators, where the size n is large and the system shows a conditioning growing polynomially with n (see [42, 79, 43, 25, 54, 55, 108, 3, 95, 53] and the references therein). Under suitable mild assumptions, the considered techniques are optimal showing linear or almost linear ($O(n \log n)$ arithmetic operations as the celebrated fast Fourier transform (FFT)) complexity for reaching the solution within a preassigned accuracy and a convergence rate independent of the size n of the involved system.

These excellent features are identical in the multilevel setting, as already known for linear systems arising in the context of elliptic ordinary and partial differential equations (see [50, 77, 114, 85] and references therein). In particular, if the underlying structures are also sparse as in the multi-level banded case, then the cost of solving the involved linear system is proportional to the order of the coefficient matrix, with constant depending linearly on the bandwidths at each level. We mention that the cost of direct methods is $O(n \log n)$ operations in the case of trigonometric matrix algebras (circulant, τ, ...) and it is $O\left(n^{\frac{3d-1}{d}}\right)$ for d-level Toeplitz matrices (see [57]). Concerning multi-level Toeplitz structures, superfast methods represent a good alternative, even if the algorithmic error has to be controlled, and the cost grows as $O\left(n^{\frac{3d-2}{d}} \log^2(n)\right)$ ops: in fact, the computational burden is really competitive for $d = 1$, while the deterioration is evident for $d > 1$, since it is nontrivial to exploit the structure at the inner levels (see [28] and references therein and refer to [122] for recent interesting findings on the subject). Moreover, in the last case the most popular preconditioning strategies, with preconditioners taken in matrix algebras with unitary transform, can be far from being optimal in the multi-dimensional case (see [97] and [86, 70, 122] for further results). On the other hand, multigrid methods are optimal also for polynomially ill-conditioned multi-dimensional problems. Furthermore, these techniques can be extended to the case of low-rank corrections of the considered structured matrices, allowing to deal also with the modified Strang

preconditioner widely used in the literature (see [24] and references therein).

The main novelty contained in the literature treating structured matrices is the use of the symbol. Indeed, starting from the initial proposal in [42], we know that the convergence analysis of the two-grid and V-cycle can be handled in a compact and elegant manner by studying few analytical properties of the symbol (so the study does not involve explicitly the entries of the matrix and, more importantly, the size n of the system). Already in the two-grid convergence analysis, it is evident that the optimality can be reached only if the symbol f has a finite number of zeros of finite order and not located at mirror points: more explicitly, if x_0 is a zero of f then $f(x_0 + \pi)$ must be greater than zero.

In Chapter 7 we show that the second requirement is not essential since it depends on the choice of projecting the original matrix of size n into a new one of size $\frac{n}{2}$. The latter is not compulsory so that, by choosing a different size reduction from n to $\frac{n}{g}$ and $g > 2$, we can overcome the pathology induced by the mirror points. A different approach for dealing with such pathologies was proposed in [25] and further analyzed in [55], by exploiting a projection strategy based on matrix-valued symbols.

The third research topic concerns the saddle point problems. Large symmetric linear systems in saddle point form arise in many scientific and engineering applications. Their efficient solution by means of iterative methods heavily relies on exploiting the matrix structure. Constraint preconditioners are among the most successful structure-oriented preconditioning strategies, especially when dealing with optimization problems. In Chapter 8 we provide a full spectral characterization of the constraint-based preconditioned matrix by means of the Weyr canonical form. We also derive estimates for the spectrum when the preconditioner needs to be modified to cope with possible high computational costs of its original version.

This Ph.D. thesis has been divided into two parts, a first part which explains the methodologies used and the theoretical results obtained and a second part which contains some applications.

Part I - Spectral Distributions of Structured Matrix-Sequences: Tools

- In Chapter 1 we introduce useful definitions and the notations used throughout the thesis. We will see the concept of approximation class of sequences (*a.c.s.*) in order to define a basic approximation theory for matrix-sequences, the definition of distribution in the eigen/singular value sense and the clustering/attracting properties of matrix-sequences.

- In Chapter 2 we report well-known results concerning the notion of *a.c.s.* and of spectral distribution. In particular we introduce the basic result of approximating theory for sequences of matrices (used to demonstrate the most famous theorem of distribution for sequences of Toeplitz matrices: the Szegö theorem), and some results of spectral distribution for linear combinations of special sequences of matrices. In Section 2.4 we report some results of perturbation for sequences of matrices (from Golinskii and Serra-Capizzano [45]); these results will be used in Chapter 3 in order to deduce new results of approximation for sequences of matrices.

- In Chapter 3 we enlarge the set of tools for providing the existence and for characterizing explicitly the limit distribution of eigen/singular values for general (structured) matrix-sequences. In particular, we test the stability of the notion of approximating class of sequences (*a.c.s.*) for sequences of Hermitian sparsely unbounded matrices under the influence of continuous functions defined on the real axis, then under mild trace norm assumptions on the perturbing sequence, we extend the recent perturbation result based on a theorem by Mirsky see in Chapter 2 to the analysis of the eigenvalue distribution and localization of a generic (non-Hermitian) complex perturbation of a bounded Hermitian

sequence of matrices. Finally we introduce two different but equivalent formulations of Theorem 2.15 that will be used in Chapter 4 to prove a result of distribution for sequences of products of Toeplitz matrices.

- Chapter 4 is devoted entirely to Toeplitz sequences: in addition to the definitions and classical distribution results, we use the result concerning the eigenvalues of a generic (non-Hermitian) complex perturbation of a bounded Hermitian sequence of matrices proved in Chapter 3 to prove that the asymptotic spectrum of the product of Toeplitz sequences, whose symbols have a real-valued essentially bounded product h, is described by the function h in the "Szegö way". Then, using Mergelyan's theorem, we extend the result to the more general case where h belongs to the Tilli class. The same technique gives us the analogous result for sequences belonging to the algebra generated by Toeplitz sequences, if the symbols associated with the sequences are bounded and the global symbol h belongs to the Tilli class. A generalization to the case of multi-level matrix-valued symbols and a study of the case of Laurent polynomials not necessarily belonging to the Tilli class are also given.

- In Chapter 5 we introduce the definitions of g-circulant and g-Toeplitz matrix, a generalization of the classical circulant and Toeplitz matrix. We study the eigen/singular values of g-circulant matrices and provide an asymptotic analysis of the distribution results for the singular values of g-Toeplitz sequences in the case where the sequence of coefficients generating the g-Toeplitz can be interpreted as the sequence of Fourier coefficients of an integrable function f over the domain $(-\pi, \pi)$.

Part II - Spectral Distributions of Structured Matrix-Sequences: Applications

- In Chapter 6 we are interested in the preconditioning problem for g-Toeplitz matrices which is well understood and widely studied in the last three decades for $g = 1$. In particular, we consider the general case with $g \geq 2$ and the interesting result is that the preconditioned sequence $\{\mathcal{P}_n\} = \{P_n^{-1} A_n\}$, where $\{P_n\}$ is the sequence of preconditioner and $\{A_n\}$ is the sequence of g-Toeplitz matrices, cannot be clustered at 1 so that the case of $g = 1$ is exceptional. However, while a satisfactory standard preconditioning cannot be achieved, the result has a positive implication since there exist choices of g-circulant sequences which are regularizing preconditioning sequence for the corresponding g-Toeplitz structures.

- In Chapter 7, starting from the spectral analysis of g-circulant matrices given in Chapter 5, we study the convergence of a multigrid method for circulant and Toeplitz matrices with various size reductions. We assume that the size n of the coefficient matrix is divisible by $g \geq 2$ such that at the lower level the system is reduced to one of size $\frac{n}{g}$, by employing g-circulant based projectors. We perform a rigorous two-grid convergence analysis in the circulant case and we extend experimentally the results to the Toeplitz setting, by employing structure-preserving projectors. The optimality of the two-grid method and of the multigrid method is proved, when the number $\theta \in \mathbb{N}$ of recursive calls is such that $1 < \theta < g$. The previous analysis is used also to overcome some pathological cases, in which the generating function has zeros located at "mirror points" and the standard two-grid method with $g = 2$ is not optimal.

- In Chapter 8 we deal with linear systems with non-singular symmetric coefficient matrix $\mathcal{A}$ that arise in many applications associated with the numerical solution of saddle point problems. We present a spectral analysis of the preconditioned matrix $\mathcal{P}^{-1}\mathcal{A}$ where $\mathcal{P}$ is not the "ideal" preconditioner but a computationally less expensive version of this. Much is known about the spectrum in the ideal case, characterized by a rich spectral structure, with non-trivial Jordan blocks and favourable real eigenvalue distribution,

while the spectral analysis of the general though far more realistic case ($\mathcal{P}$ not "ideal") has received less attention, possibly due to the difficulty of dealing with Jordan block perturbations. In this chapter we want to fill this gap.

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 [39, 101, 34, 68, 69, 99, 93, 92].

Part I

Spectral Distributions of Structured Matrix-Sequences: Tools

Chapter 1

Notations and definitions

The aim of this introductory chapter is simply to fix the notations used throughout the thesis, to recall some basic concepts of linear algebra and to illustrate the definitions that are necessary to work with sequences of matrices.

In particular, in Section 1.3 we introduce the definition of spectral distribution of a sequence of matrices; this concept links the collective behavior of the eigenvalues or singular values of all the matrices in the sequence to the behavior of a given function (or measure). Moreover we illustrate some properties and characterizations of the spectrum of a sequence of matrices such as "clustering" and "attraction". We conclude the chapter with the definitions of sparsely unbounded and sparsely vanishing sequence of matrices that will be frequently used in Chapter 6.

1.1 Some concepts of linear algebra

We denote by $M_{m,n}(\mathbb{C})$ ($M_{m,n}(\mathbb{R})$) the vector space of matrices of dimension $m \times n$ with complex (real) elements, where the matrices are square, i.e. $m = n$, we simply write $M_n(\mathbb{C})$ ($M_n(\mathbb{R})$). In the following, for matrices and vectors, $^\top$ denote the transpose operator while * the transpose conjugate operator.

Definition 1.1. *A matrix $A \in M_n(\mathbb{C})$ is* Hermitian *if $A = A^*$, and is* skew-Hermitian *if $A = -A^*$. Given a Hermitian matrix $A \in M_n(\mathbb{C})$, if we have that $x^*Ax > 0$ ($x^*Ax \geq 0$) for all $x \in \mathbb{C}^n$, $x \neq 0$, then A is* positive definite (semidefinite).

Definition 1.2. *Let $A, B \in M_n(\mathbb{C})$ be two Hermitian matrices, then the notation $A < B$ ($A \leq B$) means that $B - A$ is positive definite (semidefinite).*

For Hermitian matrices the following theorem holds.

Theorem 1.3. *Let $A \in M_n(\mathbb{C})$ be a Hermitian matrix with eigenvalues $\lambda_1(A), \ldots, \lambda_n(A)$. Then A is positive definite (semidefinite) if and only if $\lambda_j(A) > 0$ ($\lambda_j(A) \geq 0$), for $j = 1, 2, \ldots, n$.*

Any matrix $A \in M_n(\mathbb{C})$ always can be uniquely written as a Hermitian matrix plus a skew-Hermitian matrix (in analogy to case of scalar complex numbers). More precisely we have

$$A = \mathrm{Re}(A) + \mathrm{i}\,\mathrm{Im}(A), \qquad \mathrm{i}^2 = -1, \tag{1.1}$$

$$\mathrm{Re}(A) = \frac{A + A^*}{2}, \tag{1.2}$$

$$\mathrm{Im}(A) = \frac{A - A^*}{2\mathrm{i}}, \tag{1.3}$$

where $\mathrm{Re}(A)$ and $\mathrm{Im}(A)$ are Hermitian matrices so that $\mathrm{i}\,\mathrm{Im}(A)$ is skew-Hermitian.

For a matrix $A \in M_n(\mathbb{C})$ with eigenvalues $\lambda_j(A)$, $j = 1, \ldots, n$, and for a matrix $B \in M_{n,m}(\mathbb{C})$ with singular values $\sigma_j(B)$, $j = 1, \ldots, k$, $k = \min\{m, n\}$, we set

$$\begin{aligned} \Lambda(A) &= \{\lambda_1(A), \lambda_2(A), \ldots, \lambda_n(A)\}, \\ \Omega(B) &= \{\sigma_1(B), \sigma_2(B), \ldots, \sigma_k(B)\}. \end{aligned} \tag{1.4}$$

There is a relation between the singular values of a matrix $B \in M_{n,m}(\mathbb{C})$ and the eigenvalues of $B^*B \in M_m(\mathbb{C})$: firstly the matrix B^*B is positive semidefinite, since $x^*(B^*B)x = \|Bx\|_2^2 \geq 0$ for all $x \in \mathbb{C}^m$ (see the definition of $\|\cdot\|_2$ in (1.12)), then, by Theorem 1.3, the eigenvalues $\lambda_1(B^*B) \geq \lambda_2(B^*B) \geq \cdots \geq \lambda_m(B^*B)$ are non-negative and can, therefore, be written in the form

$$\lambda_j(B^*B) = \sigma_j^2, \tag{1.5}$$

with $\sigma_j \geq 0$, $j = 1, \ldots, m$. The numbers $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_k \geq 0$, $k = \min\{m, n\}$, are the singular values of B, that is $\sigma_j = \sigma_j(B)$, $j = 1, \ldots, k$, and if $m > k$, then $\lambda_j(B^*B) = 0$, $j = k+1, \ldots, m$. For a more general statement, refer to the singular value decomposition (SVD) theorem (see, e.g., the classical book by Golub and Van Loan [46]).

The following theorem reduces the calculation of the eigenvalues of a singular matrix $A \in M_n(\mathbb{C})$ with $\text{rank}(A) = k \leq n$ to the calculation of the eigenvalues of a smaller matrix $\tilde{A} \in M_k(\mathbb{C})$.

Theorem 1.4. *Let $A \in M_n(\mathbb{C})$ be a matrix which can be written as $A = XY^*$, where $X, Y \in M_{n,k}(\mathbb{C})$, with $k \leq n$. Then the n eigenvalues of the matrix A are given by k eigenvalues of the matrix $Y^*X \in M_k(\mathbb{C})$ and $n - k$ zero eigenvalues:*

$$\Lambda(A) = \Lambda(Y^*X) \cup \{0 \textit{ with geometric multiplicity } n-k\}.$$

Proof. The singular values decomposition (SVD) of the matrix $X \in M_{n,k}(\mathbb{C})$, $k \leq n$, is given by

$$X = U\Sigma V^* = U \begin{bmatrix} \sigma_1 & & & \\ & \ddots & & \\ & & \sigma_t & \\ & & & 0 \\ \hline & & 0 & \end{bmatrix} V^*,$$

where $U \in M_n(\mathbb{C})$ and $V \in M_k(\mathbb{C})$ are unitary matrices, $\Sigma \in M_{n,k}(\mathbb{R})$, $t = \text{rank}(X)$, then $t \leq k$, and $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_t > 0$.

Consider the matrix $X_\epsilon \in M_n(\mathbb{C})$, perturbation of the matrix X, defined in this way:

$$X_\epsilon = U \begin{bmatrix} \sigma_1 & & & \\ & \ddots & & \\ & & \sigma_t & \\ & & & \epsilon \\ & & & & \ddots \\ & & & & & \epsilon \end{bmatrix} \left[\begin{array}{c|c} V^* & 0 \\ \hline 0 & I_{n-k} \end{array}\right] = U\Sigma_\epsilon \widetilde{V}^*, \tag{1.6}$$

where $I_{n-k} \in M_{(n-k)}(\mathbb{R})$ is the identity matrix, $U, \widetilde{V} \in M_n(\mathbb{C})$ are unitary matrices and $\Sigma_\epsilon \in M_n(\mathbb{R})$, with $\epsilon > 0$ and $\sigma_t \geq \epsilon$ ((1.6) is an SVD for X_ϵ); it is immediate to observe that the matrix X_ϵ is invertible.

We build now the matrix $A_\epsilon \in M_n(\mathbb{C})$ in this way:

$$A_\epsilon = X_\epsilon \left[\begin{array}{c} Y^* \\ \hline 0 \end{array}\right]. \tag{1.7}$$

Since X_ϵ is invertible, the matrix A_ϵ is similar (has the same eigenvalues) to the matrix $A'_\epsilon = X_\epsilon^{-1} A_\epsilon X_\epsilon$ where

$$A'_\epsilon = X_\epsilon^{-1} A_\epsilon X_\epsilon = X_\epsilon^{-1} X_\epsilon \left[\frac{Y^*}{0}\right] X_\epsilon = \left[\frac{Y^*}{0}\right] X_\epsilon, \tag{1.8}$$

then $\Lambda(A_\epsilon) = \Lambda\left(A'_\epsilon\right)$, and also the characteristic polynomials of the two matrices are the same (we recall that the characteristic polynomial of a matrix A is the polynomial whose roots are precisely the eigenvalues of A),

$$p_{A_\epsilon}(\lambda) = p_{A'_\epsilon}(\lambda) = p_\epsilon(\lambda). \tag{1.9}$$

We observe that, since the matrix $\left[\frac{Y^*}{0}\right]$ in (1.7) and (1.8) does not depend on ϵ, and X_ϵ is a matrix whose entries depend linearly from ϵ, we have that the entries of the two matrices A_ϵ and A'_ϵ are linear functions in the variable ϵ, this means that the characteristic polynomial $p_\epsilon(\lambda)$, by construction, is continuous with respect to ϵ. Now, for $\epsilon \to 0$ we have that

$$\begin{aligned}
X_\epsilon &= U \begin{bmatrix} \sigma_1 & & & \\ & \ddots & & \\ & & \sigma_t & \\ & & & \ddots \\ & & & & \epsilon \end{bmatrix} \left[\begin{array}{c|c} V^* & 0 \\ \hline 0 & I_{n-k} \end{array}\right] \\
&\xrightarrow[\epsilon\to 0]{} U \begin{bmatrix} \sigma_1 & & & \\ & \ddots & & \\ & & \sigma_t & \\ & & & \ddots \\ & & & & 0 \end{bmatrix} \left[\begin{array}{c|c} V^* & 0 \\ \hline 0 & I_{n-k} \end{array}\right] \\
&= \left[\begin{array}{c|c} U \begin{bmatrix} \sigma_1 & & & \\ & \ddots & & \\ & & \sigma_t & \\ & & & \ddots \\ & & & & 0 \\ \hline & & & & 0 \end{bmatrix} V^* & 0 \end{array}\right] = \left[\begin{array}{c|c} X & 0 \end{array}\right];
\end{aligned}$$

and so

$$A_\epsilon = X_\epsilon \left[\frac{Y^*}{0}\right] \xrightarrow[\epsilon\to 0]{} \left[\begin{array}{c|c} X & 0 \end{array}\right] \left[\frac{Y^*}{0}\right] = XY^* = A; \tag{1.10}$$

$$A'_\epsilon = \left[\frac{Y^*}{0}\right] X_\epsilon \xrightarrow[\epsilon\to 0]{} \left[\frac{Y^*}{0}\right] \left[\begin{array}{c|c} X & 0 \end{array}\right] = \left[\begin{array}{c|c} Y^*X & 0 \\ \hline 0 & 0 \end{array}\right]. \tag{1.11}$$

Now, since as mentioned above the characteristic polynomial of A_ϵ is continuous with respect to ϵ, by (1.10) is true that

$$p_{A_\epsilon}(\lambda) \xrightarrow[\epsilon\to 0]{} p_A(\lambda),$$

and from (1.9) we obtain

$$p_{A'_\epsilon}(\lambda) \xrightarrow[\epsilon\to 0]{} p_A(\lambda),$$

that is, the eigenvalues of the matrix $A = XY^*$ in (1.10) are the same as those of the matrix in (1.11). □

1.2 The Schatten p-norms and functional norms

Let $\mathbb{C}^n$ be the complex vector space of dimension n, $\forall x \in \mathbb{C}^n$ the class of p vector norm, $p \in [1, \infty]$, is defined as

$$\begin{aligned} \|x\|_p &= \left(\sum_{j=1}^{n} |x_j|^p \right)^{\frac{1}{p}}, \qquad p \in [1, +\infty), \\ \|x\|_\infty &= \max_{j=1,\ldots,n} |x_j|, \qquad p = \infty, \end{aligned} \tag{1.12}$$

where for $p = 2$ we obtain the so-called Euclidean norm.

If $A \in M_n(\mathbb{C})$ has singular values $\sigma_1(A) \geq \sigma_2(A) \geq \cdots \geq \sigma_n(A)$, we define $\|A\|_p$, with $p \in [1, \infty]$, the Schatten p-norm of A to be the p vector norm of the vector of the singular values of A:

$$\begin{aligned} \|A\|_p &= \left(\sum_{j=1}^{n} (\sigma_j(A))^p \right)^{\frac{1}{p}}, \qquad p \in [1, +\infty), \\ \|A\|_\infty &= \sigma_1(A), \qquad p = \infty. \end{aligned}$$

We will be especially interested in the norm $\|\cdot\|_1$ which is known as the trace norm (i.e. the sum of all the singular values of a matrix), the norm $\|\cdot\|_2$ called Frobenius or Euclidean (induced matrix) norm and the norm $\|\cdot\|_\infty$ which is equal to the usual operator (spectral) norm $\|\cdot\|$

$$\|A\| = \sup_{\|x\|_2 = 1} \|Ax\|,$$

(in the following we will use the notation $\|\cdot\|$ for the spectral norm). For the Schatten p-norms, Hölder's inequality applies: if $A, B \in M_n(\mathbb{C})$ then

$$\|AB\|_1 \leq \|A\| \, \|B\|_1 . \tag{1.13}$$

Another well-known inequality involving the Schatten 1-norm of a matrix $A \in M_n(\mathbb{C})$ is the following

$$|\mathrm{tr}(A)| \leq \|A\|_1 , \tag{1.14}$$

where $\mathrm{tr}(A)$ is the trace of A, i.e., the sum of all its diagonal entries (or equivalently the sum of all its eigenvalues).

A simple proof of (1.14) is as follows. Let $A = U\Sigma V^*$ be the singular value decomposition of A [15]. Then by a similarity argument

$$\mathrm{tr}(A) = \mathrm{tr}(U\Sigma V^*) = \mathrm{tr}(\Sigma V^* U) = \mathrm{tr}(\Sigma W),$$

with $W = V^*U$ being unitary. So, $|\mathrm{tr}(A)| = |\sigma_1 w_1 + \sigma_2 w_2 + \cdots + \sigma_n w_n|$, with $\sigma_1, \sigma_2, \ldots, \sigma_n$ being the singular values of A and where $w_1, w_2, \ldots, w_n$ are the diagonal entries of W: all of them bounded by 1. Hence the application of the triangle inequality yields (1.14).

Finally, if $A \in M_n(\mathbb{C})$ is positive definite, then

$$\|\cdot\|_A = \left\| A^{\frac{1}{2}} \cdot \right\|_2 ,$$

denotes the Euclidean norm weighted by A on $\mathbb{C}^n$ and the associated induced matrix norm. We recall that, if $A \in M_n(\mathbb{C})$ is a positive definite matrix then, by Schur decomposition, it can be written as $A = UDU^*$ with U unitary and $D = \underset{j=0,\ldots,n-1}{\mathrm{diag}} (\lambda_j(A))$ diagonal, real

and positive (see Theorem 1.3), so the matrix $A^{\frac{1}{2}}$ is defined as $A^{\frac{1}{2}} = UD^{\frac{1}{2}}U^*$ with $D^{\frac{1}{2}} = \underset{j=0,\ldots,n-1}{\operatorname{diag}} \left(\sqrt{\lambda_j(A)}\right)$.

In the following we use some functional norms. Let $f : Q \to \mathbb{C}$, $Q \subseteq K^d$, with K equals $\mathbb{R}$ or $\mathbb{C}$, $d \geq 1$, Q Lebesgue measurable. We define the $L^p(Q)$ spaces, subspace of measurable functions on Q, and the respective norms, as follows:

$$\text{if } 1 \leq p < \infty \quad L^p(Q) = \left\{ f : Q \to \mathbb{C} : \int_Q |f(x)|^p \, \mathrm{d}x < \infty \right\},$$

$$\text{with} \quad \|f\|_{L^p} = \left[\int_Q |f(x)|^p \, \mathrm{d}x \right]^{\frac{1}{p}},$$

$$\text{if } p = \infty \quad L^\infty(Q) = \{ \text{Space of essentially bounded functions on } Q \},$$

$$\text{with} \quad \|f\|_{L^\infty} = \sup_{x \in Q} |f(x)|,$$

(according to the Haar measure).

For these norms, Hölder's inequality applies: if $v \in L^p(Q)$ and $u \in L^q(Q)$ with $p, q \geq 1$ and $\frac{1}{p} + \frac{1}{q} = 1$, then $vu \in L^1(Q)$ and

$$\|vu\|_{L^1} \leq \|v\|_{L^p} \|u\|_{L^q}. \tag{1.15}$$

1.3 Sequences of matrices

Throughout this thesis we speak of *matrix-sequences* as sequences $\{A_k\}$ where A_k is an $n(k) \times m(k)$ matrix with $\min\{n(k), m(k)\} \to \infty$ as $k \to \infty$. When $n(k) = m(k)$, that is, all the involved matrices are square, and this will occur often in the thesis, we will not need the extra parameter k and will consider simply matrix-sequences of the form $\{A_n\}$.

Concerning the matrix-sequences the notion of approximating class of sequences was introduced as reported below.

Definition 1.5. [83] *Suppose a sequence of matrices $\{A_n\}$, $A_n \in M_{d_n}(\mathbb{C})$ ($d_k < d_{k+1}$ for each k) is given. We say that $\{\{B_{n,m}\} : m \geq 0\}$, $m \in \hat{\mathbb{N}} \subset \mathbb{N}$, $\#\hat{\mathbb{N}} = \infty$, $B_{n,m} \in M_{d_n}(\mathbb{C})$, is an* approximating class of sequences (*a.c.s.*) *for $\{A_n\}$ if, for all sufficiently large $m \in \hat{\mathbb{N}}$, the following splitting holds*

$$A_n = B_{n,m} + R_{n,m} + N_{n,m}, \qquad \forall n > n_m,$$

with

$$\operatorname{rank}(R_{n,m}) \leq d_n c(m), \quad \|N_{n,m}\| \leq \omega(m), \tag{1.16}$$

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

$$\lim_{m \to \infty} \omega(m) = 0, \qquad \lim_{m \to \infty} c(m) = 0. \tag{1.17}$$

The idea behind the concept of *a.c.s.* was to define a basic approximation theory for matrix-sequences with respect to the global distribution of eigenvalues and singular values. More precisely, given a "difficult" sequence $\{A_n\}$, the goal is to recover its global spectral behavior from the spectral behavior of simpler approximating sequences, as we shall see in the next chapter.

Remark 1.6. *The matrix $R_{n,m}+N_{n,m}$ can be a full rank matrix (invertible) with spectral norm arbitrarily large. Indeed, let us consider, for example, the matrix $L_{n,m} \in M_n(\mathbb{C})$ defined as*

$$L_{n,m} = \begin{bmatrix} \phi(n) & & & & \\ & \frac{1}{m+1} & & & \\ & & \frac{1}{m+1} & & \\ & & & \ddots & \\ & & & & \frac{1}{m+1} \end{bmatrix}, \tag{1.18}$$

with $\phi(n) = n!$. It is immediate to observe that $L_{n,m}$ is a full rank matrix with $\|L_{n,m}\| = \phi(n)$, however, we can write $L_{n,m}$ as

$$L_{n,m} = \begin{bmatrix} \phi(n) & & & \\ & 0 & & \\ & & \ddots & \\ & & & 0 \end{bmatrix} + \begin{bmatrix} 0 & & & & \\ & \frac{1}{m+1} & & & \\ & & \frac{1}{m+1} & & \\ & & & \ddots & \\ & & & & \frac{1}{m+1} \end{bmatrix} = L_{n,m}^{(1)} + L_{n,m}^{(2)},$$

with $\operatorname{rank}\left(L_{n,m}^{(1)}\right) = 1$ *and* $\left\|L_{n,m}^{(2)}\right\| = \frac{1}{m+1}$*, that is $L_{n,m}$ is the sum of two matrices: a matrix of small norm, for m large enough, and the other of low-rank $\forall m \geq 1$.*

For matrix-sequences an important notion is that of spectral distribution in the eigenvalue or singular value sense, linking the collective behavior of the eigenvalues or singular values of all the matrices in the sequence to the behavior of a given function (or measure). First we need some notations.

For any function F defined on $\mathbb{C}$ and for any matrix $A \in M_n(\mathbb{C})$, the symbol $\Sigma_\lambda(F, A)$ stands for the mean

$$\Sigma_\lambda(F, A) := \frac{1}{n}\sum_{j=1}^{n} F(\lambda_j(A)) = \frac{1}{n}\sum_{\lambda\in\Lambda(A)} F(\lambda),$$

similarly, for any function F defined on $\mathbb{R}_0^+$ and for any matrix $A \in M_{n,m}(\mathbb{C})$, the symbol $\Sigma_\sigma(F, A)$ stands for the mean

$$\Sigma_\sigma(F, A) := \frac{1}{\min\{n,m\}}\sum_{j=1}^{\min\{n,m\}} F(\sigma_j(A)) = \frac{1}{\min\{n,m\}}\sum_{\sigma\in\Omega(A)} F(\sigma). \tag{1.19}$$

Definition 1.7. *Let $\mathcal{C}_0(\mathbb{C})$ be the set of continuous functions with bounded support defined over the complex field, let d be a positive integer, and let θ be a complex-valued measurable function defined on a set $G \subset \mathbb{R}^d$ of finite and positive Lebesgue measure $m\{G\}$. Here G will be equal to $(-\pi,\pi)^d$ so that $e^{\mathrm{i}\overline{G}} = \mathbb{T}^d$ with $\mathbb{T}$ denoting the complex unit circle and $\overline{G}$ denoting the closure of G. A matrix-sequence $\{A_n\}$, $A_n \in M_{d_n}(\mathbb{C})$ ($d_k < d_{k+1}$ for each k), is said to be* distributed (in the sense of the eigenvalues) as the pair (θ, G), *or to* have the distribution function θ*, if, for every $F \in \mathcal{C}_0(\mathbb{C})$, the following limit relation holds*

$$\lim_{n\to\infty}\Sigma_\lambda(F, A_n) = \frac{1}{m\{G\}}\int_G F(\theta(t))\,\mathrm{d}t, \qquad t = (t_1,\ldots,t_d), \tag{1.20}$$

and in this case we write that $\{A_n\} \sim_\lambda (\theta, G)$.

If (1.20) holds for every $F \in \mathcal{C}_0\left(\mathbb{R}_0^+\right)$ in place of $F \in \mathcal{C}_0(\mathbb{C})$, with the singular values $\sigma_j(A_n)$, $j = 1,\ldots,n$, in place of the eigenvalues, and with $|\theta(t)|$ in place of $\theta(t)$, we say that $\{A_n\} \sim_\sigma (\theta, G)$ or that the matrix-sequence $\{A_n\}$ is distributed (in the sense of the singular values) as the pair (θ, G)*: more specifically for every $F \in \mathcal{C}_0\left(\mathbb{R}_0^+\right)$ we have*

$$\lim_{n\to\infty}\Sigma_\sigma(F, A_n) = \frac{1}{m\{G\}}\int_G F(|\theta(t)|)\,\mathrm{d}t, \qquad t = (t_1,\ldots,t_d). \tag{1.21}$$

When considering θ taking values in $M_{p,q}(\mathbb{C})$ and a function is considered to be measurable if and only if the component functions are, we say that $\{A_n\} \sim_\sigma (\theta, G)$ when for every $F \in \mathcal{C}_0\left(\mathbb{R}_0^+\right)$ we have

$$\lim_{n\to\infty} \Sigma_\sigma(F, A_n) = \frac{1}{m\{G\}} \int_G \frac{\sum_{j=1}^{\min\{p,q\}} F(\sigma_j(\theta(t)))}{\min\{p,q\}} \mathrm{d}t, \qquad t = (t_1, \ldots, t_d),$$

with $\sigma_j(\theta(t)) = \sqrt{\lambda_j(\theta(t)\theta^*(t))}$. If $p = q$ we say that $\{A_n\} \sim_\lambda (\theta, G)$ when for every $F \in \mathcal{C}_0(\mathbb{C})$ we have

$$\lim_{n\to\infty} \Sigma_\lambda(F, A_n) = \frac{1}{m\{G\}} \int_G \frac{\sum_{j=1}^{q} F(\lambda_j(\theta(t)))}{q} \mathrm{d}t, \qquad t = (t_1, \ldots, t_d), \tag{1.22}$$

where $\lambda_i(\theta(t))$ in equation (1.22) are the eigenvalues of the matrix-valued function $\theta(t)$.

Finally we say that two sequences $\{A_n\}$ and $\{B_n\}$ are equally distributed *in the sense of eigenvalues (λ) or in the sense of singular values (σ) if, $\forall F \in \mathcal{C}_0(\mathbb{C})$, we have*

$$\lim_{n\to\infty} [\Sigma_\nu(F, B_n) - \Sigma_\nu(F, A_n)] = 0, \qquad \text{with} \quad \nu = \lambda \text{ or } \nu = \sigma.$$

Notice that two sequences having the same distribution function are equally distributed. On the other hand, two equally distributed sequences may not be associated with a distribution function at all: consider any diagonal matrix-sequence $\{A_n\}$ and let $\{B_n\}$ be a sequence of the form $B_n = A_n + \epsilon_n I_n$ with $\epsilon_n \to 0$ when $n \to \infty$. Then, if the original $\{A_n\}$ does not have an eigenvalue distribution function (e.g. $A_n = (-1)^n I_n$), we will have $\{A_n\}$ and $\{B_n\}$ equally distributed, even though it is impossible to associate a distribution function with either of them. On the other hand, if one of them has a distribution function, then the other necessarily has the same one. This is easy to show using the definitions (or see [84, Remark 6.1]).

Now, we can observe that a matrix-sequence $\{A_n\}$ is distributed as the pair (θ, G) if and only if the sequence of linear functionals $\{\phi_n\}$ defined by $\phi_n(F) = \Sigma_\lambda(F, A_n)$ converges weak-* to the functional $\phi(F) = \frac{1}{m\{G\}} \int_G F(\theta(t))\,\mathrm{d}t$ as in (1.20). In order to describe what this really means about the asymptotic qualities of the spectrum, we will derive more concrete characterizations of $\{\Lambda(A_n)\}$ such as "clustering" and "attraction", where $\Lambda(A_n)$ is defined in (1.4).

Definition 1.8. [116] *A matrix-sequence $\{A_n\}$, $A_n \in M_{d_n}(\mathbb{C})$ ($d_k < d_{k+1}$ for each k), is* strongly clustered at $s \in \mathbb{C}$ *(in the eigenvalue sense), if for any $\varepsilon > 0$ the number of the eigenvalues of A_n off the disc*

$$D(s, \varepsilon) := \{z : |z - s| < \varepsilon\}, \tag{1.23}$$

can be bounded by a constant q_ε possibly depending on ε, but not on n. In other words

$$q_\varepsilon(n, s) := \#\{j : \lambda_j(A_n) \notin D(s, \varepsilon)\} = O(1), \quad n \to \infty.$$

If every A_n has only real eigenvalues (at least for large n) then we may assume that s is real and that the disc $D(s, \varepsilon)$ is the interval $(s - \varepsilon, s + \varepsilon)$. A matrix-sequence $\{A_n\}$ is said to be strongly clustered at a non-empty closed set $S \subset \mathbb{C}$ *(in the eigenvalue sense) if for any $\varepsilon > 0$*

$$q_\varepsilon(n, S) := \#\{j : \lambda_j(A_n) \notin D(S, \varepsilon)\} = O(1), \quad n \to \infty, \tag{1.24}$$

where $D(S, \varepsilon) := \bigcup_{s\in S} D(s, \varepsilon)$ is the ε-neighborhood of S. If every A_n has only real eigenvalues (at least for large n), then S is a non-empty closed subset of $\mathbb{R}$. We replace the term "strongly" by "weakly", if

$$q_\varepsilon(n, s) = o(d_n), \qquad (q_\varepsilon(n, S) = o(d_n)), \quad n \to \infty,$$

in the case of a point s (or a closed set S). Finally, if we replace eigenvalues with singular values, we obtain all the corresponding definitions for singular values.

To link the concept of cluster with the distribution notion it is instructive to observe that $\{A_n\} \sim_\lambda (\theta, G)$, with $\theta \equiv s$ equal to a constant function if and only if $\{A_n\}$ is weakly clustered at $s \in \mathbb{C}$ (for more results and relations between the notions of equal distribution, equal localization, spectral distribution, spectral clustering etc., see [84, Section 4]). We introduce one more notion concerning the eigenvalues of a matrix-sequence.

Definition 1.9. *Let $\{A_n\}$ be a matrix-sequence, $A_n \in M_{d_n}(\mathbb{C})$ ($d_k < d_{k+1}$ for each k), and let $\Lambda(A_n)$ defined as in (1.4). We say that $\{A_n\}$ is* strongly attracted by $s \in \mathbb{C}$ *if*

$$\lim_{n\to\infty} \operatorname{dist}(s, \Lambda(A_n)) = 0, \tag{1.25}$$

where $\operatorname{dist}(X, Y)$ *is the usual Euclidean distance between two subsets X and Y of the complex plane. Furthermore, if we order the eigenvalues according to their distance from s, i.e.,*

$$|\lambda_1(A_n) - s| \le |\lambda_2(A_n) - s| \le \cdots \le |\lambda_{d_n}(A_n) - s|,$$

then we say that the attraction to s is of order $r(s) \in \mathbb{N}$, $r(s) \ge 1$ is a fixed number, if

$$\lim_{n\to\infty} \left|\lambda_{r(s)}(A_n) - s\right| = 0, \qquad \liminf_{n\to\infty} \left|\lambda_{r(s)+1}(A_n) - s\right| > 0,$$

and that the attraction is of order $r(s) = \infty$ if

$$\lim_{n\to\infty} |\lambda_j(A_n) - s| = 0,$$

for every fixed j. Finally, one defines weak attraction by replacing lim *with* liminf *in (1.25).*

It is not hard to see that, if $\{A_n\}$ is at least weakly clustered at a point s, then s strongly attracts $\{A_n\}$ with infinite order. Indeed, if there is an attraction of finite order to s then

$$\lim_{n\to\infty} \frac{\#\{\lambda \in \Lambda(A_n) : \lambda \notin D(s, \delta)\}}{d_n} = 1,$$

for some $\delta > 0$ and this is impossible if $\{A_n\}$ is weakly clustered at s. On the other hand, there are sequences which are strongly attracted by s with infinite order, but not even weakly clustered at s. Indeed, the notion of weak clustering does not tell anything concerning weak attraction or attraction of finite order.

In the following we introduce the definitions of sparsely unbounded and sparsely vanishing matrix-sequences and of sparsely unbounded and sparsely vanishing functions, which will be widely used in Chapter 6.

Definition 1.10. *A sequence of matrices $\{A_n\}$, $A_n \in M_{d_n}(\mathbb{C})$ ($d_k < d_{k+1}$ for each k), is said to be* sparsely unbounded *if there exists a non-negative function $x(s)$ with $\lim_{s\to 0} x(s) = 0$ so that $\forall \varepsilon > 0$ $\exists n_\varepsilon \in \mathbb{N}$ such that $\forall n \ge n_\varepsilon$*

$$\frac{1}{d_n} \#\left\{i : \sigma_i(A_n) \ge \frac{1}{\varepsilon}\right\} \le x(\varepsilon), \tag{1.26}$$

where $\sigma_i \in \Omega(A_n)$, $i = 1, 2, \ldots, n$ (see (1.4)).

Analogously, a sequence of matrices $\{A_n\}$ is said to be sparsely vanishing *if there exists a non-negative function $x(s)$ with $\lim_{s\to 0} x(s) = 0$ so that $\forall \varepsilon > 0$ $\exists n_\varepsilon \in \mathbb{N}$ such that $\forall n \ge n_\varepsilon$*

$$\frac{1}{d_n} \#\{i : \sigma_i(A_n) \le \varepsilon\} \le x(\varepsilon).$$

Definition 1.11. *A function θ is* sparsely unbounded *if*

$$\lim_{\eta \to 0} m \left\{ x : |\theta(x)| > \frac{1}{\eta} \right\} = 0,$$

with $m\{\cdot\}$ denoting the usual Lebesgue measure. Analogously, a function θ is sparsely vanishing *if*

$$\lim_{\eta \to 0} m \{ x : |\theta(x)| < \eta \} = 0.$$

We conclude this chapter with the notion of the essential range of a function, which plays an important role in the study of the asymptotic properties of the spectrum of sequences of matrices.

Definition 1.12. *Given a measurable complex-valued function θ defined on a Lebesgue measurable set G, the* essential range of θ *is the set $\mathcal{S}(\theta)$ of points $s \in \mathbb{C}$ such that, for every $\varepsilon > 0$, the Lebesgue measure of the set $\theta^{(-1)}(D(s,\varepsilon)) := \{t \in G : \theta(t) \in D(s,\varepsilon)\}$ is positive, with $D(s,\varepsilon)$ as in (1.23). The function θ is* essentially bounded *if its essential range is bounded. Furthermore, if θ is real-valued, then the essential supremum (infimum) is defined as the supremum (infimum) of its essential range. Finally if the function θ is $q \times q$ matrix-valued and measurable, then the essential range of θ is the union of the essential ranges of the complex-valued eigenvalues $\lambda_j(\theta)$, $j = 1, \ldots, q$.*

We note that $\mathcal{S}(\theta)$ is clearly a closed set (it is easy to see that its complement is open), and moreover

$$\mathcal{S}(\theta) = \cap \left\{ B : \text{closed set with } m\left\{\theta^{(-1)}(B)\right\} = m\{G\} \right\}.$$

Chapter 2

Known tools for general matrix-sequences

In this chapter we will introduce the most important tools, known in the literature, which are used to calculate the spectral distribution of sequences of matrices: starting from the main result of distribution (Theorem 2.1) and from the extensibility to combinations of matrices of the concept of *a.c.s.*, we will see some special cases of distributions of combinations of matrices and some results concerning the sequences of non-square matrices and of sparsely unbounded and sparsely vanishing sequences of matrices.

In the concluding section of this chapter some distribution results are presented concerning the perturbation of sequences of Hermitian matrices; we shall see in Chapter 3 that these results of perturbation can be "converted" in the distribution results for sequences of matrices.

2.1 Main distribution theorems for sequences of matrices

The importance of the concept of *a.c.s.* introduced in the previous chapter is well emphasized in the following theorem which is a basic result of approximation theory for matrix-sequences: the existence of a distributional result for any of the (simpler) sequences $\{B_{n,m}\}$ implies a distributional result for $\{A_n\}$, as long as $\{\{B_{n,m}\} : m \geq 0\}$ is an *a.c.s.* for $\{A_n\}$.

Theorem 2.1. [83] *Suppose a sequence of Hermitian matrices $\{A_n\}$ is given such that $A_n \in M_{d_n}(\mathbb{C})$ ($d_k < d_{k+1}$ for each k) and $\{\{B_{n,m}\} : m \geq 0\}$, $m \in \hat{\mathbb{N}} \subset \mathbb{N}$, $\#\hat{\mathbb{N}} = \infty$, $B_{n,m} \in M_{d_n}(\mathbb{C})$, is an a.c.s. for $\{A_n\}$ in the sense of Definition 1.5, with all $B_{n,m}$ being Hermitian. Suppose that $\{B_{n,m}\} \sim_\lambda (h_m, K)$ and that h_m converges in measure to the measurable function h over K, K of finite and positive measure. Then necessarily*

$$\{A_n\} \sim_\lambda (h, K). \tag{2.1}$$

If we lose the Hermitian character either of A_n or of $B_{n,m}$, then the same statement is true in full generality for the singular values, that is

$$\{A_n\} \sim_\sigma (h, K),$$

(see Definition 1.7).

The following lemma is a particular result on the distribution in the sense of the singular values for sequences of non-square matrices. This result is used in Chapter 5 to find the spectral distribution in the singular value sense for sequences of g-Toeplitz matrices.

Lemma 2.2. *Let f be a measurable complex-valued function on a set K, and consider the measurable function $\sqrt{|f|} : K \to \mathbb{R}^+$. Let $\{A_{n,m}\}$, with $A_{n,m} \in M_{d_n,d'_n}(\mathbb{C})$, $d'_n \leq d_n$, be a sequence of matrices of strictly increasing dimension: $d'_n < d'_{n+1}$ and $d_n \leq d_{n+1}$. If the*

sequence of matrices $\left\{A_{n,m}^{*}A_{n,m}\right\}$*, with* $A_{n,m}^{*}A_{n,m} \in M_{d_n'}(\mathbb{C})$ *and* $d_n' < d_{n+1}'$*, is distributed in the singular value sense as the function* f *over a suitable set* $G \subset K$ *in the sense of Definition 1.7:* $\left\{A_{n,m}^{*}A_{n,m}\right\} \sim_\sigma (f,G)$*, then the sequence* $\{A_{n,m}\}$ *is distributed in the singular value sense as the function* $\sqrt{|f|}$ *over the same* G*:* $\{A_{n,m}\} \sim_\sigma \left(\sqrt{|f|},G\right)$.

Proof. From the SVD, we can write $A_{n,m}$ as

$$A_{n,m} = U\Sigma V^{*} = U \left[\begin{array}{cccc} \sigma_1 & & & \\ & \sigma_2 & & \\ & & \ddots & \\ & & & \sigma_{d_n'} \\ \hline & & 0 & \end{array}\right] V^{*},$$

with U and V unitary matrices, $U \in M_{d_n}(\mathbb{C})$, $V \in M_{d_n'}(\mathbb{C})$, and $\Sigma \in M_{d_n,d_n'}(\mathbb{R})$, $\sigma_j \geq 0$. By multiplying $A_{n,m}^{*}A_{n,m}$ we obtain

$$\begin{aligned} A_{n,m}^{*}A_{n,m} = V\Sigma^{\top}U^{*}U\Sigma V^{*} &= V\Sigma^{\top}\Sigma V^{*} = V\Sigma^{(2)}V^{*} \\ &= V \left[\begin{array}{cccc} \sigma_1^2 & & & \\ & \sigma_2^2 & & \\ & & \ddots & \\ & & & \sigma_{d_n'}^2 \end{array}\right] V^{*}, \end{aligned} \tag{2.2}$$

with V unitary matrix, $V \in M_{d_n'}(\mathbb{C})$, and $\Sigma^{(2)} \in M_{d_n'}(\mathbb{R})$, $\sigma_j^2 \geq 0$. We observe that (2.2) is an SVD for $A_{n,m}^{*}A_{n,m}$, that is, the singular values $\sigma_j\left(A_{n,m}^{*}A_{n,m}\right)$ of $A_{n,m}^{*}A_{n,m}$ are the square of singular values $\sigma_j(A_{n,m})$ of $A_{n,m}$. Since $\left\{A_{n,m}^{*}A_{n,m}\right\} \sim_\sigma (f,G)$, by definition it follows that for every $F \in \mathcal{C}\left(\mathbb{R}_0^{+}\right)$,

$$\begin{aligned} \lim_{n\to\infty} \frac{1}{d_n'} \sum_{i=1}^{d_n'} F\left(\sigma_i\left(A_{n,m}^{*}A_{n,m}\right)\right) &= \frac{1}{m\{G\}} \int_G F\left(|f(t)|\right) \mathrm{d}t \\ &= \frac{1}{m\{G\}} \int_G H\left(\sqrt{|f(t)|}\right) \mathrm{d}t, \end{aligned} \tag{2.3}$$

where H is such that $F = H \circ \sqrt{\cdot}$, that is, $H(x) = F\left(x^2\right)$ for every x. Owing to $\sigma_j(A_{n,m}) = \sqrt{\sigma_j\left(A_{n,m}^{*}A_{n,m}\right)}$, we obtain

$$\begin{aligned} \lim_{n\to\infty} \frac{1}{d_n'} \sum_{i=1}^{d_n'} F\left(\sigma_i\left(A_{n,m}^{*}A_{n,m}\right)\right) &= \lim_{n\to\infty} \frac{1}{d_n'} \sum_{i=1}^{d_n'} F\left(\sigma_i^2(A_{n,m})\right) \\ &= \lim_{n\to\infty} \frac{1}{d_n'} \sum_{i=1}^{d_n'} H\left(\sigma_i(A_{n,m})\right). \end{aligned} \tag{2.4}$$

In conclusion, by combining (2.3) and (2.4) we deduce

$$\lim_{n\to\infty} \frac{1}{d_n'} \sum_{i=1}^{d_n'} H\left(\sigma_i(A_{n,m})\right) = \frac{1}{m\{G\}} \int_G H\left(\sqrt{|f(t)|}\right) \mathrm{d}t,$$

for every $H \in \mathcal{C}\left(\mathbb{R}_0^{+}\right)$, so $\{A_{n,m}\} \sim_\sigma \left(\sqrt{|f(t)|},G\right)$. □

2.2 Combinations of sequences of matrices

The notion of *a.c.s.* introduced in the previous chapter is stable under inversion, linear combinations and product, whenever natural and mild conditions are satisfied.

Proposition 2.3. *Let* $\left\{A_n^{(1)}\right\}$ *and* $\left\{A_n^{(2)}\right\}$ *be two sequences of matrices,* $A_n^{(i)} \in M_{d_n}(\mathbb{C})$ *(*$d_k < d_{k+1}$ *for each* k*),* $i = 1, 2$*. Suppose that*

$$\left\{\left\{B_{n,m}^{(1)}\right\} : m \geq 0\right\} \quad \text{and} \quad \left\{\left\{B_{n,m}^{(2)}\right\} : m \geq 0\right\},$$

$m \in \hat{\mathbb{N}}^{(i)} \subset \mathbb{N}$, $\#\hat{\mathbb{N}}^{(i)} = \infty$, $B_{n,m}^{(i)} \in M_{d_n}(\mathbb{C})$, $i = 1, 2$, *are two a.c.s. for* $\left\{A_n^{(1)}\right\}$ *and* $\left\{A_n^{(2)}\right\}$*, respectively. Then any linear combination of* $\left\{\left\{B_{n,m}^{(1)}\right\} : m \geq 0\right\}$ *and* $\left\{\left\{B_{n,m}^{(2)}\right\} : m \geq 0\right\}$ *is an a.c.s. for the same linear combination of* $\left\{A_n^{(1)}\right\}$ *and* $\left\{A_n^{(2)}\right\}$*; if in addition* $\left\{A_n^{(i)}\right\}$*,* $i = 1, 2$*, are sparsely unbounded matrix-sequences (see Definition 1.10), then* $\left\{\left\{B_{n,m}^{(1)} B_{n,m}^{(2)}\right\} : m \geq 0\right\}$ *is an a.c.s. for the sequence* $\left\{A_n^{(1)} A_n^{(2)}\right\}$*. Furthermore, suppose that* $\left\{A_n^{(1)}\right\}$ *is sparsely vanishing and each* $A_n^{(1)}$ *is invertible together with* $B_{n,m}^{(1)}$*. Then*

$$\left\{\left\{\left[B_{n,m}^{(1)}\right]^{-1}\right\} : m \geq 0\right\},$$

is an a.c.s. for the sequence $\left\{\left[A_n^{(1)}\right]^{-1}\right\}$*.*

The proof of the first two parts concerning linear combinations and the (component-wise) product of sequences can be found in [83]. The second part can be found in [94].

Using Proposition 2.3, we can demonstrate the results of distribution for particular products or sums of sequences of matrices.

Proposition 2.4. [83, 89] *If* $\{A_n\}$ *and* $\{B_n\}$ *are two sequences of matrices,* $A_n, B_n \in M_{d_n}(\mathbb{C})$ *(*$d_k < d_{k+1}$ *for each* k*), such that* $\{A_n\} \sim_\sigma (\theta, G)$ *and* $\{B_n\} \sim_\sigma (0, G)$*, then*

$$\{A_n + B_n\} \sim_\sigma (\theta, G).$$

Lemma 2.5. *Let* $\{A_n\}$ *and* $\{Q_n\}$ *be two sequences of matrices* $A_n, Q_n \in M_{d_n}(\mathbb{C})$ *(*$d_k < d_{k+1}$ *for each* k*), where* Q_n *are all unitary matrices* $(Q_n Q_n^* = I_n)$*. If* $\{A_n\} \sim_\sigma (0, G)$*, then* $\{A_n Q_n\} \sim_\sigma (0, G)$ *and* $\{Q_n A_n\} \sim_\sigma (0, G)$*.*

Proof. Putting $B_n = A_n Q_n$, assuming that

$$A_n = U_n \Sigma_n V_n^*,$$

is an SVD for A_n, and taking into account that the product of two unitary matrices is still a unitary matrix, we deduce that the writing

$$B_n = A_n Q_n = U_n \Sigma_n V_n^* Q_n = U_n \Sigma_n \widehat{V}_n^*,$$

is an SVD for B_n. The latter implies that A_n and B_n have exactly the same singular values, so that the two sequences $\{A_n\}$ and $\{B_n\}$ are distributed in the same way. □

If Q is unitary, then $\|Q_n\| = 1$, so Lemma 2.5 is a special case of the following lemma.

Lemma 2.6. *Let* $\{A_n\}$ *and* $\{Q_n\}$ *be two sequences of matrices,* $A_n, Q_n \in M_{d_n}(\mathbb{C})$ *(*$d_k < d_{k+1}$ *for each* k*). If* $\{A_n\} \sim_\sigma (0, G)$ *and* $\|Q_n\| \leq K$ *for some non-negative constant* K *independent of* n*, then* $\{A_n Q_n\} \sim_\sigma (0, G)$ *and* $\{Q_n A_n\} \sim_\sigma (0, G)$*.*

Proof. Since $\{A_n\} \sim_\sigma (0, G)$, then $\{0_n\}$ (sequence of zero matrices) is an *a.c.s.* for $\{A_n\}$; this means (by Definition 1.5) that we can write, for every m sufficiently large, $m \in \mathbb{N}$

$$A_n = 0_n + R_{n,m} + N_{n,m}, \qquad \forall n > n_m, \tag{2.5}$$

with

$$\text{rank}\,(R_{n,m}) \leq d_n c\,(m)\,, \qquad \|N_{n,m}\| \leq \omega\,(m)\,,$$

where $n_m \geq 0$, $c\,(m)$ and $\omega\,(m)$ depend only on m, and, moreover,

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

Now consider the matrix $A_n Q_n$; from (2.5) we obtain

$$A_n Q_n = 0_n + R_{n,m} Q_n + N_{n,m} Q_n, \qquad \forall n > n_m,$$

with

$$\begin{aligned} &\text{rank}\,(R_{n,m}Q_n) \leq \min\{\text{rank}\,(R_{n,m})\,, \text{rank}\,(Q_n)\} \leq \text{rank}\,(R_{n,m}) \leq d_n c\,(m)\,, \\ &\|N_{n,m}Q_n\| \leq \|N_{n,m}\|\,\|Q_n\| \leq K\omega\,(m)\,, \end{aligned}$$

where

$$\lim_{m\to\infty} c\,(m) = 0, \qquad \lim_{m\to\infty} K\omega\,(m) = 0,$$

then $\{0_n\}$ is an *a.c.s.* for the sequence $\{A_n Q_n\}$ and, by Theorem 2.1, $\{A_n Q_n\} \sim_\sigma (0, G)$. □

2.3 Sparsely vanishing and sparsely unbounded sequences of matrices

In Chapter 1 we have introduced the definition of sparsely vanishing and sparsely unbounded matrix-sequence and of sparsely vanishing and sparsely unbounded function. In this section we will show how these two objects (matrix-sequences and functions) are "linked" by the notion of spectral distribution, in particular we prove that a sequence of matrices $\{A_n\}$ spectrally distributed as a sparsely vanishing function is sparsely vanishing and a sequence of matrices $\{A_n\}$ spectrally distributed as a sparsely unbounded function is sparsely unbounded in the sense of Definition 1.10.

Proposition 2.7. *Let $\{A_n\}$ be a sequence of matrices, $A_n \in M_{d_n}(\mathbb{C})$ ($d_k < d_{k+1}$ for each k), spectrally distributed as a* sparsely vanishing (sparsely unbounded) *function f over the set K. Then the sequence $\{A_n\}$ is* sparsely vanishing (sparsely unbounded).

Proof. First, we consider the case of a sparsely vanishing function f. For any $\varepsilon > 0$ we define the non-negative test function

$$G_\varepsilon\,(y) = \begin{cases} \frac{y}{c} + 1 & \text{for } -c \leq y \leq 0, \\ 1 & \text{for } 0 \leq y \leq \varepsilon, \\ -\frac{y}{\varepsilon} + 2 & \text{for } \varepsilon \leq y \leq 2\varepsilon, \\ 0 & \text{otherwise.} \end{cases}$$

Now, since

$$\begin{aligned} \frac{1}{d_n}\sum_{i=1}^{d_n} G_\varepsilon\,(\sigma_i\,(A_n)) &= \frac{1}{d_n}\left[\sum_{i\in\{j:\,\sigma_j(A_n)\leq\varepsilon\}} 1 + \sum_{i\in\{j:\,\varepsilon<\sigma_j(A_n)\leq 2\varepsilon\}} G_\varepsilon\,(\sigma_i\,(A_n))\right] \\ &\geq \frac{1}{d_n}\sum_{i\in\{j:\,\sigma_j(A_n)\leq\varepsilon\}} 1, \end{aligned}$$

we find that

$$\frac{1}{d_n}\#\left\{j:\,\sigma_j\left(A_n\right)\le\varepsilon\right\}\le\frac{1}{d_n}\sum_{i=1}^{d_n}G_\varepsilon\left(\sigma_i\left(A_n\right)\right).$$

Moreover,

$$\begin{aligned}\lim_{n\to\infty}\frac{1}{d_n}\sum_{i=1}^{d_n}G_\varepsilon\left(\sigma_i\left(A_n\right)\right) &= \frac{1}{m\left\{K\right\}}\int_K G_\varepsilon\left(\left|f\left(t\right)\right|\right)\mathrm{d}t\\ &\le \frac{1}{m\left\{K\right\}}m\left\{x\in K:\,\left|f\left(x\right)\right|\le 2\varepsilon\right\}.\end{aligned}$$

By recalling that the assumption f *sparsely vanishing* implies that

$$\lim_{\eta\to 0}m\left\{x\in K:\,\left|f\left(x\right)\right|\le\eta\right\}=0,$$

the thesis follows by considering $x\left(s\right)=\frac{1}{m\{K\}}m\left\{x\in K:\,\left|f\left(x\right)\right|\le s\right\}$ in Definition 1.10.

Now, we consider the case of a sparsely unbounded function f. For any $\varepsilon>0$ we define the non-negative test function

$$F_\varepsilon\left(y\right)=\begin{cases}\frac{y}{c}+1 & \text{for } -c\le y\le 0,\\ 1 & \text{for } 0\le y\le\frac{1}{2\varepsilon},\\ -2\varepsilon y+2 & \text{for } \frac{1}{2\varepsilon}\le y\le\frac{1}{\varepsilon},\\ 0 & \text{otherwise.}\end{cases}$$

By taking into account the relations below

$$\begin{aligned}\frac{1}{d_n}\sum_{i=1}^{d_n}F_\varepsilon\left(\sigma_i\left(A_n\right)\right) &= \frac{1}{d_n}\left[\sum_{i\in\left\{j:\,\sigma_j\left(A_n\right)\le\frac{1}{2\varepsilon}\right\}}1+\sum_{i\in\left\{j:\,\frac{1}{2\varepsilon}<\sigma_j\left(A_n\right)\le\frac{1}{\varepsilon}\right\}}F_\varepsilon\left(\sigma_i\left(A_n\right)\right)\right]\\ &\le \frac{1}{d_n}\sum_{i\in\left\{j:\,\sigma_j\left(A_n\right)\le\frac{1}{\varepsilon}\right\}}1,\end{aligned}$$

we easily deduce that

$$\frac{1}{d_n}\#\left\{j:\,\sigma_j\left(A_n\right)<\frac{1}{\varepsilon}\right\}\ge\frac{1}{d_n}\sum_{i=1}^{d_n}F_\varepsilon\left(\sigma_i\left(A_n\right)\right).$$

Moreover,

$$\begin{aligned}\lim_{n\to\infty}\frac{1}{d_n}\sum_{i=1}^{d_n}F_\varepsilon\left(\sigma_i\left(A_n\right)\right) &= \frac{1}{m\left\{K\right\}}\int_K F_\varepsilon\left(\left|f\left(t\right)\right|\right)\mathrm{d}t\\ &\ge \frac{1}{m\left\{K\right\}}m\left\{x\in K:\,\left|f\left(x\right)\right|\le\frac{1}{2\varepsilon}\right\}\\ &= 1-\frac{1}{m\left\{K\right\}}m\left\{x\in K:\,\left|f\left(x\right)\right|>\frac{1}{2\varepsilon}\right\}.\end{aligned}$$

By the inequality

$$\frac{1}{d_n}\#\left\{j:\,\sigma_j\left(A_n\right)\ge\frac{1}{\varepsilon}\right\}=1-\frac{1}{d_n}\#\left\{j:\,\sigma_j\left(A_n\right)<\frac{1}{\varepsilon}\right\}\le 1-\frac{1}{d_n}\sum_{i=1}^{d_n}F_\varepsilon\left(\sigma_i\left(A_n\right)\right),$$

and by recalling that the assumption f *sparsely unbounded* implies that

$$\lim_{\eta\to 0} m\left\{x \in K:\ |f(x)| \geq \frac{1}{\eta}\right\} = 0,$$

the thesis follows by considering $x(\varepsilon) = 1 - \frac{1}{m\{K\}} m\left\{x \in K:\ |f(x)| \geq \frac{1}{2\varepsilon}\right\}$ in Definition 1.10.

It is worth noticing that essentially the same proof applies in the case of a sequence of Hermitian matrices with a real-valued function f when considering the eigenvalues instead of the singular values. The only change is in the definition of the test functions F_ε and G_ε: in fact it is enough to take new test functions $\hat{T}_\varepsilon = \hat{T}_\varepsilon(y)$ that coincides with $T_\varepsilon(y)$ if the argument y is non-negative and coincides with $T_\varepsilon(-y)$ otherwise. Here the symbol "T" means "F" or "G" according to the previous notations. □

The following result is very useful in practical manipulations in order to give norm bounds from above.

Lemma 2.8. *Consider a sequence of matrices* $\{A_n\}$, $A_n \in M_{d_n}(\mathbb{C})$ *($d_k < d_{k+1}$ for each k). The following are equivalent.*

- *The sequence* $\{A_n\}$ *is* sparsely unbounded.
- *There exists a non-negative function* $x(s)$ *with* $\lim_{s\to\infty} x(s) = 0$ *so that* $\forall \varepsilon > 0\ \exists n_\varepsilon \in \mathbb{N}$ *such that* $\forall n \geq n_\varepsilon$ *it holds that* $A_n = R_n + L_n$, *where* $\|R_n\| < \frac{1}{\varepsilon}$ *and* $\operatorname{rank}(L_n) \leq x(\varepsilon)\, d_n$.

Proof. The result trivially follows by using the singular value decomposition properties of the involved matrices and the singular values interlacing properties [46]. □

The next technical lemmas deal with the product and the inversion of sparsely unbounded and sparsely vanishing sequences of matrices and will be useful for performing the spectral analysis of preconditioned matrices in Chapter 6.

Lemma 2.9. *Let* $\{A_n\}$ *and* $\{B_n\}$ *be two* sparsely unbounded *matrix-sequences,* $A_n, B_n \in M_{d_n}(\mathbb{C})$ *($d_k < d_{k+1}$ for each k). Then the sequences* $\{A_nB_n\}$ *and* $\{A_n + B_n\}$ *are* sparsely unbounded *(the latter implies that the notion* sparsely unbounded *sequence is stable under linear combinations).*

Proof. Under these assumptions, we can consider the following splittings

$$\begin{aligned} A_n &= \hat{R}_n + \hat{L}_n, \\ B_n &= \tilde{R}_n + \tilde{L}_n, \end{aligned}$$

where $\forall \hat{\delta} > 0\ \exists n_{\hat{\delta}} \in \mathbb{N}$ such that $\forall n \geq n_{\hat{\delta}}$ it holds that $\left\|\hat{R}_n\right\| < \frac{1}{\hat{\delta}}$ and $\operatorname{rank}\left(\hat{L}_n\right) \leq \hat{x}\left(\hat{\delta}\right) d_n$ with $\lim_{s\to 0} \hat{x}(s) = 0$ and where $\forall \tilde{\delta} > 0\ \exists n_{\tilde{\delta}} \in \mathbb{N}$ such that $\forall n \geq n_{\tilde{\delta}}$ it holds that $\left\|\tilde{R}_n\right\| < \frac{1}{\tilde{\delta}}$ and $\operatorname{rank}\left(\tilde{L}_n\right) \leq \tilde{x}\left(\tilde{\delta}\right) d_n$ with $\lim_{s\to 0} \tilde{x}(s) = 0$ (see Lemma 2.8). Therefore, the matrices A_nB_n can be written as

$$A_nB_n = R_n + L_n,$$

with

$$\begin{aligned} R_n &= \tilde{R}_n\hat{R}_n, \\ L_n &= \tilde{L}_n\left(\hat{R}_n + \hat{L}_n\right) + \tilde{R}_n\hat{L}_n, \end{aligned}$$

where, for n large enough, we find

$$\begin{aligned} \|R_n\| &< \frac{1}{\tilde{\delta}\hat{\delta}}, \\ \operatorname{rank}(L_n) &\leq \left(\tilde{x}\left(\tilde{\delta}\right) + \hat{x}\left(\hat{\delta}\right)\right) d_n. \end{aligned}$$

For the arbitrariness of $\tilde{\delta}$ and $\hat{\delta}$ the first part of the claimed thesis follows by virtue of Lemma 2.8.

The matrices $A_n + B_n$ can be written as

$$A_n + B_n = \breve{R}_n + \breve{L}_n,$$

with

$$\begin{aligned} \breve{R}_n &= \tilde{R}_n + \hat{R}_n, \\ \breve{L}_n &= \tilde{L}_n + \hat{L}_n, \end{aligned}$$

where, for n large enough, we find

$$\begin{aligned} \left\|\breve{R}\right\| &< \frac{1}{\tilde{\delta}} + \frac{1}{\hat{\delta}} < 2\left(\min\left\{\tilde{\delta}, \hat{\delta}\right\}\right)^{-1}, \\ \operatorname{rank}\left(\breve{L}_n\right) &\leq \left(\tilde{x}\left(\tilde{\delta}\right) + \hat{x}\left(\hat{\delta}\right)\right) d_n. \end{aligned}$$

For the arbitrariness of $\tilde{\delta}$ and $\hat{\delta}$ the second part of the claimed thesis follows again by Lemma 2.8. □

Remark 2.10. *Lemma 2.9 tells us that the set of* sparsely unbounded *sequences forms an algebra (that is closed under linear combinations and products). On the other side, Lemma 2.14 can be read by saying that the set of sequences which are* clustered at zero *forms a two-sided ideal in the algebra of* sparsely unbounded *sequences.*

Lemma 2.11. *Let $\{A_n\}$ be a sequence of invertible matrices, $A_n \in M_{d_n}(\mathbb{C})$ ($d_k < d_{k+1}$ for each k). If the sequence $\{A_n\}$ is* sparsely vanishing *then the sequence $\{A_n^{-1}\}$ is* sparsely unbounded *and vice versa.*

Proof. The result trivially follows by using the singular value decomposition properties of the involved matrices. □

Lemma 2.12. *Let $\{A_n\}$ and $\{B_n\}$ be two* sparsely vanishing *matrix-sequences of invertible matrices, $A_n, B_n \in M_{d_n}(\mathbb{C})$ ($d_k < d_{k+1}$ for each k). Then the sequence $\{A_nB_n\}$ is* sparsely vanishing. *This is not true for the sequence $\{A_n + B_n\}$, that is, the notion* sparsely vanishing *sequence is not stable under linear combinations.*

Proof. Since $\{A_n\}$ and $\{B_n\}$ are both sequences of invertible matrices, from $(A_nB_n)^{-1} = (B_n)^{-1}(A_n)^{-1}$, the first part trivially follows from Lemma 2.9 by recalling Lemma 2.11. The second part is straightforward by considering $B_n = -A_n$, so that $A_n + B_n \equiv 0$ is not sparsely vanishing. □

Remark 2.13. *The assumption of invertibility in Lemma 2.11 and Lemma 2.12 can be removed by considering the pseudo-inverse of Moore-Penrose* [64, 73] *instead of the usual inverse matrix.*

Lemma 2.14. *Let $\{A_n\}$ and $\{B_n\}$ be two matrix-sequences, $A_n, B_n \in M_{d_n}(\mathbb{C})$ ($d_k < d_{k+1}$ for each k). Suppose that the sequence $\{A_n\}$ is* sparsely unbounded *and the sequence $\{B_n\}$ is clustered at 0 with respect to the singular values (see Definition 1.8). Then both the sequences $\{A_nB_n\}$ and $\{B_nA_n\}$ are clustered at 0.*

Proof. Under these assumptions, we have that $\forall \hat{\varepsilon} > 0\ \exists n_{\hat{\varepsilon}} \in \mathbb{N}$ such that $\forall n \geq n_{\hat{\varepsilon}}$ it holds that

$$A_n = R_n + L_n,$$

where $\|R_n\| < \frac{1}{\hat{\varepsilon}}$ and $\operatorname{rank}(L_n) \leq x(\hat{\varepsilon})\, d_n$ with $\lim_{s\to 0} x(s) = 0$ and $\forall \varepsilon > 0\ \exists n_\varepsilon \in \mathbb{N}$ such that $\forall n \geq n_\varepsilon$ we have

$$B_n = N_n + P_n,$$

where $\|N_n\| \leq \varepsilon$ and $\operatorname{rank}(P_n) \leq y(\varepsilon)\, d_n$ with $\lim_{s\to 0} y(s) = 0$. Now, by splitting the matrices as

$$A_n B_n = \tilde{N}_n + \tilde{P}_n,$$

with

$$\begin{aligned} \tilde{N}_n &= R_n N_n, \\ \tilde{P}_n &= R_n P_n + L_n (N_n + P_n), \end{aligned}$$

where

$$\begin{aligned} \left\|\tilde{N}_n\right\| &< \frac{\varepsilon}{\hat{\varepsilon}}, \\ \operatorname{rank}\left(\tilde{P}_n\right) &\leq (x(\hat{\varepsilon}) + y(\varepsilon))\, d_n, \end{aligned}$$

and for the arbitrariness of $\hat{\varepsilon}$ and ε, by choosing $\hat{\varepsilon} = \sqrt{\varepsilon}$, the desired result plainly follows. The case $\{B_n A_n\}$ can be proved in the same manner. □

2.4 Some distribution results

In [45, Theorem 2.2] Golinskii and Serra-Capizzano address the problem of finding new instruments, different from Theorem 2.1, to derive the spectral distribution (in the sense of eigenvalues) for sequences of matrices $\{A_n\}$ bounded in the operator norm; in particular, they try to find a way for relating formula (1.20), with F being an arbitrary polynomial, to the same formula in its full extent, i.e., with $F \in \mathcal{C}_0(\mathbb{C})$ being a continuous function.

In order to do this, they use the Mergelyan's Theorem, which requires some hypothesis on the essential range of the symbol θ and a priori assumptions on the clustering properties of the sequence $\{A_n\}$. The reason for these requirements is in part due to the barrier given by the Mergelyan's Theorem stating that the closure in the uniform norm of the polynomials on a compact set S is given by the set of all continuous functions on S, which are holomorphic in its interior, provided that $\mathbb{C}\backslash S$ is connected (for the proof see [76, Theorem 20.5]). Therefore, the polynomial space is able to approximate every continuous function on S if and only if S has empty interior in $\mathbb{C}$ and $\mathbb{C}\backslash S$ is connected.

Now we rewrite the Theorem 2.2 from [45] in a slightly different, but equivalent way. The basic ideas used here come from the paper [112], where the same questions were considered in a different context (but they were also known in a certain form to the operator theory community (see [120, top of p. 39]) and were extensively developed by Böttcher, Roch, SeLegue, Silbermann etc., see [20]).

Theorem 2.15. [45] *Let $\{A_n\}$ be a matrix-sequence, $A_n \in M_{d_n}(\mathbb{C})$ ($d_k < d_{k+1}$ for each k), and S a subset of $\mathbb{C}$. If:*

(a1) *S is a compact set and $\mathbb{C}\backslash S$ is connected;*

(a2) *the matrix-sequence $\{A_n\}$ is weakly clustered at S;*

(a3) *the spectra* $\Lambda(A_n)$ *of* A_n *are uniformly bounded, i.e.,* $\exists C \in \mathbb{R}^+$ *such that* $|\lambda| < C$, $\lambda \in \Lambda(A_n)$, *for all* n;

(a4) *there exists a function* θ *measurable, bounded, and defined on a set* G *of positive and finite Lebesgue measure, such that, for every non-negative integer* L, *we have*

$$\lim_{n\to\infty} \frac{\operatorname{tr}\left(A_n^L\right)}{d_n} = \frac{1}{m\{G\}} \int_G \theta^L(t)\,\mathrm{d}t,$$

i.e., relation (1.20) holds with F *being any polynomial of an arbitrary fixed degree;*

(a5) *the essential range of* θ *(see Definition 1.12) is contained in* S;

then relation (1.20) is true for every continuous function F *with bounded support which is holomorphic in the interior of* S. *If it is also true that the interior of* S *is empty then the sequence* $\{A_n\}$ *is distributed as* θ, *on its domain* G, *in the sense of the eigenvalues.*

The following theorem ([45, Theorem 2.4]) will allow us to deduce weak clustering and strong attraction for sequences of matrices distributed as a measurable function θ in the sense of Definition 1.7.

Theorem 2.16. *Let* θ *be a measurable function defined on* G *with finite and positive Lebesgue measure, and* $\mathcal{S}(\theta)$ *be the essential range of* θ. *Let* $\{A_n\}$ *be a matrix-sequence distributed as* θ *in the sense of eigenvalues; in that case the following facts are true:*

a) $\mathcal{S}(\theta)$ *is a weak cluster for* $\{A_n\}$;

b) *each point* $s \in \mathcal{S}(\theta)$ *strongly attracts the spectra* $\Lambda(A_n)$ *with infinite order* $r(s) = \infty$;

c) *there exists a sequence* $\left\{\lambda^{(n)}\right\}$, *where* $\lambda^{(n)}$ *is an eigenvalue of* A_n, *such that*

$$\liminf_{n\to\infty} \left|\lambda^{(n)}\right| \geq \|\theta\|_{L^\infty}.$$

The same statements hold in the case of a $q \times q$ *matrix-valued function* θ.

Proof. For items **a)** and **b)** see [45, Theorem 2.4], for a proof. Then notice that, by **b)**, each point $s \in \mathcal{S}(\theta)$ is a limit of a sequence $\left\{\lambda^{(n)}\right\}$ where $\lambda^{(n)}$ is an eigenvalue of A_n. Hence item **c)** follows from the definition of $\mathcal{S}(\theta)$. The extension to the matrix-valued case is trivial. □

The next result ([45, Theorem 3.4]), based on Mirsky theorem (see [15, Proposition III, Section 5.3]), establishes a link between distributions of non-Hermitian perturbations of Hermitian matrix-sequences and the distribution of the original sequence.

Theorem 2.17. [119], [45, Theorem 3.4] *Let* $\{B_n\}$ *and* $\{C_n\}$ *be two matrix-sequences,* $B_n, C_n \in M_{d_n}(\mathbb{C})$ *(*$d_k < d_{k+1}$ *for each* k*), where* B_n *is Hermitian and* $A_n = B_n + C_n$. *Assume further that* $\{B_n\}$ *is distributed as* (θ, G) *in the sense of the eigenvalues, where* G *is of finite and positive Lebesgue measure, both* $\|B_n\|$ *and* $\|C_n\|$ *are uniformly bounded by a positive constant* $\widehat{C}$ *independent of* n, *and* $\|C_n\|_1 = o(d_n)$, $n \to \infty$. *Then* θ *is real-valued and* $\{A_n\}$ *is distributed as* (θ, G) *in the sense of the eigenvalues. In particular, if* $\mathcal{S}(\theta)$ *is the essential range of* θ, *then* $\{A_n\}$ *is weakly clustered at* $\mathcal{S}(\theta)$, *and* $\mathcal{S}(\theta)$ *strongly attracts the spectra of* $\{A_n\}$ *with an infinite order of attraction for any of its points.*

We conclude this section with a theorem which is a slight extension of a theorem from [45] concerning strong clustering.

Theorem 2.18. *Let $\{B_n\}$ and $\{C_n\}$ be two matrix-sequences, $B_n, C_n \in M_{d_n}(\mathbb{C})$ ($d_k < d_{k+1}$ for each k), where B_n is Hermitian and $A_n = B_n + C_n$. Let E be a compact subset of the real line. Assume that $\{B_n\}$ is strongly clustered at E, $\|C_n\|_1 = O(1)$, $n \to \infty$ and $\|A_n\|$ is uniformly bounded by a positive constant $\widehat{C}$ independent of n. Then $\{A_n\}$ is strongly clustered at E.*

Proof. The case where the compact set E is a union of m disjoint closed intervals (possibly, degenerate) has been treated in [45, Theorem 3.6] and a similar result has been established without using the Mirsky theorem in [119]. The general case follows since, for the notion of strong clustering we have to consider the ϵ fattening of E, or $D(E, \epsilon)$ defined as in relation (1.24). It is clear that for every compact set E, the closure of $D(E, \epsilon)$ is a finite union of closed intervals and so the general case is reduced to that handled in [45]. □

Chapter 3

New tools for general matrix-sequences

In this chapter we enlarge the set of tools, presented in Chapter 2, for proving the existence and for characterizing explicitly the limit distribution of eigenvalues and singular values for general (structured) matrix-sequences.

The new results presented in this chapter are obtained as generalizations of some theorems and Propositions in the previous chapter: ranging from the spectral distribution in the sense of eigenvalues for sequences of non-Hermitian matrices (extension of Theorem 2.1), to the generalization of the concept of *a.c.s.* for sequences of functions of Hermitian matrices (extension of Proposition 2.3).

In Section 3.3 we present some variants of Theorem 2.15 that will be used in Chapter 4 to find the spectral distribution of sequences of products of Toeplitz matrices.

3.1 Generalization of Theorem 2.1

In Theorem 2.1 we have seen that if we lose the Hermitian character either of A_n or of $B_{n,m}$, then the same statements as (2.1) is true in full generality for the singular values (see [83, 90] for details, more results and applications), but become false in general when considering the eigenvalues; see [98] for a striking counterexample.

The goal of this section is to give new more restrictive conditions under which a more severe notion of approximating class of sequences still enables to derive the spectral distribution of a "difficult" sequence from those of simpler approximating sequences.

More precisely, under mild trace norm assumptions on the perturbing sequence and taking into consideration the definition of *a.c.s.*, we extend the perturbation result of Theorem 2.17. The analysis concerns the localization and the distribution of the eigenvalues of a generic (non-Hermitian) complex perturbation of a bounded Hermitian sequence of matrices.

Theorem 3.1. *Let $\{\{B_{n,m}\} : m \geq 0\}$, $m \in \hat{\mathbb{N}} \subset \mathbb{N}$, $\#\hat{\mathbb{N}} = \infty$ be an a.c.s. for $\{A_n\}$, with $A_n, B_{n,m} \in M_{d_n}(\mathbb{C})$ ($d_k < d_{k+1}$ for each k), such that $E_{n,m} = N_{n,m} + R_{n,m}$, $B_{n,m}$ are Hermitian and*

$$\begin{aligned} &\{B_{n,m}\} \sim_\lambda (h_m, K), \qquad 0 < m\{K\} < \infty, \\ &\lim_{m\to\infty} h_m = h \quad \text{in measure on } K, \end{aligned} \tag{3.1}$$

with

$$\begin{aligned} &\sup_m \sup_n \|B_{n,m}\| = \widetilde{C}, \\ &\sup_m \sup_n \|E_{n,m}\| = \widehat{C}, \\ &C = \max\left\{\widetilde{C}, \widehat{C}\right\}. \end{aligned}$$

Here $\widetilde{C}$, $\widehat{C}$ are positive universal constants, $\|E_{n,m}\|_1 \leq c(m)\, d_n$ with $\lim_{m\to\infty} c(m) = 0$.

Then h is real-valued and $\{A_n\}$ is distributed in the eigenvalue sense as h, i.e., $\{A_n\} \sim_\lambda (h, K)$ or, equivalently,

$$\lim_{n\to\infty} \Sigma_\lambda(F, A_n) = \frac{1}{m\{K\}} \int_K F(h(x))\, \mathrm{d}x, \tag{3.2}$$

$\forall F \in \mathcal{C}_0(\mathbb{C})$.

Proof. We define the functionals acting on $\mathcal{C}_0(\mathbb{C})$ as follows

$$\Phi_m(F) = \frac{1}{m\{K\}} \int_K F(h_m(x))\, \mathrm{d}x,$$
$$\Phi(F) = \frac{1}{m\{K\}} \int_K F(h(x))\, \mathrm{d}x,$$

where the function F is continuous with bounded support, i.e., $F \in \mathcal{C}_0(\mathbb{C})$. It is immediate to check that relation (3.2) is equivalent to write that $\forall \epsilon > 0$, $\exists \overline{n} > 0$ such that $\forall n \geq \overline{n}$ and $\forall F \in \mathcal{C}_0(\mathbb{C})$ we have $|\Sigma_\lambda(F, A_n) - \Phi(F)| < \epsilon$. By allowing the parameter m, the latter is equivalent to state that there exists a non-negative function $k(m)$ with $\lim_{m\to\infty} k(m) = 0$ such that $\forall F \in \mathcal{C}_0(\mathbb{C})$, $\forall m \in \hat{\mathbb{N}}$, $\exists \overline{n}_m \in \mathbb{N}$ and the inequalities

$$|\Sigma_\lambda(F, A_n) - \Phi(F)| < k(m), \qquad \forall n \geq \overline{n}_m, \tag{3.3}$$

are fulfilled. Let us consider the left-hand side of (3.3) and let us decompose it in basic quantities to be studied separately. In fact, by proper manipulations we find

$$\begin{aligned} |\Sigma_\lambda(F, A_n) - \Phi(F)| &= |\Sigma_\lambda(F, A_n) - \Sigma_\lambda(F, B_{n,m}) + \\ &\quad + \Sigma_\lambda(F, B_{n,m}) - \Phi_m(F) + \\ &\quad + \Phi_m(F) - \Phi(F)| \\ &\leq \alpha_{n,m} + \beta_{n,m} + \gamma_m, \end{aligned}$$

where

$$\begin{aligned} \alpha_{n,m} &= |\Sigma_\lambda(F, A_n) - \Sigma_\lambda(F, B_{n,m})|, \\ \beta_{n,m} &= |\Sigma_\lambda(F, B_{n,m}) - \Phi_m(F)|, \\ \gamma_m &= |\Phi_m(F) - \Phi(F)|. \end{aligned}$$

First we focus on the quantities $\beta_{n,m}$. From the assumptions in (3.1), for all fixed m, $\beta_{n,m}$ converges to zero as $n \to \infty$; then we can take a value $\widehat{n}_m$ sufficiently large in such a way that $\beta_{n,m} \leq \frac{1}{m}$ $\forall n \geq \widehat{n}_m$. In other words we can write $\limsup_{n\to\infty} \beta_{n,m} \leq \frac{1}{m}$. By following the same reasoning on the quantities γ_m, by (3.1) and from [100, Remark 5.1.3], $\lim_{m\to\infty} \Phi_m(F) = \Phi(F)$, so we can write $\gamma_m = |\Phi_m(F) - \Phi(F)| \leq \upsilon(m)$, with $\lim_{m\to\infty} \upsilon(m) = 0$.

As a consequence, the proof of (3.3) (and a fortiori of (3.2)) is reduced to check whether $\lim_{m\to\infty} \left[\limsup_{n\to\infty} \alpha_{n,m}\right] = 0$, that is

$$|\Sigma_\lambda(F, A_n) - \Sigma_\lambda(F, B_{n,m})| \leq \delta(m), \quad \text{with } \lim_{m\to\infty} \delta(m) = 0. \tag{3.4}$$

Now, remembering that, as mentioned above, to check the relation (3.2) is sufficient to prove that (3.4) is true, we continue the proof by verifying that the assumptions of Theorem 2.15 are met, that is:

I. the spectrum of all A_n is uniformly bounded, that is $\exists Q$ positive constant such that $|\lambda_j(A_n)| < Q\ \forall n$ ($\lambda_j \in \Lambda(A_n)$ where $\Lambda(A_n)$ is defined in (1.4));

II. relation (3.2) is verified whenever F is a polynomial of arbitrary fixed degree;

III. the sequence $\{A_n\}$ is weakly clustered, in the eigenvalue sense, at a compact set $S \subset \mathbb{C}$ with empty interior, such that $\mathbb{C}\backslash S$ is a connected set, and $\mathcal{S}(h) \subset S$, with $\mathcal{S}(h)$ denoting the essential range of h (see Definition 1.12).

Item I. From the assumptions we have

$$\|A_n\| = \|B_{n,m} + E_{n,m}\| \leq \|B_{n,m}\| + \|E_{n,m}\| \leq 2C, \qquad \forall n, m,$$

and hence the spectra of the sequences $\{A_n\}$, $\{B_{n,m}\}$, and $\{E_{n,m}\}$ lie all in the closed disk $\{|z| \leq 2C\}$. In particular, the spectrum of all A_n is uniformly bounded since $\forall n\ |\lambda_j(A_n)| \leq 2C$, $2C$ constant independent of n, $\lambda_j \in \Lambda(A_n)$.

Item II. Since

$$\mathrm{tr}(X) = \sum_{\lambda \in \Lambda(X)} \lambda = \sum_{k=1}^{d_n} [X]_{k,k},$$

and since $\mathrm{tr}(\cdot)$ is a linear functional, the assumption $A_n = B_{n,m} + E_{n,m}$ implies that

$$\mathrm{tr}(A_n) - \mathrm{tr}(B_{n,m}) = \mathrm{tr}(E_{n,m}).$$

Consequently

$$\begin{aligned}
\left| \frac{1}{d_n} \sum_{\lambda \in \Lambda(A_n)} \lambda - \frac{1}{d_n} \sum_{\lambda \in \Lambda(B_{n,m})} \lambda \right| &= \left| \frac{1}{d_n} \sum_{\lambda \in \Lambda(E_{n,m})} \lambda \right| \\
&\underset{\text{(a)}}{\leq} \frac{1}{d_n} \|E_{n,m}\|_1 \\
&\underset{\text{(b)}}{\leq} \frac{1}{d_n} c(m)\, d_n = c(m), \quad \text{with } \lim_{m \to \infty} c(m) = 0,
\end{aligned}$$

where (a) follows from (1.14) and (b) follows from the assumptions; therefore, by invoking also (3.4), we deduce that (3.2) is satisfied in the special case where $F(z) = z$ (defined on the whole complex field $\mathbb{C}$, hence with non-compact support, but which can be considered an admissible test function since the spectra of all A_n are uniformly bounded).

We now prove that (3.2) is satisfied by taking as test function F any arbitrary polynomial of fixed degree. To this end, from the linearity in the first variable of the operators $\Sigma_\lambda(\cdot,\cdot)$ and $\Phi(\cdot)$ (i.e. $\Sigma_\lambda(aG + bH, \cdot) = a\Sigma_\lambda(G,\cdot) + b\Sigma_\lambda(H,\cdot)$ and $\Phi(aG + bH) = a\Phi(G) + b\Phi(H)$), it is sufficient to consider the case of monomials, i.e., $F(z) = z^q$ for all non-negative integers q. For $q = 0, 1$ the result is valid, so that we focus our attention to the case where $q \geq 2$. Relation $A_n = B_{n,m} + E_{n,m}$ implies

$$A_n^q = (B_{n,m} + E_{n,m})^q = B_{n,m}^q + \widetilde{E}_{n,m},$$

where $\widetilde{E}_{n,m}$ is a term of the form

$$\widetilde{E}_{n,m} = \sum_{X_i \in \{B_{n,m}, E_{n,m}\}} (X_1 \cdots X_q) - B_{n,m}^q. \tag{3.5}$$

In other words, the error matrix $\widetilde{E}_{n,m}$ is the sum of all possible combinations of products of j matrices $B_{n,m}$ and k matrices $E_{n,m}$, with $j + k = q$ and the exception of $j = q$ (obviously it is understood that all the addends are pair-wise different). By using a simple Hölder's inequality

involving Schatten p-norms (see (1.13)), for every summand R in (3.5), we deduce that there exists $j \geq 1$, $k = q - j$ for which

$$\begin{aligned} \|R\|_1 &\leq \|B_{n,m}\|^k \|E_{n,m}\|^{j-1} \|E_{n,m}\|_1 \\ &\leq C^k C^{j-1} c(m)\, d_n. \end{aligned} \tag{3.6}$$

Therefore by the triangle inequality and by applying inequality (3.6) to any summand in (3.5), we find $\left\|\widetilde{E}_{n,m}\right\|_1 \leq \widehat{K} c(m)\, d_n$, with $\widehat{K} = \widehat{K}(q)$ constant independent of n and m. Consequently $\mathrm{tr}(A_n^q) - \mathrm{tr}\left(B_{n,m}^q\right) = \mathrm{tr}\left(\widetilde{E}_{n,m}\right)$, and, since $\lambda(X^q) = \lambda^q(X)$, we have

$$\begin{aligned} \left|\frac{1}{d_n}\sum_{\lambda\in\Lambda(A_n)}\lambda^q - \frac{1}{d_n}\sum_{\lambda\in\Lambda(B_{n,m})}\lambda^q\right| &= \left|\frac{1}{d_n}\sum_{\lambda\in\Lambda\left(\widetilde{E}_{n,m}\right)}\lambda\right| \\ &\leq \frac{1}{d_n}\left\|\widetilde{E}_{n,m}\right\|_1 \\ &\leq \frac{1}{d_n}\widehat{K} c(m)\, d_n = \widehat{K} c(m), \quad \text{with } \lim_{m\to\infty} c(m) = 0. \end{aligned}$$

The latter, joint with relation (3.4), proves that (3.2) is satisfied with $F(z) = z^q$ for any non-negative integer q and, a fortiori, with any polynomial F of fixed degree.

Item III. According to the standard notations in (1.1)–(1.3), we write the matrix $E_{n,m}$ as

$$E_{n,m} = \mathrm{Re}(E_{n,m}) + \mathrm{iIm}(E_{n,m}).$$

Clearly we have

$$\|\mathrm{Re}(E_{n,m})\|_1 \leq \|E_{n,m}\|_1 \leq c(m)\, d_n, \tag{3.7}$$

$$\|\mathrm{Im}(E_{n,m})\|_1 \leq \|E_{n,m}\|_1 \leq c(m)\, d_n. \tag{3.8}$$

We recall that every matrix $B_{n,m}$ is Hermitian and the same is obviously true for $\mathrm{Re}(A_n)$ and $\mathrm{Re}(E_{n,m})$. Since using (3.7) we have

$$\|\mathrm{Re}(A_n) - B_{n,m}\|_1 = \|\mathrm{Re}(E_{n,m})\|_1 \leq c(m)\, d_n,$$

from [100, Lemma 5.1.3 and Corollary 5.1.2] we deduce that $\{\{B_{n,m}\} : m \geq 0\}$ is an *a.c.s.* for the sequence $\{\mathrm{Re}(A_n)\}$. From (3.1) and Theorem 2.1, it follows that

$$\{\mathrm{Re}(A_n)\} \sim_\lambda (h, K).$$

As a consequence from Theorem 2.16, $\{\mathrm{Re}(A_n)\}$ is weakly clustered at the essential range $\mathcal{S}(h)$ of h, which is a compact subset of $[-2C, 2C]$ (recall that $\max_{j=1,\ldots,d_n} |\lambda_j(\mathrm{Re}(A_n))| = \|\mathrm{Re}(A_n)\| \leq \|A_n\| \leq 2C$). Therefore all the eigenvalues of the Hermitian matrix $\mathrm{Re}(A_n)$ belong to the same interval $[-2C, 2C]$.

We now consider the matrix $\mathrm{Im}(A_n) = \mathrm{Im}(E_{n,m})$. From (3.8) we have

$$\|\mathrm{Im}(A_n)\|_1 = \|\mathrm{Im}(E_{n,m})\|_1 \leq c(m)\, d_n,$$

from [100, Lemma 5.1.3] it follows that $\{0_n\}$ (sequence of null matrices) is an *a.c.s.* for the sequence $\{\mathrm{Im}(A_n)\}$, and since $\{0_n\} \sim_\lambda (0, K)$, from Theorem 2.1, we deduce that

$$\{\mathrm{Im}(A_n)\} \sim_\lambda (0, K),$$

then Theorem 2.16 implies that $\{\mathrm{Im}(A_n)\}$ is weakly clustered a $\mathcal{S}(0) = \{0\}$. Therefore, by the definition of weak cluster (see Definition 1.8), for all $\epsilon > 0$, we obtain that

$$\#\{j : \lambda_j(\mathrm{Im}(A_n)) \notin D(0, \epsilon)\} = o(d_n). \tag{3.9}$$

Finally, from Ky Fan-Mirsky Theorem (see [45, Theorem 3.1]), and from (3.9), we plainly deduce

$$\#\left\{j \,:\, \mathrm{Im}\left(\lambda_j\left(A_n\right)\right) \notin D\left(0,\epsilon\right)\right\} = o\left(d_n\right).$$

Consequently from [45, (17), (18) and Corollary 3.3] we infer that all eigenvalues of A_n show real parts in the interval $[-2C, 2C]$ and

$$\#\left\{j \,:\, \lambda_j\left(A_n\right) \notin D\left(\left[-2C, 2C\right], \epsilon\right)\right\} = o\left(d_n\right),$$

i.e., $\{A_n\}$ is weakly clustered at the compact set $[-2C, 2C]$, the interior of $[-2C, 2C]$ is empty as subset of $\mathbb{C}$, $\mathbb{C}\backslash\left[-2C, 2C\right]$ is connected in $\mathbb{C}$, and $\mathcal{S}\left(h\right) \subset [-2C, 2C]$.

Since all the assumptions of Theorem 2.15 are met, we conclude that $\{A_n\}$ is distributed, in the eigenvalue sense, as the function h in its definition domain K. □

We see application of this theorem to the case of boundary value problems (BVPs) approximated by Finite Difference (FD) schemes. The idea is very general and can be applied to other local approximation methods as finite elements or finite volumes on regions of $\mathbb{R}^d$, $d \geq 1$. However, for the sake of clarity, we choose a single elliptic one-dimensional BVP for illustrating the general strategy. Let a, b, c, and γ be four given continuous functions on $[0, 1]$ and let us consider the second-order differential equation $-\left(au'\right)' + bu' + c = \gamma$ on $(0, 1)$, with given Neumann-Dirichlet boundary conditions, i.e., $u\left(0\right) = \alpha, u'\left(1\right) = 0$. We consider the centered FD formula of minimal bandwidth, with precision order two and on the equi-spaced grid $\{x_j\}_{j=0}^{n+1}$, $x_j = jh$, $h = \left(n+1\right)^{-1}$.

The coefficient matrix of the resulting linear system has the form

$$X_n = A'_n\left(a\right) + R_n, \qquad A'_n\left(a\right) = A_n\left(a\right) - a_{n-\frac{1}{2}} e_n e_{n-1}^{\top},$$

where

$$A_n\left(a\right) = \begin{bmatrix} a_{\frac{1}{2}} + a_{\frac{3}{2}} & -a_{\frac{3}{2}} & & & & & \\ -a_{\frac{3}{2}} & a_{\frac{3}{2}} + a_{\frac{5}{2}} & -a_{\frac{5}{2}} & & & & \\ & -a_{\frac{5}{2}} & \ddots & \ddots & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & \ddots & \ddots & -a_{n-\frac{1}{2}} & \\ & & & & -a_{n-\frac{1}{2}} & a_{n-\frac{1}{2}} + a_{n+\frac{1}{2}} \end{bmatrix},$$

and $R_n = h^2 \underset{j=1,\ldots,n}{\mathrm{diag}}\left(c_j\right) + h \underset{j=1,\ldots,n}{\mathrm{diag}}\left(b_j\right) T_n\left(\mathrm{i}\sin\left(t\right)\right)$, $v_j = v\left(x_j\right)$, $v \in \{a, b, c\}$ (the matrix $T_n\left(\mathrm{i}\sin\left(t\right)\right)$ is the Toeplitz matrix generated by the function $\mathrm{i}\sin\left(t\right)$, see Definition 4.1).

From the previous representation and taking into account that $nh < 1$ and $\left\|T_n\left(\mathrm{i}\sin\left(t\right)\right)\right\| \leq \left\|\mathrm{i}\sin\left(t\right)\right\|_{L^\infty} \leq 1$ ((see [4] and [96, Corollary 4.2])), it easy to check that

$$\left\| h^2 \underset{j=1,\ldots,n}{\mathrm{diag}}\left(c_j\right) \right\| \leq h^2 \left\|c\right\|_{L^\infty}, \quad \left\| h \underset{j=1,\ldots,n}{\mathrm{diag}}\left(b_j\right) T_n\left(\mathrm{i}\sin\left(t\right)\right) \right\| \leq h \left\|b\right\|_{L^\infty},$$

and

$$\left\|A'_n\left(a\right) - A_n\left(a\right)\right\| \leq \left\|a\right\|_{L^\infty},$$

with $\mathrm{rank}(A'_n\left(a\right) - A_n\left(a\right)) \leq 1$, and hence the coefficient matrix X_n equals $A_n\left(a\right)$ plus a correction whose trace norm is bounded by $\left\|a\right\|_{L^\infty} + \left\|b\right\|_{L^\infty} + h\left\|c\right\|_{L^\infty}$. Furthermore, as clearly analyzed in [111], the Hermitian (real symmetric indeed) sequence $\{A_n\left(a\right)\}$ is distributed as $a\left(x\right)\left(2 - 2\cos\left(s\right)\right)$ over $(0, 1) \times (0, \pi)$. As a consequence, by Theorem 3.1, it follows that the

non-Hermitian sequence $\{X_n\}$ shares the same distribution function with $\{A_n(a)\}$. In addition, since the trace norm of the correction $X_n - A_n(a)$ is bounded by a constant, for every fixed $\epsilon > 0$ it follows that the number of eigenvalues of X_n not belonging to an ϵ-neighborhood of the range of $a(x)(2 - 2\cos(s))$ is bounded by a constant, possibly depending on ϵ, but independent of n (see [45, Theorem 3.5]).

It is worth noticing that all these derivations go through, with minor modifications, also for higher order differential operators, in higher dimension, and by varying the boundary conditions: as an example, the previous analysis remains the same if the Neumann-Dirichlet boundary conditions are replaced by Dirichlet boundary conditions. As it can be easily argued, the only significant exception can be found when considering somehow artificial singularly perturbed problems in which the perturbation parameter is of the order of h (or of some power of h): in that case the analysis becomes more involved, since the arising matrix structures become significantly non-normal and different tools have to be taken into consideration.

3.2 Approximating class of sequences for matrix-functions

In Chapter 2 we have shown that the *a.c.s.* notion is stable under inversion, linear combinations and products, whenever natural and mild conditions are satisfied. In this section we focus our attention on the Hermitian case and we show that $\{\{f(B_{n,m})\} : m \geq 0\}$ is an *a.c.s.* for $\{f(A_n)\}$, if $\{\{B_{n,m}\} : m \geq 0\}$ is an *a.c.s.* for $\{A_n\}$, $\{A_n\}$ is sparsely unbounded, and f is a suitable continuous function defined on $\mathbb{R}$.

We recall that, if $A \in M_n(\mathbb{C})$ is a Hermitian matrix then, by Schur decomposition, it can be written as $A = UDU^*$ with U unitary and $D = \underset{j=0,\ldots,n-1}{\operatorname{diag}}(\lambda_j(A))$ diagonal and real; in this case, for every continuous function f, the matrix $f(A)$ is defined as $f(A) = Uf(D)U^*$ with $f(D) = \underset{j=0,\ldots,n-1}{\operatorname{diag}}(f(\lambda_j(A)))$.

Theorem 3.2. *Let $\{A_n\}$ be a sequence formed by Hermitian matrices, $A_n \in M_{d_n}(\mathbb{C})$ ($d_k < d_{k+1}$ for each k), and let $\{\{B_{n,m}\} : m \geq 0\}$, $m \in \hat{\mathbb{N}} \subset \mathbb{N}$, $\#\hat{\mathbb{N}} = \infty$, $B_{n,m} \in M_{d_n}(\mathbb{C})$ be one of its a.c.s. Suppose that K is a compact subset of $\mathbb{R}$ and that the sequence $\{A_n\}$ is weakly clustered at K, in the eigenvalue sense. Take any $\delta > 0$ and any function f continuous on $\overline{D(K,\delta)}$ (the closure of $D(K,\delta)$) and arbitrarily extended elsewhere. Then $\{\{f(B_{n,m})\} : m \geq 0\}$ is an a.c.s. for $\{f(A_n)\}$.*

Proof. From the definition of weak cluster (see Definition 1.8), for every $\epsilon > 0$, we know that

$$\lim_{n\to\infty} \frac{1}{d_n} \#\{i : \lambda_i(A_n) \notin D(K,\epsilon)\} = 0.$$

Take any positive r with $K \subset (-r, r)$. Then $2r$ is larger than the diameter of K and hence

$$\lim_{n\to\infty} \frac{1}{d_n} \#\{i : \lambda_i(A_n) \notin [-r, r]\} = 0.$$

The latter directly implies that

$$\lim_{r\to\infty} \limsup_{n\to\infty} \frac{1}{d_n} \#\{i : |\lambda_i(A_n)| \geq r\} = 0,$$

and then $\{A_n\}$ is sparsely unbounded according to (1.26) (note that, for Hermitian matrices, $|\lambda_i| = \sigma_i$).

Now take $\epsilon > 0$ and consider p_ϵ algebraic polynomial such that

$$\|f - p_\epsilon\|_{\infty,C} < \epsilon, \qquad C = \overline{D(K,\delta)},$$

where, by definition, $\|h\|_{\infty,S} := \sup_{x\in S} |h(x)|$ with h being a continuous function defined on the compact set S. Using a standard Schur decomposition, it is clear that

$$f(A_n) - p_\epsilon(A_n) = N_{n,\epsilon} + R_{n,\epsilon}, \qquad \|N_{n,\epsilon}\| < \epsilon, \qquad \operatorname{rank}(R_{n,\epsilon}) = o(d_n). \tag{3.10}$$

By the second part of Proposition 2.3, since $\{A_n\}$ is sparsely unbounded, it follows that for any positive integer j, fixed independently of n, $\left\{\left\{B^j_{n,m}\right\} : m \geq 0\right\}$ is an *a.c.s.* for $\{A^j_n\}$. Therefore again by Proposition 2.3, first two parts, we find that $\{\{p(B_{n,m})\} : m \geq 0\}$ is an *a.c.s.* for $\{p(A_n)\}$, for every polynomial p fixed independently of n. In particular, $\forall \epsilon > 0$, $\exists c_\epsilon(\cdot), \omega_\epsilon(\cdot), n_{\epsilon,}$. non-negative functions such that for every $m \geq 0$ we can write

$$p_\epsilon(A_n) = p_\epsilon(B_{n,m}) + R^\epsilon_{n,m} + N^\epsilon_{n,m}, \tag{3.11}$$

with

$$\begin{aligned} &\operatorname{rank}\left(R^\epsilon_{n,m}\right) \leq d_n c_\epsilon(m), \quad \left\|N^\epsilon_{n,m}\right\| \leq \omega_\epsilon(m), \qquad \forall n \geq n_{\epsilon,m}, \\ &\lim_{m\to\infty} \omega_\epsilon(m) = 0, \qquad \lim_{m\to\infty} c_\epsilon(m) = 0. \end{aligned}$$

We choose $\epsilon = g(m)$ such that $\lim_{m\to\infty} g(m) = 0$, $\lim_{m\to\infty} \omega(m) + c(m) = 0$ with $\omega(m) = \omega_{g(m)}(m)$, $c(m) = c_{g(m)}(m)$.

Therefore, by combining (3.10) and (3.11), we have

$$\begin{aligned} f(A_n) &= p_{g(m)}(A_n) + R_{n,g(m)} + N_{n,g(m)} \\ &= p_{g(m)}(B_{n,m}) + R_{n,g(m)} + N_{n,g(m)} + R^{g(m)}_{n,m} + N^{g(m)}_{n,m}. \end{aligned} \tag{3.12}$$

The above partial result allows one to write that $\left\{\left\{p_{g(m)}(B_{n,m})\right\} : m \geq 0\right\}$ is an *a.c.s.* for $\{f(A_n)\}$.

Therefore, for completing the proof, it only remains to study the behavior of the sequence $\left\{f(B_{n,m}) - p_{g(m)}(B_{n,m})\right\}$ for large m. For this end, let us order the eigenvalues of A_n and $B_{n,m}$ in a non-increasing order, i.e., $\lambda_1(X) \geq \lambda_2(X) \geq \cdots \geq \lambda_{d_n}(X)$ with X being either A_n or $B_{n,m}$. Since $\{\{B_{n,m}\} : m \geq 0\}$ is an *a.c.s.* for $\{A_n\}$, Definition 1.5 implies that there exist $\{\{R_{n,m}\} : m \geq 0\}$ and $\{\{N_{n,m}\} : m \geq 0\}$, with the conditions indicated in (1.16) and (1.17), such that $A_n = B_{n,m} + R_{n,m} + N_{n,m}$, that is $B_{n,m} = A_n - R_{n,m} - N_{n,m}$.

If we set

$$\widetilde{A_n} = A_n - N_{n,m},$$

then, by the MinMax Theorem (see, e.g., [15]), we find

$$\lambda_i(A_n) - \omega(m) \leq \lambda_i\left(\widetilde{A_n}\right) \leq \lambda_i(A_n) + \omega(m), \qquad \forall n \geq n_{g(m),m}. \tag{3.13}$$

Furthermore, writing

$$B_{n,m} = \widetilde{A_n} - R_{n,m},$$

again the MinMax Theorem leads to

$$\lambda_{i+2\lceil c(m)\rceil d_n}\left(\widetilde{A_n}\right) \leq \lambda_i(B_{n,m}) \leq \lambda_{i-2\lceil c(m)\rceil d_n}\left(\widetilde{A_n}\right), \qquad \forall n \geq n_{g(m),m}. \tag{3.14}$$

By combining formulae (3.13) and (3.14) we deduce

$$\lambda_{i+2\lceil c(m)\rceil d_n}(A_n) - \omega(m) \leq \lambda_i(B_{n,m}) \leq \lambda_{i-2\lceil c(m)\rceil d_n}(A_n) + \omega(m), \qquad \forall n \geq n_{g(m),m}, \tag{3.15}$$

and for $1 + 2\lceil c(m)\rceil d_n \leq i \leq d_n - 2\lceil c(m)\rceil d_n$, where $\omega(m)$ and $c(m)$ are the quantities indicated in Definition 1.5. Relation (3.15) is the classical interlacing property between the ordered spectrum of $B_{n,m}$ and that of A_n.

Let us consider the compact interval K_m defined as the closure of $D\left(\tilde{K}, \omega(m)\right)$ where $\tilde{K}$ is the convex hull of K. Since K is a (weak) cluster for the spectra of $\{A_n\}$ and given the interlacing between the spectra of A_n and $B_{n,m}$ (see (3.15)), the set K_m contains all the eigenvalues of $B_{n,m}$ except for at most $v(m) d_n$ of them with $v(m) \geq 0$ having zero limit as m tends to infinity. Therefore

$$\lim_{m\to\infty} \limsup_{n\to\infty} \frac{1}{d_n} \#\{i : \lambda_i(B_{n,m}) \notin K_m\} = 0.$$

Taking a positive r_m for which $K_m \subset (-r_m, r_m)$, we infer that $2r_m$ exceeds the diameter of K_m and hence

$$\lim_{r_m\to\infty} \limsup_{n\to\infty} \frac{1}{d_n} \#\{i : |\lambda_i(B_{n,m})| \geq r_m\} = 0.$$

As a consequence, for m large enough, $\{B_{n,m}\}$ behaves as a sparsely unbounded sequence in accordance with (1.26).

We observe that for any fixed $\delta > 0$, there exist $\overline{m}_\delta, \overline{\epsilon}_\delta > 0$ such that $\forall m > \overline{m}_\delta$ and $\forall \epsilon < \overline{\epsilon}_\delta$, it holds

$$\overline{D(K_m, \epsilon)} \subset \overline{D(K, \delta)}.$$

Therefore, from the choice of the polynomial $p_{g(m)}$, by exploiting the Schur decomposition, we have

$$f(B_{n,m}) - p_{g(m)}(B_{n,m}) = N_{n,m}(g(m)) + R_{n,m}(g(m)),$$

with $\|N_{n,m}(g(m))\| < g(m)$ and $\operatorname{rank}(R_{n,m}(g(m))) = o(d_n)$.

Putting together the latter expression with the formula (3.12), we plainly infer

$$\begin{aligned}
f(A_n) &= p_{g(m)}(A_n) + R_{n,g(m)} + N_{n,g(m)} \\
&= p_{g(m)}(B_{n,m}) + R_{n,g(m)} + N_{n,g(m)} + R^{g(m)}_{n,m} + N^{g(m)}_{n,m} \\
&= f(B_{n,m}) - N_{n,m}(g(m)) - R_{n,m}(g(m)) + R_{n,g(m)} + \\
&\quad + N_{n,g(m)} + R^{g(m)}_{n,m} + N^{g(m)}_{n,m} \\
&= f(B_{n,m}) + N_{n,g(m)} + R_{n,g(m)},
\end{aligned}$$

where

$$\begin{aligned}
N_{n,g(m)} &= N_{n,g(m)} + N^{g(m)}_{n,m} - N_{n,m}(g(m)), \\
R_{n,g(m)} &= R_{n,g(m)} + R^{g(m)}_{n,m} - R_{n,m}(g(m)),
\end{aligned}$$

$$\begin{aligned}
\left\|N_{n,g(m)}\right\| &\leq \left\|N_{n,g(m)}\right\| + \left\|N^{g(m)}_{n,m}\right\| + \|N_{n,m}(g(m))\| \\
&\leq 2g(m) + \omega_{g(m)}(m), \\
\operatorname{rank}\left(R_{n,g(m)}\right) &\leq \operatorname{rank}\left(R_{n,g(m)}\right) + \operatorname{rank}\left(R^{g(m)}_{n,m}\right) + \operatorname{rank}(R_{n,m}(g(m))) \\
&\leq o(d_n) + c_{g(m)}(m) d_n,
\end{aligned} \tag{3.16}$$

and with

$$\lim_{m\to\infty} 2g(m) + \omega_{g(m)}(m) = 0.$$

We now conclude by recalling that Definition 1.5 requires that all relations should hold for $n \geq n_m$, for a certain n_m. However from the definition of the Landau symbol $o(\cdot)$, for every $g(n) = o(d_n)$, there exists a value $\overline{n}_m$ for which

$$g(n) \leq \frac{d_n}{m}, \qquad \text{for } n \geq \overline{n}_m.$$

In conclusion (3.16) implies

$$\operatorname{rank}\left(R_{n,g(m)}\right) \leq d_n \left(\frac{1}{m} + c_{g(m)}(m)\right),$$

with

$$\lim_{m\to\infty} \frac{1}{m} + c_{g(m)}(m) = 0,$$

and the claimed thesis follows, i.e., $\{\{f(B_{n,m})\} : m \geq 0\}$ is an *a.c.s.* for $\{f(A_n)\}$ in accordance with Definition 1.5. □

Theorem 3.3. *Let $\{A_n\}$ be a sparsely unbounded matrix-sequence formed by Hermitian matrices, $A_n \in M_{d_n}(\mathbb{C})$ ($d_k < d_{k+1}$ for each k), and let $\{\{B_{n,m}\} : m \geq 0\}$, $m \in \hat{\mathbb{N}} \subset \mathbb{N}$, $\#\hat{\mathbb{N}} = \infty$, $B_{n,m} \in M_{d_n}(\mathbb{C})$ be one of its a.c.s. Take any function f continuous on $\mathbb{R}$. Then $\{\{f(B_{n,m})\} : m \geq 0\}$ is an a.c.s. for $\{f(A_n)\}$.*

Proof. The only difference with respect to the assumptions of the previous theorem concerns the fact that $\{A_n\}$ is not weakly clustered to a compact set, in the sense of Definition 1.8. For any $\epsilon > 0$, by definition of sparsely unbounded sequence, we have

$$\#\left\{i : |\lambda_i(A_n)| > \frac{1}{\epsilon}\right\} \leq \gamma(\epsilon)\, d_n, \qquad \text{with } \lim_{\epsilon\to 0} \gamma(\epsilon) = 0.$$

Therefore we choose p_ϵ standard polynomial approximating f with infinity norm error bounded by ϵ on the domain $\left[-\frac{1}{\epsilon}, \frac{1}{\epsilon}\right]$:

$$\|f - p_\epsilon\|_{\infty,\left[-\frac{1}{\epsilon},\frac{1}{\epsilon}\right]} < \epsilon,$$

in such a way (3.10) is replaced by

$$f(A_n) - p_\epsilon(A_n) = N_{n,\epsilon} + R_{n,\epsilon}, \quad \|N_{n,\epsilon}\| < \epsilon, \quad \operatorname{rank}(R_{n,\epsilon}) \leq \gamma(\epsilon)\, d_n.$$

From now we can work on a bounded compact interval (depending on ϵ) so that the proof is reduced to the one of Theorem 3.2. In fact, relations (3.11) and (3.12) are worked similarly as well as the expression of $\left\{f(B_{n,m}) - p_{g(m)}(B_{n,m})\right\}$. □

The result of Theorem 3.3 is of interest for applying concretely the Krylov convergence analysis proposed in [8, 59]. In this sense, the above analysis could be applied for proving that the class of Hermitian Generalized Locally Toeplitz (GLT) sequences [89] is closed under, e.g., square root or other noteworthy functions of interest in applications. We recall that the GLT class includes virtually any Finite Difference or Finite Element approximation of PDEs (see [111, 89, 90]) and therefore this result could have impact in stability issues in the Von Neumann/Lax sense, in providing spectral information for devising efficient multigrid solvers, or in providing spectral information on large preconditioned systems, when both the matrix and the preconditioner belong to the GLT class (see the series of applications discussed in [90]): as an example of the preconditioning issue, refer to the discussion on the diagonal-plus-structured preconditioning in [9, end of Section 3.2] for a concrete use of the results of this section.

3.3 Other versions of the Theorem 2.15

In this section we illustrate how we can achieve the same result of Theorem 2.15, weakening, strengthening and/or slightly modifying the hypotheses **(a1)**-**(a5)**.

In the first version we show that the hypotheses **(a3)** and (a slightly stronger form of) **(a4)** imply **(a1)**, **(a2)**, and **(a3)** for the set S defined by "filling in" the essential range of the function θ from **(a4)** (or its strengthened version). This will show that, when our set $\mathcal{S}(\theta)$ has empty interior our matrix-sequence has the desired distribution. When we say "filling in" we mean taking the "Area" in the following sense:

Definition 3.4. Let K be a compact subset of $\mathbb{C}$, then its complement has just one unbounded connected component. In other words, one has

$$\mathbb{C}\backslash K = U_0 \cup \bigcup_{j=1}^{\infty} U_j, \qquad U_i \cap U_j = \emptyset \text{ if } i \neq j,$$

where each U_j, $j \geq 1$, is a connected bounded open set, and U_0 is an unbounded connected open set (it may hold $U_j = \emptyset$ for some or even all $j \geq 1$). We define the *Area* of K as

$$Area\,(K) = \mathbb{C}\backslash U_0.$$

In other words, $Area\,(K)$ is the union of K and all the bounded components of its complement (intuitively, $Area\,(K)$ is the region of $\mathbb{C}$ that is delimited by K).

Theorem 3.5. *Let $\{A_n\}$ be a matrix-sequence, $A_n \in M_{d_n}(\mathbb{C})$ ($d_k < d_{k+1}$ for each k). If*

(b1) *the spectra $\Lambda(A_n)$ of A_n are uniformly bounded, i.e., $\exists C \in \mathbb{R}^+$ such that $|\lambda| < C$, $\lambda \in \Lambda(A_n)$, for all n;*

(b2) *there exists a function θ measurable, bounded, and defined on a set G of positive and finite Lebesgue measure, such that, for all non-negative integers L and l, we have*

$$\lim_{n\to\infty} \frac{\mathrm{tr}\left(\left(A_n^*\right)^l A_n^L\right)}{d_n} = \frac{1}{m\{G\}} \int_G \overline{\theta^l(t)}\theta^L(t)\,\mathrm{d}t;$$

then $\mathcal{S}(\theta)$ is compact, the matrix-sequence $\{A_n\}$ is weakly clustered at $Area\,(\mathcal{S}(\theta))$, and relation (1.20) is true for every continuous function F with bounded support which is holomorphic in the interior of $S = Area\,(\mathcal{S}(\theta))$.

If it is also true that $\mathbb{C}\backslash\mathcal{S}(\theta)$ is connected and the interior of $\mathcal{S}(\theta)$ is empty then the sequence $\{A_n\}$ is distributed as θ on its domain G, in the sense of the eigenvalues.

Proof. Since θ is bounded, $\mathcal{S}(\theta)$ is bounded, and so, since the essential range is always closed, the set $\mathcal{S}(\theta)$ is compact. Hence we can define $S = Area\,(\mathcal{S}(\theta))$.

We prove that S is a weak cluster for the spectra of $\{A_n\}$. First, we notice that the compact set $S_C = \{z \in \mathbb{C} : |z| \leq C\}$ is a strong cluster for the spectra of $\{A_n\}$ since by **(b1)** it contains all the eigenvalues. Moreover C can be chosen such that S_C contains S. Therefore, we will have proven that S is a weak cluster for $\{A_n\}$ if we prove that, for every $\varepsilon > 0$, the compact set $S_C\backslash D(S,\varepsilon)$ contains at most only $o(d_n)$ eigenvalues, with $D(S,\varepsilon)$ as in Definition 1.8. By compactness, for any $\delta > 0$, there exists a finite covering of $S_C\backslash D(S,\varepsilon)$ made of balls $D(z,\delta)$, $z \in S_C\backslash S$ with $D(z,\delta) \cap S = \emptyset$, and so, it suffices to show that, for a particular δ, at most $o(d_n)$ eigenvalues lie in $D(z,\delta)$. Let $F(t)$ be the characteristic function of the compact set $\overline{D(z,\delta)}$ (the closure of $D(z,\delta)$). Then restricting our attention to the compact set $\overline{D(z,\delta)} \cup S$, Mergelyan's theorem [76] implies that for each $\epsilon > 0$ there exists a polynomial P_ϵ such that

$|F(t) - P_\epsilon(t)|$ is bounded by ϵ on $\overline{D(z,\delta)} \cup S$. Therefore, setting $\gamma_n(z,\delta)$ equal to the number of eigenvalues of A_n belonging to $\overline{D(z,\delta)}$, we find

$$
\begin{aligned}
(1-\epsilon)\,\gamma_n(z,\delta) &\leq \sum_{i=1}^{d_n} F(\lambda_i)\,|P_\epsilon(\lambda_i)| && (3.17)\\
&\leq \left(\sum_{i=1}^{d_n} F^2(\lambda_i)\right)^{\frac{1}{2}} \left(\sum_{i=1}^{d_n} |P_\epsilon(\lambda_i)|^2\right)^{\frac{1}{2}} && (3.18)\\
&= \left(\sum_{i=1}^{d_n} F(\lambda_i)\right)^{\frac{1}{2}} \left(\sum_{i=1}^{d_n} |P_\epsilon(\lambda_i)|^2\right)^{\frac{1}{2}} && (3.19)\\
&= (\gamma_n(z,\delta))^{\frac{1}{2}} \left(\sum_{i=1}^{d_n} |P_\epsilon(\lambda_i)|^2\right)^{\frac{1}{2}} && (3.20)\\
&\leq (\gamma_n(z,\delta))^{\frac{1}{2}} \|P_\epsilon(A_n)\|_2 && (3.21)\\
&= (\gamma_n(z,\delta))^{\frac{1}{2}} \left(\mathrm{tr}\left(P_\epsilon^*(A_n)\,P_\epsilon(A_n)\right)\right)^{\frac{1}{2}} && (3.22)\\
&= (\gamma_n(z,\delta))^{\frac{1}{2}} \left(\mathrm{tr}\left(\sum_{l,L=0}^{K} \overline{c_l} c_L\,(A_n^*)^l\,A_n^L\right)\right)^{\frac{1}{2}} && (3.23)\\
&= (\gamma_n(z,\delta))^{\frac{1}{2}} \left(\sum_{l,L=0}^{K} \overline{c_l} c_L \mathrm{tr}\left((A_n^*)^l\,A_n^L\right)\right)^{\frac{1}{2}}, && (3.24)
\end{aligned}
$$

where inequality (3.17) follows from the definition of F and from the approximation properties of P_ϵ, inequality (3.18) is Cauchy-Schwarz, relations (3.19)–(3.20) come from the definitions of F and $\gamma_n(z,\delta)$, (3.21) is a consequence of the Schur decomposition and of the unitary invariance of the Schatten p-norms for each p, identities (3.22)–(3.24) follow from the entry-wise definition of the Schatten 2-norm (the Frobenius norm), from the monomial expansion of the polynomial P_ϵ, and from the linearity of the trace.

Given $\epsilon_2 > 0$, we choose $\epsilon_1 > 0$ so that equation

$$
\epsilon_1 \sum_{l,L=0}^{K} |c_l|\,|c_L| \leq \epsilon_2,
$$

is true and then, using **(b2)**, we choose N so that for $n > N$, the equation

$$
\left| \frac{\mathrm{tr}\left((A_n^*)^l\,A_n^L\right)}{d_n} - \frac{1}{m\{G\}} \int_G \overline{\theta^l(t)} \theta^L(t)\,\mathrm{d}t \right| < \epsilon_1,
$$

is true. Then, picking up from equation (3.24), we have that, for $n > N$

$$
\begin{aligned}
(1-\epsilon)\,\gamma_n(z,\delta) &\leq (\gamma_n(z,\delta))^{\frac{1}{2}} \left(d_n \left(\epsilon_2 + \frac{1}{m\{G\}} \int_G \sum_{l,L=0}^{K} \overline{c_l} c_L \left(\overline{\theta^l(t)} \theta^L(t)\right)\mathrm{d}t\right)\right)^{\frac{1}{2}} && (3.25)\\
&= (\gamma_n(z,\delta))^{\frac{1}{2}} \left(d_n \left(\epsilon_2 + \frac{1}{m\{G\}} \int_G |P_\epsilon(\theta(t))|^2\,\mathrm{d}t\right)\right)^{\frac{1}{2}} && (3.26)\\
&\leq (\gamma_n(z,\delta))^{\frac{1}{2}}\, d_n^{\frac{1}{2}} \left(\epsilon^2 + \epsilon_2\right)^{\frac{1}{2}}, && (3.27)
\end{aligned}
$$

where inequality (3.25) is assumption **(b2)**, the latter two inequalities are again consequences of the monomial expansion of P_ϵ and of the approximation properties of P_ϵ over the area delimited

by the range of θ, and ϵ_2 is arbitrarily small. So, choosing $\epsilon_2 = \epsilon^2$, we see that (3.17)–(3.27) imply that, for n sufficiently large,

$$\gamma_n(z,\delta) \leq 2d_n\epsilon^2(1-\epsilon)^{-2},$$

which means that: $\gamma_n(z,\delta) = o(d_n)$.

Thus, hypotheses **(a1)**–**(a5)** of Theorem 2.15 hold with $S = Area(\mathcal{S}(\theta))$, which is necessarily compact and with connected complement, and consequently the first conclusion of Theorem 2.15 holds. Finally if $\mathbb{C}\backslash\mathcal{S}(\theta)$ is connected and the interior of $\mathcal{S}(\theta)$ is empty then $Area(\mathcal{S}(\theta)) = \mathcal{S}(\theta)$ and so all the hypotheses of Theorem 2.15 are satisfied. We can now conclude that the sequence $\{A_n\}$ is distributed in the sense of the eigenvalues as θ on its domain G. □

Now, we give a second version, replacing hypotheses **(a1)**–**(a5)** with only **(a3)**, **(a4)**, and a condition on the Schatten p-norm for a certain p.

Theorem 3.6. *Let $\{A_n\}$ be a matrix-sequence, $A_n \in M_{d_n}(\mathbb{C})$ ($d_k < d_{k+1}$ for each k). If*

(c1) *the spectra $\Lambda(A_n)$ of A_n are uniformly bounded, i.e., $|\lambda| < C$, $\lambda \in \Lambda(A_n)$, for all n;*

(c2) *there exists a function θ measurable, bounded, and defined over G having positive and finite Lebesgue measure, such that, for every non-negative integer L, we have*

$$\lim_{n\to\infty} \frac{\operatorname{tr}\left(A_n^L\right)}{d_n} = \frac{1}{m\{G\}} \int_G \theta^L(t)\,\mathrm{d}t;$$

(c3) *there exists a positive real number $p \in [1,\infty)$, independent of n, such that, for every polynomial P there exists $N \in \mathbb{N}$ such that, for $n > N$,*

$$\|P(A_n)\|_p^p \leq 2d_n \frac{1}{m\{G\}} \int_G |P(\theta(t))|^p\,\mathrm{d}t;$$

then the matrix-sequence $\{A_n\}$ is weakly clustered at $Area(\mathcal{S}(\theta)) := \mathbb{C}\backslash U$ (see Definition 3.4) and relation (1.20) is true for every continuous function F with bounded support which is holomorphic in the interior of $S = Area(\mathcal{S}(\theta))$. If, moreover

(c4) *$\mathbb{C}\backslash\mathcal{S}(\theta)$ is connected and the interior of $\mathcal{S}(\theta)$ is empty;*

then the sequence $\{A_n\}$ is distributed as θ on its domain G, in the sense of the eigenvalues.

Proof. The proof goes as in Theorem 3.5 until relation (3.17). Then with q the conjugate of p (i.e., $\frac{1}{q} + \frac{1}{p} = 1$) we have

$$\begin{aligned}
(1-\epsilon)\gamma_n(z,\delta) &\leq \left(\sum_{i=1}^{d_n} F^q(\lambda_i)\right)^{\frac{1}{q}} \left(\sum_{i=1}^{d_n} |P_\epsilon(\lambda_i)|^p\right)^{\frac{1}{p}} && (3.28)\\
&= \left(\sum_{i=1}^{d_n} F(\lambda_i)\right)^{\frac{1}{q}} \left(\sum_{i=1}^{d_n} |P_\epsilon(\lambda_i)|^p\right)^{\frac{1}{p}} && (3.29)\\
&= (\gamma_n(z,\delta))^{\frac{1}{q}} \left(\sum_{i=1}^{d_n} |P_\epsilon(\lambda_i)|^p\right)^{\frac{1}{p}} && (3.30)\\
&\leq (\gamma_n(z,\delta))^{\frac{1}{q}} \|P_\epsilon(A_n)\|_p && (3.31)\\
&\leq (\gamma_n(z,\delta))^{\frac{1}{q}} \left(\frac{2d_n}{m\{G\}} \int_G |P_\epsilon(\theta(t))|^p\,\mathrm{d}t\right)^{\frac{1}{p}} && (3.32)\\
&\leq (\gamma_n(z,\delta))^{\frac{1}{q}} (2d_n)^{\frac{1}{p}}\,\epsilon, && (3.33)
\end{aligned}$$

where relation (3.28) is the Hölder's inequality (see (1.15)), relations (3.29)–(3.30) come from the definitions of F and $\gamma_n(z,\delta)$, (3.31) comes from the fact that, for any square matrix, the vector with the moduli of the eigenvalues is weakly majorized by the vector of the singular values (see [15] for the precise definition and for the result), inequality (3.32) is assumption **(c3)** (which holds for any polynomial of fixed degree), and finally inequality (3.33) follows from the approximation properties of P_ϵ over the area delimited by the range of θ. Therefore

$$\gamma_n(z,\delta) \leq 2d_n\epsilon^p(1-\epsilon)^{-p},$$

and since ϵ is arbitrary we have the desired result, i.e., $\gamma_n(z,\delta) = o(d_n)$.

The rest of the proof is the same as in Theorem 3.5. □

The next result tells us that the key assumption **(c3)** follows from the distribution in the singular value sense of $\{P(A_n)\}$ and that the latter is equivalent to the very same limit relation with only polynomial test functions.

Theorem 3.7. *If the sequence $\{A_n\}$, $A_n \in M_{d_n}(\mathbb{C})$ ($d_k < d_{k+1}$ for each k), is uniformly bounded in spectral norm then $\{A_n\} \sim_\sigma (\theta, G)$ is true whenever condition (1.21) holds for all polynomial test functions. Moreover, if $\{P(A_n)\} \sim_\sigma (P(\theta), G)$ for every polynomial P then claim **(c3)** is true for every value $p \in [1,\infty)$.*

Proof. The first claim is proved by using the fact that one can approximate any continuous function defined on a compact set contained in the (positive) real line by polynomials. The second claim follows from taking as test function the function z^p, with positive p, and exploiting the limit relation from the assumption $\{P(A_n)\} \sim_\sigma (P(\theta), G)$. Indeed, the sequence $\{P(A_n)\}$ is uniformly bounded since $\{A_n\}$ is, so we are allowed to use as test functions continuous functions with no restriction on the support. Therefore, by definition (see (1.21)), $\{P(A_n)\} \sim_\sigma (P(\theta), G)$ implies that

$$\lim_{n\to\infty} \frac{1}{d_n} \sum_{j=1}^{d_n} \sigma_j^p(P(A_n)) = \frac{1}{m\{G\}} \int_G |P(\theta(t))|^p \, \mathrm{d}t.$$

Hence, by observing that $\sum_{j=1}^{d_n} \sigma_j^p(P(A_n))$ is by definition $\|P(A_n)\|_p^p$ and using the definition of limit, we see that, for every $\epsilon > 0$, there exists an integer $\bar{n}_\epsilon$ such that

$$\|P(A_n)\|_p^p \leq d_n \frac{1+\epsilon}{m\{G\}} \int_G |P(\theta(t))|^p \, \mathrm{d}t, \qquad \forall n \geq \bar{n}_\epsilon,$$

and since, without loss of generality, we can assume that $\epsilon < 1$, we get **(c3)**. □

The two previous theorems will be used in Chapter 4 to find the spectral distribution of sequences of products of Toeplitz matrices.

Chapter 4

Sequences of Toeplitz matrices

In this chapter we study the asymptotic spectral behavior of a product of Toeplitz sequences (in the usual, matrix-valued, and multi-level cases), by using and extending tools from matrix theory and finite-dimensional linear algebra.

The notion of distribution in the eigen/singular value sense goes back to Weyl and has been investigated by many authors in the Toeplitz and locally Toeplitz context (see the book by Böttcher and Silbermann [20] where many classical results by the authors, Szegö, Avram, Parter, Widom, Tyrtyshnikov, and many other can be found, and more recent results in [22, 21, 45, 60, 103, 118, 109, 112]).

It is well-known that the product of Toeplitz operators is rarely equal to a Toeplitz operator (see [23, 62]), but, it turns out that the sequence of eigenvalues or singular values of the product of two Toeplitz sequences is often related to the product of the two symbols in a Szegö-type way. For the singular values the result is known as long as all the involved symbols are essentially bounded and, in fact, for any linear combination of products of Toeplitz operators, the distribution function is exactly the linear combination of the products of the symbols of the sequences: the latter goes back to the work of Roch and Silbermann (see [20, Sections 4.6 and 5.7]). The previous results have been extended by considering integrable symbols, not necessarily bounded [89, 90], and by considering (pseudo) inversion and the related algebra of sequences (see [90, 94]). Of course for the eigenvalues much less is known, and one simple reason is that much less is true, as another basic example discussed at the beginning of Section 2 in [98] shows. However, quite recently, using the Ky Fan-Mirsky theorem which says that the real (or imaginary) parts of the eigenvalues are majorized by the eigenvalues of the real (or imaginary) part of the matrix (see [15]), Golinskii and Serra-Capizzano found a method for deducing the eigenvalue distribution of sequences obtained as generic perturbations of Hermitian sequences, when the trace norm of the perturbation is asymptotically negligible with respect to size of the involved matrices (see Theorem 2.17, Theorem 2.18 and [45]). We recall that a real vector v of size n is said to be majorized by a real vector w of the same size if, for each k, the sum of the largest k entries of v is bounded by the sum of the k largest entries of w and equality holds for $k = n$.

By using [45, Lemma 3.2], Golinskii and Serra-Capizzano proved that the eigenvalues of a non-Hermitian complex perturbation of a Jacobi matrix-sequence, which are not necessarily real, are still distributed as the real-valued function $2\cos t$ on $[0, \pi]$, which characterizes the non-perturbed case where the Jacobi sequence is of course real and symmetric: see [45], and [52, 90] for a further application of Lemma 3.2 in [45] to a (pseudo) differential setting. In this chapter, we apply these results to certain products of Toeplitz sequences, then discuss, apply and extend more general tools introduced by Tilli [112], and based on the Mergelyan's theorem, see [76]. Furthermore, the case of Laurent polynomials not necessarily in the Tilli class is sketched and a generalization to the case of multi-level Toeplitz sequences and sequences $T_n(f)$ where f is a matrix-valued function is also given: we have to emphasize that these multi-level and matrix-valued extensions are of interest in the Engineering context where the number

of levels refers to multiple inputs (Multi-Input systems) and size of the basic blocks, i.e., the size of the matrix-valued symbol refers to multiple outputs (Multi-Output systems). Following the Engineering terminology, we are talking of SIMO and MIMO systems, see, e.g., [44, 49] for details and the references therein.

4.1 Toeplitz sequences: definition and previous distribution results

We begin this section by introducing the definition of multi-level Toeplitz matrix.

Definition 4.1. *Let f be a Lebesgue integrable function defined over Q^d, where $Q = (-\pi, \pi)$, and taking values in $M_{p,q}(\mathbb{C})$, for given positive integers p and q. Then, for d-indices $r = (r_1, \ldots, r_d), j = (j_1, \ldots, j_d), n = (n_1, \ldots, n_d)$, $e = (1, \ldots, 1)$, $\underline{0} = (0, \ldots, 0)$, the multi-level Toeplitz matrix $T_n(f) \in M_{p\hat{n},q\hat{n}}(\mathbb{C})$, $\hat{n} = n_1 n_2 \cdots n_d$, is defined as follows (see* [109]*):*

$$T_n(f) = \sum_{j_1=-n_1+1}^{n_1-1} \cdots \sum_{j_d=-n_d+1}^{n_d-1} J_{n_1}^{(j_1)} \otimes \cdots \otimes J_{n_d}^{(j_d)} \otimes \tilde{f}_{(j_1,\ldots,j_d)}, \tag{4.1}$$

where $J_m^{(\ell)} \in M_m(\mathbb{R}), (-m+1 \le \ell \le m-1)$ is the matrix whose (i,j)th entry is 1 if $i - j = \ell$ and 0 otherwise (thus $\{J_{-m+1}, \ldots, J_{m-1}\}$ is the natural basis for the space of $m \times m$ Toeplitz matrices), $\otimes$ denotes the tensor or Kronecker product of matrices and $\tilde{f}_k$ are the Fourier coefficients of f defined by

$$\tilde{f}_k = \tilde{f}_{(k_1,\ldots,k_d)} = \frac{1}{(2\pi)^d} \int_{Q^d} f(t_1, \ldots, t_d)\, e^{-\mathrm{i}(k_1 t_1 + \cdots + k_d t_d)} \mathrm{d}t_1 \cdots \mathrm{d}t_d, \qquad \left(\mathrm{i}^2 = -1\right), \tag{4.2}$$

for integers k_ℓ such that $-\infty < k_\ell < \infty$ for $1 \le \ell \le d$. Since f is a matrix-valued function of d variables whose component functions are all integrable, then the $(k_1, \ldots, k_d)$th Fourier coefficient is considered to be the matrix whose (u,v)th entry is the $(k_1, \ldots, k_d)$th Fourier coefficient of the function $(f(t_1, \ldots, t_d))_{u,v}$. In the usual multi-level indexing language, we can rewrite (4.1) more compactly as

$$T_n(f) = \left[\tilde{f}_{r-j}\right]_{r,j=\underline{0}}^{n-e},$$

where f is called symbol *or* generating function *of the Toeplitz matrix $T_n(f)$.*

In the following we write $n \to \infty$ to indicate that $\min_{r=1,\ldots,d} n_r \to \infty$.

Throughout this chapter we speak of *Toeplitz sequences* as matrix-sequences of the form $\{A_n\}$ with $A_n = T_n(f)$ and $T_n(f)$ defined as above.

For the sake of clarity, whenever the extension from scalar to matrix-valued functions is simple enough, we shall prove our theorems (especially the ones concerning multi-level Toeplitz) only in the case $p = q = 1$ (see [109] for a detailed matrix definition and [20] for an explanation with examples in the case $d = 2$), then $T_n(f) \in M_{\hat{n}}(\mathbb{C})$.

The asymptotic distribution of eigenvalues and singular values of a sequence of Toeplitz matrices has been deeply studied in the last century, and strictly depends on the generating function f (see, for example, [20, 109, 118] and the references therein). Now, let $\{f_{\alpha,\beta}\}$ be a finite set of $L^1\left(Q^d\right)$ functions and define the measurable function h by:

$$h = \sum_{\alpha=1}^{\rho} \prod_{\beta=1}^{v_\alpha} f_{\alpha,\beta}^{s(\alpha,\beta)}, \qquad s(\alpha,\beta) \in \{\pm 1\}, \tag{4.3}$$

where $f_{\alpha,\beta}$ is sparsely vanishing (see Definition 1.11) when $s(\alpha,\beta) = -1$. The function h may not belong to L^1 in which case $\{T_n(h)\}$ is not defined according to the rule in (4.1) simply

because the Fourier coefficients (4.2) are not well-defined. However, for the sequence of matrices $\left\{\sum_{\alpha=1}^{\rho}\prod_{\beta=1}^{v_\alpha} T_n^{s(\alpha,\beta)}(f_{\alpha,\beta})\right\}$, the following result holds.

Proposition 4.2. [89, 90, 117, 83] *With the above assumptions we have that*

$$\left\{\sum_{\alpha=1}^{\rho}\prod_{\beta=1}^{v_\alpha} T_n^{s(\alpha,\beta)}(f_{\alpha,\beta})\right\} \sim_\sigma \left(h, Q^d\right),$$

and

$$\left\{\sum_{\alpha=1}^{\rho}\prod_{\beta=1}^{v_\alpha} T_n^{s(\alpha,\beta)}(f_{\alpha,\beta})\right\} \sim_\lambda \left(h, Q^d\right),$$

if the matrices $\sum_{\alpha=1}^{\rho}\prod_{\beta=1}^{v_\alpha} T_n^{s(\alpha,\beta)}(f_{\alpha,\beta})$ *are Hermitian, at least for* n *large enough (which implies necessarily that* $p = q$*). In this context, the symbol* $T_n^{s(\alpha,\beta)}(f_{\alpha,\beta})$ *with* $s(\alpha,\beta) = -1$ *and* $f_{\alpha,\beta}$ *sparsely vanishing means that we are (pseudo) inverting the matrix in the sense of Moore-Penrose (see* [15]*), since* $T_n(f_{\alpha,\beta})$ *is not necessarily invertible, but the number of zero singular values is at most* $o(\hat{n})$*, for* $n \to \infty$*.*

We should mention here that the distribution results in the singular value sense are much easier to obtain and to prove [89, 90, 111, 109, 118], thanks to the higher stability of singular values under perturbations [121].

Notice that in defining the symbol h when matrix-valued symbols are involved, it is necessary to consider compatible dimensions and also one has to be careful in respecting the correct ordering in the products, owing to the lack of commutativity in the matrix context.

When $\rho = 1$, $p = q = 1$, and $v_1 = 1$ this result concerns standard Toeplitz sequences and is attributed to Tyrtyshnikov, Zamarashkin and Tilli [109, 116, 118]; see also [83] and the references therein for the evolution of the subject. The case where $s(\alpha,\beta) = 1$ for every α and β is considered and solved in [83, 117] by using matrix theory techniques. We stress that the Hermitian case where h is defined as in (4.3) has been treated in two different ways in [20, 90], for both singular values and eigenvalues. In the following section we first assume only that the symbol of the product (h) is real-valued, then in Section 4.3 we extend these results to functions with "thin spectrum" or that belong to the Tilli class.

Definition 4.3. *Let* D *be any domain equipped with a positive measure and let us consider the space* $L^\infty(D)$ *of complex-valued essentially bounded functions. The* Tilli class *is the subset of* $L^\infty(D)$ *made by functions* f *whose (essential) range* $\mathcal{S}(f)$ *has empty interior and does not disconnect the complex plane.*

It is clear that the condition defining the Tilli class does not involve any regularity of the function, but it is more related to the topology/geometry of the range (see also [20, Example 5.39] and [120, top of p. 390]); by the way it is evident that the Tilli class includes properly all the real-valued L^∞ functions.

Remark 4.4. *It should be noted that, according to* [90]*, the distribution result for singular values holds for any sequence belonging to the algebra generated by Toeplitz sequences with* $L^1\left(Q^d\right)$ *symbols, where the allowed algebraic operations are linear combination, product, and (pseudo) inversion. In order to formally define this algebra* $\mathcal{A}_T$ *we say that* $\mathcal{A}_T = \bigcup_{j=0}^{\infty} \mathcal{A}_T^{(j)}$ *where Toeplitz sequences with* $L^1\left(Q^d\right)$ *symbols form the set* $\mathcal{A}_T^{(0)}$ *and* $\{A_n\} \in \mathcal{A}_T^{(j)}$*,* $j \geq 1$*, if*

there exists a finite set of sequences $\left\{A_n^{(\alpha,\beta)}\right\}$*, with measurable symbols* $f_{\alpha,\beta}$*, belonging to* $\mathcal{A}_T^{(k)}$*,* $0 \leq k < j$*, such that*

$$A_n = \sum_{\alpha=1}^{\rho} \prod_{\beta=1}^{v_\alpha} \left(A_n^{(\alpha,\beta)}\right)^{s(\alpha,\beta)}, \qquad s(\alpha,\beta) \in \{\pm 1\},$$

where every sequence which is (pseudo) inverted $(s(\alpha,\beta) = -1)$ *should have sparsely vanishing symbol; the new symbol of* $\{A_n\}$ *is recursively defined as*

$$h = \sum_{\alpha=1}^{\rho} \prod_{\beta=1}^{v_\alpha} f_{\alpha,\beta}^{s(\alpha,\beta)}, \qquad s(\alpha,\beta) \in \{\pm 1\}.$$

The general result in [90] *is that* $\{A_n\} \sim_\sigma \left(h, Q^d\right)$ *and* $\{A_n\} \sim_\lambda \left(h, Q^d\right)$*, if all the matrices* A_n *are Hermitian, at least for* n *large enough.*

Finally it is worth mentioning that the above results also hold when starting from the set of block multi-level sequences generated by matrix-valued $p \times q$ *symbols; see* [90] *for general integrable symbols (i.e. all the singular values of the symbol are integrable on* Q^d*) and* [21] *for the case of bounded symbols with* $p = q$ *and without pseudo inversion, but where the distribution result for eigenvalues is extended to the case in which the involved sequences are normal (the Hermitian case for general integrable symbols and with pseudo inversion can be found in* [90]*).*

4.2 Preliminary results for sequences of Toeplitz matrices

In this section we present some simple technical results which are useful in our next study in search of the spectral distribution of sequences of products of Toeplitz matrices.

The first lemma is due to SeLegue: it can explicitly be found in [20, Lemma 5.16]. We present an elementary matrix proof as an alternative to the (elementary) operator theory proof given in [20] . This proof seems to be the most natural one to extend to the multi-level case, as explained in the proof of Lemma 4.6.

Lemma 4.5. *Let* $f, g \in L^\infty(Q)$*,* $A_n = T_n(f)\,T_n(g)$*, and let* $h = fg$*. Then* $\|A_n - T_n(h)\|_1 = o(n)$*.*

Proof. In order to estimate $\|A_n - T_n(h)\|_1$, i.e., the Schatten 1-norm of $A_n - T_n(h)$, we will use some classical results from approximation theory.

For a given $\theta \in L^1(Q)$, let $p_{k,\theta}$ be its Cesaro sum of degree k, i.e., the arithmetic average of Fourier sums of order q with $q \leq k$ (see [16, 123]). From standard trigonometric series theory we know that $p_{k,\theta}$ converges in L^1 norm to θ as k tends to infinity and also that $\|p_{k,\theta}\|_{L^\infty} \leq \|\theta\|_{L^\infty}$, whenever $\theta \in L^\infty(Q)$ with $L^\infty(Q) \subset L^1(Q)$. Furthermore, the norm inequality $\|T_n(\theta)\|_p \leq \left(\frac{n}{2\pi}\right)^{\frac{1}{p}} \|\theta\|_{L^p}$ holds for every $\theta \in L^p(Q)$ if $1 \leq p \leq \infty$ (see [4] and [96, Corollary 4.2]). Now, by adding and subtracting and by using the triangle inequality several times we get:

$$\begin{aligned} \|A_n - T_n(h)\|_1 \;\leq\; & \|A_n - T_n(p_{k,f})\,T_n(g)\|_1 + \|T_n(p_{k,f})\,T_n(g) - T_n(p_{k,f})\,T_n(p_{k,g})\|_1 + \\ & + \|T_n(p_{k,f})\,T_n(p_{k,g}) - T_n(p_{k,f}p_{k,g})\|_1 + \|T_n(p_{k,f}p_{k,g}) - T_n(h)\|_1, \end{aligned} \tag{4.4}$$

and, by using Hölder's inequality for the Schatten p-norms (see (1.13)) and the previously

mentioned norm inequality from [96], we infer that

$$\begin{aligned}
\left\|A_n - T_n\left(p_{k,f}\right) T_n\left(g\right)\right\|_1 &= \left\|\left(T_n\left(f\right) - T_n\left(p_{k,f}\right)\right) T_n\left(g\right)\right\|_1 \\
&\leq \left\|T_n\left(f\right) - T_n\left(p_{k,f}\right)\right\|_1 \left\|T_n\left(g\right)\right\| \\
&\leq \frac{n}{2\pi} \left\|f - p_{k,f}\right\|_{L^1} \left\|g\right\|_{L^\infty}; \\
\left\|T_n\left(p_{k,f}\right) T_n\left(g\right) - T_n\left(p_{k,f}\right) T_n\left(p_{k,g}\right)\right\|_1 &= \left\|T_n\left(p_{k,f}\right)\left(T_n\left(g\right) - T_n\left(p_{k,g}\right)\right)\right\|_1 \\
&\leq \left\|T_n\left(g\right) - T_n\left(p_{k,g}\right)\right\|_1 \left\|T_n\left(p_{k,f}\right)\right\| \\
&\leq \left\|T_n\left(g - p_{k,g}\right)\right\|_1 \left\|p_{k,f}\right\|_{L^\infty} \\
&\leq \frac{n}{2\pi} \left\|g - p_{k,g}\right\|_{L^1} \left\|f\right\|_{L^\infty}; \\
\left\|T_n\left(p_{k,f} p_{k,g}\right) - T_n\left(h\right)\right\|_1 &= \left\|T_n\left(h - p_{k,f} p_{k,g}\right)\right\|_1 \\
&\leq \frac{n}{2\pi} \left\|h - p_{k,f} p_{k,g}\right\|_{L^1}.
\end{aligned}$$

Thus, we see that the sum of the first, second and fourth terms of (4.4) equals $\epsilon\left(k\right) n$ where, since the Cesaro operator converges to the identity in the L^1 topology, we have

$$\lim_{k\to\infty} \epsilon\left(k\right) = 0.$$

We treat the third term of (4.4) in a different way. Let us recall that $p_{k,f}$ and $p_{k,g}$ are trigonometric polynomials of degree at most k, so that

$$p_{k,f}\left(t\right) = \sum_{j=-k}^{k} a_j e^{\mathrm{i}jt}, \qquad p_{k,g}\left(t\right) = \sum_{j=-k}^{k} b_j e^{\mathrm{i}jt},$$

and their product is

$$p_{k,f}\left(t\right) p_{k,g}\left(t\right) = \sum_{j=-2k}^{2k} \gamma_j e^{\mathrm{i}jt}, \qquad \text{with} \quad \gamma_j = \sum_{l+L=j} a_l b_L.$$

Now from the definition of Toeplitz matrix generated by a symbol, we find

$$\left(T_n\left(p_{k,f} p_{k,g}\right)\right)_{r,s} = \gamma_{r-s} = \sum_{l+L=r-s} a_l b_L = \sum_{l=-k}^{k} a_l b_{r-s-l}, \tag{4.5}$$

and

$$\begin{aligned}
\left(T_n\left(p_{k,f}\right) T_n\left(p_{k,g}\right)\right)_{r,s} &= \sum_{v=1}^{n} \left(T_n\left(p_{k,f}\right)\right)_{r,v} \left(T_n\left(p_{k,g}\right)\right)_{v,s} \\
&= \sum_{v=1}^{n} a_{r-v} b_{v-s} \\
&= \sum_{l=-r}^{n-r} a_l b_{r-s-l}.
\end{aligned} \tag{4.6}$$

The summations in (4.5) and (4.6) coincide when $k \leq r \leq n-k$ (remember that $a_l = 0$ if $l > k$ or $l < -k$). Since r is the row index, the latter remark implies that the two matrices $T_n\left(p_{k,f} p_{k,g}\right)$ and $T_n\left(p_{k,f}\right) T_n\left(p_{k,g}\right)$ differ only on the first and on the last $k-1$ rows so that

$$\operatorname{rank}\left(T_n\left(p_{k,f} p_{k,g}\right) - T_n\left(p_{k,f}\right) T_n\left(p_{k,g}\right)\right) \leq 2\left(k-1\right) < 2k.$$

Now, since the trace norm is bounded by the the rank times the spectral or operator norm, we see that:

$$\begin{aligned}
\left\|T_n\left(p_{k,f}\right)T_n\left(p_{k,g}\right)-T_n\left(p_{k,f}p_{k,g}\right)\right\|_1 &\leq 2k\left\|T_n\left(p_{k,f}\right)T_n\left(p_{k,g}\right)-T_n\left(p_{k,f}p_{k,g}\right)\right\| \\
&\leq 2k\left(\left\|T_n\left(p_{k,f}\right)\right\|\left\|T_n\left(p_{k,g}\right)\right\|+\left\|T_n\left(p_{k,f}p_{k,g}\right)\right\|\right) \\
&\leq 2k\left(\left\|p_{k,f}\right\|_{L^\infty}\left\|p_{k,g}\right\|_{L^\infty}+\left\|p_{k,f}p_{k,g}\right\|_{L^\infty}\right) \\
&\leq 2k\left(\left\|p_{k,f}\right\|_{L^\infty}\left\|p_{k,g}\right\|_{L^\infty}+\left\|p_{k,f}\right\|_{L^\infty}\left\|p_{k,g}\right\|_{L^\infty}\right) \\
&= 4k\left\|p_{k,f}\right\|_{L^\infty}\left\|p_{k,g}\right\|_{L^\infty} \\
&\leq 4k\left\|f\right\|_{L^\infty}\left\|g\right\|_{L^\infty},
\end{aligned}$$

for each $k\in\mathbb{N}$. Thus, if $G=4\left\|g\right\|_{L^\infty}\left\|f\right\|_{L^\infty}$ and $\epsilon\left(k\right)$ is defined above, then

$$\left\|A_n-T_n\left(h\right)\right\|_1\leq\epsilon\left(k\right)n+kG, \tag{4.7}$$

for each $k\in\mathbb{N}$. Now, for each $\epsilon>0$, by first choosing k_0 so that $\epsilon\left(k_0\right)<\frac{\epsilon}{2}$ then choosing $\tilde{N}>\frac{2Gk_0}{\epsilon}$, we see that $n\geq\tilde{N}$ gives

$$\frac{\left\|A_n-T_n\left(h\right)\right\|_1}{n}\leq\epsilon,$$

which finishes the proof. □

Next, we notice that the reasoning above applies to multi-level Toeplitz matrices, where $T_n\left(f\right)\in M_{\hat{n}}\left(\mathbb{C}\right)$ represent the multi-level Toeplitz matrix with symbol f (as in Definition 4.1).

Lemma 4.6. *Let $f,g\in L^\infty\left(Q^d\right)$, $n=\left(n_1,\ldots,n_d\right)\in\mathbb{N}^d$ and $\hat{n}=n_1n_2\cdots n_d$. Then for $A_n=T_n\left(f\right)T_n\left(g\right)$ and $h=fg$ we have:*

$$\left\|A_n-T_n\left(h\right)\right\|_1=o\left(\hat{n}\right).$$

The only part of the proof which is slightly different from that of Lemma 4.5 is the treatment of the third term of (4.4). To get the analogous inequality, we remember that all the involved symbols are trigonometric polynomials of degree not exceeding k, a direct check shows that the two matrices $T_n\left(p_{k,f}\right)T_n\left(p_{k,g}\right)$ and $T_n\left(p_{k,f}p_{k,g}\right)$ can differ only on the first k_1 block-rows and on the last k_1 block-rows of size $\frac{\hat{n}}{n_1}$; moreover on every block of size $\frac{\hat{n}}{n_1}$ the two matrices can differ only the first k_2 block-rows and on the last k_2 block-rows of size $\frac{\hat{n}}{n_1n_2}$ and so on. Therefore, setting $\left\|k\right\|_\infty=\max\limits_{j=1,\ldots,d}k_j$, the trace norm of $T_n\left(p_{k,f}\right)T_n\left(p_{k,g}\right)-T_n\left(p_{k,f}p_{k,g}\right)$ is bounded by the rank times the spectral norm, i.e.,

$$\left\|T_n\left(p_{k,f}\right)T_n\left(p_{k,g}\right)-T_n\left(p_{k,f}p_{k,g}\right)\right\|_1\leq 4\left\|k\right\|_\infty\left\|g\right\|_{L^\infty}\left\|f\right\|_{L^\infty}\frac{\hat{n}}{\min\limits_{j=1,\ldots,d}n_j},$$

we can replace equation (4.7) with the equation:

$$\left\|A_n-T_n\left(h\right)\right\|_1\leq\epsilon\left(k\right)\hat{n}+\gamma\left(k\right)G,$$

where $\gamma\left(k\right)=\left\|k\right\|_\infty\frac{\hat{n}}{\min\limits_{j=1,\ldots,d}n_j}$, for each $k\in\mathbb{N}^d$, and choose, for $\epsilon>0$, a d-tuple k such that $\epsilon\left(k\right)<\frac{\epsilon}{2}$ and an $\tilde{N}$ such that $\tilde{N}>\frac{2G\gamma(k)}{\epsilon}$. Then, if $\hat{n}>\tilde{N}$ we will have

$$\frac{\left\|A_n-T_n\left(h\right)\right\|_1}{\hat{n}}\leq\epsilon,$$

which shows that $\left\|A_n-T_n\left(h\right)\right\|_1=o\left(\hat{n}\right)$ and finishes the proof.

Now we consider the results of distribution in the case of a sequence $\{A_n\}$ where $A_n = T_n(f)\,T_n(g)\,;\, f,g \in L^\infty(Q)$ such that fg is real-valued (even though f and g are not necessarily real-valued; for the simpler, all real-valued case, see [90] or [117]). Symbols of this type are studied, e.g., in statistics (see [13, 12]). The idea is to look at A_n as the Hermitian matrix $T_n(h)$, $h = fg$, plus a correction term C_n such that $\|C_n\|_1 = o(n)$ as $n \to \infty$, where each of the matrix-sequences is uniformly bounded in operator norm (see Lemma 4.5). This will permit us to use the powerful Theorem 2.17.

Theorem 4.7. *Let $f,g \in L^\infty(Q)$ be such that $h = fg$ is real-valued. Then $\{A_n\} \sim_\lambda (h,Q)$ with $A_n = T_n(f)\,T_n(g)$, $\mathcal{S}(h)$ is a weak cluster for $\{A_n\}$, and any $s \in \mathcal{S}(h)$ strongly attracts the spectra of $\{A_n\}$ with infinite order.*

Proof. It is well-known (see [48]) that $\{T_n(h)\} \sim_\lambda (h,Q)$ and $\|T_n(\theta)\| \le \|\theta\|_{L^\infty}$ for every $\theta \in L^\infty(Q)$. Thus $\|T_n(h)\| \le \|h\|_{L^\infty}$ and $\|A_n\| \le \|T_n(f)\|\,\|T_n(g)\| \le \|f\|_{L^\infty}\|g\|_{L^\infty}$. As a consequence, since $\|A_n - T_n(h)\|_1 = o(n)$ by Lemma 4.5, the desired results follow by applying Theorem 2.17 with $B_n = T_n(h)$, $C_n = A_n - T_n(h)$. □

Now we once again notice that the same theorem holds for multi-level Toeplitz matrices.

Theorem 4.8. *Let $d \in \mathbb{N}^+$ and let $f,g \in L^\infty\left(Q^d\right)$ be such that $h = fg$ is real-valued. Then, if $A_n = T_n(f)\,T_n(g)$, we have that $\{A_n\} \sim_\lambda \left(h,Q^d\right)$, $\mathcal{S}(h)$ is a weak cluster for $\{A_n\}$, and any $s \in \mathcal{S}(h)$ strongly attracts the spectra of $\{A_n\}$ with infinite order.*

Proof. In 1993, Tyrtyshnikov showed that the relation (1) holds for multi-level Toeplitz sequences (see [20, Theorem 6.41]) so that we once again have $\{T_n(h)\} \sim_\lambda \left(h,Q^d\right)$. Also, by the definition of the Toeplitz operators it is again true that

$$\|T_n(h)\| \le \|h\|_{L^\infty},$$

(see [48]) and $\|A_n\| \le \|T_n(f)\|\,\|T_n(g)\| \le \|f\|_{L^\infty}\|g\|_{L^\infty}$. As a consequence, since we have that $\|A_n - T_n(h)\|_1 = o(\hat{n})$ by Lemma 4.6, the desired results follow by applying Theorem 2.17 with $B_n = T_n(h)$ and $C_n = A_n - T_n(h)$. □

Remark 4.9. *Let f,g,h and A_n be defined as in Theorem 4.7 and suppose that either f or g is a Laurent polynomial (see (4.17)) of degree k. Then, if $h = fg$, by the same type of reasoning as above, $A_n - T_n(h)$ has rank less than or equal to $2k$. Therefore, again using the fact that the sequences $\{\|A_n\|\}$ and $\{\|T_n(h)\|\}$ are both bounded by $\|f\|_{L^\infty}\|g\|_{L^\infty}$ and the Schur decomposition, it follows that $\|A_n - T_n(h)\|_1 \le 4k\,\|f\|_{L^\infty}\|g\|_{L^\infty}$. As a consequence, since $\mathcal{S}(h)$ is a compact real set, Theorem 2.18 implies that $\mathcal{S}(h)$ is a strong cluster for the spectra of $\{A_n\}$.*

Remark 4.10. *Lemma 4.5 and Theorem 4.7 remain valid in a block multi-dimensional setting, i.e., when considering symbols belonging to $L_q^\infty\left(Q^d\right)$ with $d \ge 2$, $q \ge 2$. In fact, we can follow verbatim the same proof as in Lemma 4.5 (see also* [21]*) and in Theorem 4.7 since all the tools concerning the Cesaro operator and the trace norm estimates have a natural counterpart in several dimensions and in the matrix-valued setting (see* [96, 123]*). The only change is of notational type: in fact all the terms $o(n)$ will become $o(\hat{n})$, since the involved dimensions in the multi-dimensional Toeplitz setting are $q\hat{n}$, with $\hat{n} = n_1 n_2 \cdots n_d$ and with $n = (n_1,\ldots,n_d)$ being a multi-index, see Section 4.1.*

In light of the previous remark, it is natural to state the following generalization without proof.

Theorem 4.11. *Let $f,g \in L_q^\infty\left(Q^d\right)$ be such that $h = fg$ is Hermitian-valued (real-valued for $q = 1$). Then $\{A_n\} \sim_\lambda \left(h,Q^d\right)$ with $A_n = T_n(f)\,T_n(g)$, $\mathcal{S}(h)$ is a weak cluster for $\{A_n\}$, and any $s \in \mathcal{S}(h)$ strongly attracts the spectra of $\{A_n\}$ with infinite order.*

Theorem 4.11 is the basis for the subsequent general result concerning the algebra generated by Toeplitz sequences with $L_q^\infty\left(Q^d\right)$ symbols. Its proof works by induction on the structure of h and of A_n and, more specifically, Theorem 4.11 is used for the basis of induction and for the inductive step. We do not furnish further details since, under mild additional assumptions, the same statement is proved carefully in Section 4.3 in the more general case where h belongs to the Tilli class. We recall that Hermitian-valued (real-valued if $q = 1$) $L_q^\infty\left(Q^d\right)$ functions form a proper subset of the Tilli class.

Theorem 4.12. *Let $f_{\alpha,\beta} \in L_q^\infty\left(Q^d\right)$ with $\alpha = 1, \ldots, \rho$, $\beta = 1, \ldots, v_\alpha$, $\rho, v_\alpha < \infty$. Assume that the function $\sum_{\alpha=1}^{\rho} \prod_{\beta=1}^{v_\alpha} f_{\alpha,\beta}$, is Hermitian-valued (real-valued for $q = 1$) and consider the sequence $\{A_n\}$ with $A_n = \sum_{\alpha=1}^{\rho} \prod_{\beta=1}^{v_\alpha} T_n\left(f_{\alpha,\beta}\right)$. Then $\{A_n\} \sim_\lambda \left(h, Q^d\right)$, $\mathcal{S}(h)$ is a weak cluster for $\{A_n\}$, and any $s \in \mathcal{S}(h)$ strongly attracts the spectra of $\{A_n\}$ with infinite order.*

Remark 4.13. *Theorem 4.12 nicely complements the analysis by Böttcher and coauthors in* [21]*. In fact in* [21] *the authors require that the given sequence $\{A_n\}$ is normal, i.e., every A_n satisfies $A_n^* A_n = A_n A_n^*$. This technical assumption may be difficult to verify except in the Hermitian case. For the Hermitian setting see also* [90] *and Remark 4.4.*

We are now ready to state and prove two important lemmas. An alternative proof using operator theory methods can be found in [21].

Lemma 4.14. *Let $f_\alpha \in L^\infty\left(Q^d\right)$ with $\alpha = 1, \ldots, \rho$, $\rho < \infty$, $d \geq 1$, and let $n = (n_1, \ldots, n_d)$ and $\hat{n} = n_1 n_2 \cdots n_d$. Set*

$$A_n = \prod_{\alpha=1}^{\rho} T_n\left(f_\alpha\right),$$

and $h = \prod_{\alpha=1}^{\rho} f_\alpha$. Then

$$\left\|A_n - T_n(h)\right\|_1 = o(\hat{n}), \tag{4.8}$$

$$\lim_{n\to\infty} \frac{\operatorname{tr}(A_n)}{\hat{n}} = \frac{1}{(2\pi)^d} \int_{Q^d} h(t_1, \ldots, t_d)\, \mathrm{d}t_1 \cdots \mathrm{d}t_d. \tag{4.9}$$

Proof. For proving (4.8) we proceed by induction on the positive integer ρ. If $\rho = 1$ then there is nothing to prove since $A_n - T_n(h)$ is the null matrix. For $\rho > 1$, we write $A_n = \left(\prod_{\alpha=1}^{\rho-1} T_n\left(f_\alpha\right)\right) T_n\left(f_\rho\right)$, where, by the inductive step, we have $\prod_{\alpha=1}^{\rho-1} T_n\left(f_\alpha\right) = T_n\left(h_{\rho-1}\right) + E_{n,\rho-1}$ with $h_{\rho-1} = \prod_{\alpha=1}^{\rho-1} f_\alpha$ and $\left\|E_{n,\rho-1}\right\|_1 = o(\hat{n})$. As a consequence

$$A_n = T_n\left(h_{\rho-1}\right) T_n\left(f_\rho\right) + E_{n,\rho-1} T_n\left(f_\rho\right),$$

where

$$\left\|E_{n,\rho-1} T_n\left(f_\rho\right)\right\|_1 \leq \left\|E_{n,\rho-1}\right\|_1 \left\|T_n\left(f_\rho\right)\right\| \leq \left\|E_{n,\rho-1}\right\|_1 \left\|f_\rho\right\|_{L^\infty},$$

by the Hölder's inequality (see (1.13)) and by the inequality $\left\|T_n(g)\right\| \leq \|g\|_{L^\infty}$, see, e.g., [20]. Furthermore, thanks to Lemma 4.6, we have

$$\left\|T_n\left(h_{\rho-1}\right) T_n\left(f_\rho\right) - T_n(h)\right\|_1 = o(\hat{n}),$$

since $h = h_{\rho-1} f_\rho$. In conclusion $A_n = T_n(h) + E_{n,\rho}$ where

$$E_{n,\rho} = E_{n,\rho-1} T_n(f_\rho) + T_n(h_{\rho-1}) T_n(f_\rho) - T_n(h)$$

so that by the triangle inequality $\|E_{n,\rho}\|_1 = o(\hat{n})$, and therefore the proof of the first part is concluded.

The proof of the second part, i.e., relation (4.9) is plain since the statement is a straightforward consequence of the first part. In fact

$$\operatorname{tr}(T_n(h)) = \hat{n}\tilde{h}_0 = \frac{\hat{n}}{(2\pi)^d} \int_{Q^d} h(t_1, \ldots, t_d)\, \mathrm{d}t_1 \cdots \mathrm{d}t_d,$$

where $\tilde{h}_0$ is the Fourier coefficient defined in (4.2), and, by (1.14) and (4.8),

$$\operatorname{tr}(A_n) = \operatorname{tr}(T_n(h)) + o(\hat{n}) = \frac{\hat{n}}{(2\pi)^d} \int_{Q^d} h(t_1, \ldots, t_d)\, \mathrm{d}t_1 \cdots \mathrm{d}t_d + o(\hat{n}),$$

which implies (4.9). □

Lemma 4.15. *Let $f_{\alpha,\beta} \in L^\infty\left(Q^d\right)$ with $\alpha = 1, \ldots, \rho$, $\beta = 1, \ldots, v_\alpha$, $\rho, v_\alpha < \infty$, $d \geq 1$, and let $n = (n_1, \ldots, n_d)$ and $\hat{n} = n_1 n_2 \cdots n_d$. Set*

$$A_n = \sum_{\alpha=1}^{\rho} \prod_{\beta=1}^{v_\alpha} T_n(f_{\alpha,\beta}),$$

and $h = \sum_{\alpha=1}^{\rho} \prod_{\beta=1}^{v_\alpha} f_{\alpha,\beta}$. Then

$$\|A_n - T_n(h)\|_1 = o(\hat{n}),$$

$$\lim_{n\to\infty} \frac{\operatorname{tr}(A_n)}{\hat{n}} = \frac{1}{(2\pi)^d} \int_{Q^d} h(t_1, \ldots, t_d)\, \mathrm{d}t_1 \cdots \mathrm{d}t_d. \tag{4.10}$$

Proof. The first claim is a trivial consequence of Lemma 4.14. For the second claim, just observe that the linearity of the trace operator and of the limit operation implies that (4.10) is equivalent to the statement that

$$\sum_{\alpha=1}^{\rho} \lim_{n\to\infty} \frac{1}{\hat{n}} \operatorname{tr}\left(\prod_{\beta=1}^{v_\alpha} T_n(f_{\alpha,\beta})\right) = \sum_{\alpha=1}^{\rho} \frac{1}{(2\pi)^d} \int_{Q^d} \prod_{\beta=1}^{v_\alpha} f_{\alpha,\beta}(t_1, \ldots, t_d)\, \mathrm{d}t_1 \cdots \mathrm{d}t_d.$$

Hence, setting $g_\alpha = \prod_{\beta=1}^{v_\alpha} f_{\alpha,\beta}$, $\alpha = 1, \ldots, \rho$, the desired result follows from

$$\lim_{n\to\infty} \frac{1}{\hat{n}} \operatorname{tr}\left(\prod_{\beta=1}^{v_\alpha} T_n(f_{\alpha,\beta})\right) = \frac{1}{(2\pi)^d} \int_{Q^d} g_\alpha(t_1, \ldots, t_d)\, \mathrm{d}t_1 \cdots \mathrm{d}t_d,$$

which is a consequence of Lemma 4.14. □

4.3 The Tilli class and the algebra generated by Toeplitz sequences

Many mathematicians have worked to obtain generalizations of the Szegö theorem to functions with "thin spectrum", a concept which varies a bit from one author to another. Here we work with the definition used by Tilli given in Definition 4.3.

In the paper [112] Tilli was able to show that the distribution in the sense of the eigenvalues of the Toeplitz sequence $\{T_n(f)\}$ is valid whenever the symbol f lies in the Tilli class. Indeed the proof is given in one dimension ($d = 1$) but the extension in several dimensions is plain.

Now we want to extend the result of Tilli to the case of linear combinations of products of sequences of Toeplitz matrices whose "cumulative" symbol (derived from the same linear combination of the products of generating functions/symbols of the Toeplitz matrices involved) belongs to the Tilli class; to this aim we work along two different ways depending on the tools used: the first theorem (Theorem 4.16) is obtained by extending Theorem 4.8 from the subset of real-valued symbols to the whole Tilli class (the proof is not given in detail, but only by providing the main steps), the second (Theorem 4.18), however, is obtained by application of Theorem 3.6 and powerful Lemmas 4.14 and 4.15.

Theorem 4.16. *Let $f, g \in L^\infty\left(Q^d\right)$ be such that $h = fg$ belongs to the Tilli class, $d \geq 1$. Assume that a function ϕ can be found continuous on $\mathcal{S}(h)$, the range of h, such that it is injective and the range of $\phi(h)$ lies on the real line, i.e., $\mathcal{S}(\phi(h))$ is compact set of $\mathbb{R}$. Then $\{A_n\} \sim_\lambda \left(h, Q^d\right)$ with $A_n = T_n(f)\, T_n(g)$.*

Proof. We consider five steps:

Step1. Given $\epsilon > 0$ consider ϕ_ϵ polynomial such that $\|\phi - \phi_\epsilon\|_{L^\infty, \mathcal{S}(h)} < \epsilon$, ϕ_ϵ is injective, the latter implying that its range does not disconnect the complex plane. In such a way the range of $\phi_\epsilon(h)$ lies in a ϵ-neighborhood of a compact subset of the real line.

Step2. Therefore, $T_n(\phi(h))$ is Hermitian since $\phi(h)$ is real-valued and has real eigenvalues contained in the interval $[r, R]$ with r being the essential infimum of $\phi(h)$ and R being the essential supremum of $\phi(h)$. Moreover, since $\|T_n(\phi_\epsilon(h)) - T_n(\phi(h))\| = \|T_n(\phi_\epsilon(h) - \phi(h))\| \leq \|\phi(h) - \phi_\epsilon(h)\|_{L^\infty} = \|\phi - \phi_\epsilon\|_{L^\infty, \mathcal{S}(h)} < \epsilon$, it follows that the matrix $T_n(\phi_\epsilon(h))$ has all the eigenvalues in a ϵ-neighborhood of $[r, R]$.

Step3. Now for every polynomial P of fixed degree $\|P(T_n(f)\, T_n(g)) - P(T_n(h))\|_1 = o(\hat{n})$ and $\|P(T_n(h)) - T_n(P(h))\|_1 = o(\hat{n})$; this is not difficult in view of Theorem 4.6.

Step4. Using the previous step with any polynomial $P = \phi_\epsilon$ and since the eigenvalues of $T_n(\phi(h))$ belong to $[r, R]$, we deduce that the eigenvalues of the sequences $\{\phi_\epsilon(T_n(h))\}$ and $\{\phi_\epsilon(T_n(f)\, T_n(g))\}$ are clustered in a ϵ-neighborhood of $[r, R]$. As a consequence, the injectivity of ϕ_ϵ implies that the eigenvalues of $\{T_n(f)\, T_n(g)\}$ are clustered in a ϵ'-neighborhood of the range of $h = fg$. Since ϵ and therefore ϵ' can be chosen arbitrarily, it follows that the sequence $\{T_n(f)\, T_n(g)\}$ is clustered in the eigenvalue sense at $\mathcal{S}(h)$.

Step5. By trivial computation it is easily deduced that for every positive integer k

$$\operatorname{tr}\left(\left(T_n(f)\, T_n(g)\right)^k - \left(T_n(h)\right)^k\right) = o(\hat{n}).$$

Since $\{T_n(h)\} \sim_\lambda \left(h, Q^d\right)$ by the previous relation, for every positive integer k, it follows that

$$\lim_{n\to\infty} \frac{\operatorname{tr}\left(T_n(f)\, T_n(g)\right)^k}{\hat{n}} = \frac{1}{(2\pi)^d} \int_{Q^d} h(t)^k \, \mathrm{d}t.$$

The above limit relation, the clustering of $\{T_n(f)\,T_n(g)\}$ at $\mathcal{S}(h)$, the uniform boundedness of the spectra of $\{T_n(f)\,T_n(g)\}$, and the fact that h does not disconnect the complex plane and its range has empty interior are the assumptions of Theorem 2.15: the conclusion of Theorem 2.15 is exactly the desired claim and hence the proof is concluded.

□

The proof would have worked without technical assumptions if the following claim would have been true.

Claim. Given h bounded such that its range has empty interior and does not disconnect the complex plane, find ϕ continuous on $\mathcal{S}(h)$, the range of h, such that it is injective and the range of $\phi(h)$ lies on the real line, i.e., $\mathcal{S}(\phi(h))$ is compact set of $\mathbb{R}$.

Unfortunately this claim is generally false as the subsequent example shows.

Proposition 4.17. *Let $K \subseteq \mathbb{C}$ be the Y shaped compact set illustrated in Fig. 4.1, and let $\phi : K \to \mathbb{R}$. Then ϕ continuous implies that ϕ is not injective and viceversa, i.e., ϕ injective implies that ϕ is not continuous.*

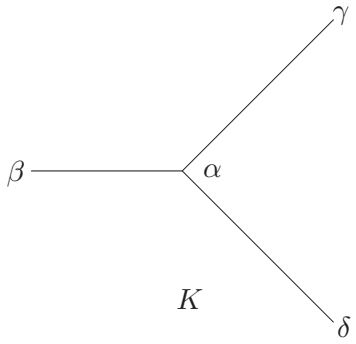

Figure 4.1: Y shaped compact set K.

Proof. Let $\phi : K \to \mathbb{R}$ be continuous. Let us consider the three edges C_1, C_2, C_3 and the point α, $\alpha \notin C_r$, $r = 1, 2, 3$, as illustrated in Fig. 4.2, where $K = C_1 \cup C_2 \cup C_3 \cup \{\alpha\}$. The continuity of ϕ implies the following relationships

$$\lim_{\substack{z \in C_1 \\ z \to \alpha}} \phi(z) = \phi(\alpha), \tag{4.11}$$

$$\lim_{\substack{z \in C_2 \\ z \to \alpha}} \phi(z) = \phi(\alpha), \tag{4.12}$$

$$\lim_{\substack{z \in C_3 \\ z \to \alpha}} \phi(z) = \phi(\alpha). \tag{4.13}$$

Assume now that ϕ is injective on K, that is $\forall z_1, z_2 \in K$, $z_1 \neq z_2$, we find $\phi(z_1) \neq \phi(z_2)$. Let us set $\phi(\alpha) = a$.

The following reasonings can be made.

Claim 1: Let us remark that, given any of the three sets C_r with $r = 1, 2, 3$, the function $\phi - a$ cannot change sign since the existence of two points $z_1, z_2 \in C_r$ with $\phi(z_1) > a$ and $\phi(z_2) < a$, would imply by continuity the existence of $\widetilde{z} \in C_r$, $\widetilde{z} \neq \alpha$ ($\alpha \notin C_r$) such that $\phi(\widetilde{z}) = \phi(\alpha) = a$. This would imply that ϕ is not injective.

Claim 2: Given C_r and C_s, $r, s = 1, 2, 3$, $r \neq s$, the following condition is impossible:

$$\forall z \in C_r,\ \phi(z) > a \quad \text{and} \quad \forall t \in C_s,\ \phi(t) > a. \tag{4.14}$$

Indeed, suppose by contradiction that (4.14) is satisfied. Since relations (4.11), (4.12), and (4.13) hold, for every $\epsilon > 0$ there exist $\widetilde{C}_r \subseteq C_r$ and $\widehat{C}_s \subseteq C_s$, with $\widetilde{C}_r, \widehat{C}_s \neq \emptyset$, $\widetilde{C}_r \cap \widehat{C}_s = \emptyset$, such that

for $z \in \widetilde{C}_r$, the range of $\phi(z)$ contains $(a, a+\epsilon)$;
for $t \in \widehat{C}_s$, the range of $\phi(t)$ contains $(a, a+\epsilon)$.

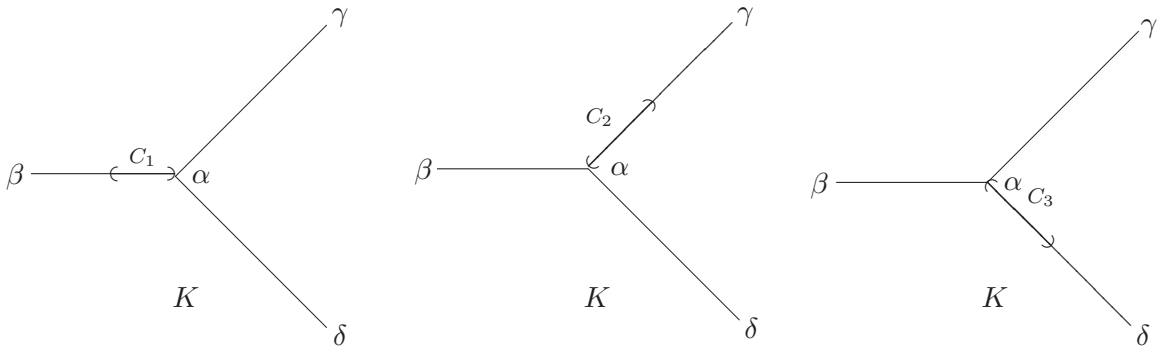

Figure 4.2: Analysis of ϕ in the 3 branches of K.

Furthermore, since ϕ is a continuous function, by varying z in the set $\widetilde{C}_r$, $\phi(z)$ has to take all the values $(a, a+\epsilon)$ and by varying t in the set $\widehat{C}_s$, $\phi(t)$ has to take all the values $(a, a+\epsilon)$. From this it follows that there exist $\widetilde{z} \in \widetilde{C}_r$ and $\widehat{t} \in \widehat{C}_s$ with $\widetilde{z} \neq \widehat{t}$ such that $\phi(\widetilde{z}) = \phi\left(\widehat{t}\right)$, and the latter would imply again that ϕ is not injective.

Claim 3: Given C_r and C_s, $r, s = 1, 2, 3$, $r \neq s$, the following condition is impossible:

$$\forall z \in C_r,\ \phi(z) < a \quad \text{and} \quad \forall t \in C_s,\ \phi(t) < a.$$

In fact, it is enough to repeat verbatim the same reasoning as in the previous claim, using the interval $(a-\epsilon, a)$ in place of $(a, a+\epsilon)$.

Let us consider for instance the subset C_1. Given relation (4.11) and Claim 1, we have two possibilities:

I) $\forall z \in C_1$, $\phi(z) > a$;

II) $\forall z \in C_1$, $\phi(z) < a$.

Let us suppose that case *I)* holds: of course the alternative case *II)* can be handled similarly. Therefore, we have

$$\forall z \in C_1, \quad \phi(z) > a. \tag{4.15}$$

Let us consider the subset C_2. With the same arguments used for C_1, from (4.12) and (4.15), Claim 2 implies that

$$\forall z \in C_2, \quad \phi(z) < a. \tag{4.16}$$

Now let us consider the subset C_3. From (4.13), given the continuity of ϕ, we deduce that

1. by Claim 1, $\phi - a$ cannot change sign in any of the subsets C_r, $r = 1, 2, 3$;
2. by Claim 2 and (4.15) $\forall z \in C_3$, $\phi(z) \not> a$;

3. by Claim 3 and (4.16) $\forall z \in C_3$, $\phi(z) \not< a$;

4. by injectivity $\forall z \in C_3$, $\phi(z) \neq a$.

The four listed claims are of course in contradiction. Therefore ϕ cannot be simultaneously continuous and injective over all K and the proof is complete. □

However, even in this case where the key assumption of Theorem 4.16 is not satisfied, the Szegö formula can be recovered in full generality.

Indeed it is enough to repeat the same proof as in Theorem 4.16, with ϕ being continuous and injective over $C_1 \cup C_2$ and with $\phi(\alpha) = \phi(z)\ \forall z \in C_3$. In that case we obtain a partial Szegö relation in which the test function is an arbitrary continuous function over $C_1 \cup C_2$, but it is constant over the remaining branch. However if we follow the same steps now choosing ϕ continuous and injective over $C_1 \cup C_3$ with $\phi(\alpha) = \phi(z)\ \forall z \in C_2$, then we obtain a new partial Szegö relation in which the test function is an arbitrary continuous function over $C_1 \cup C_3$, but it is constant over the branch C_2. If we sum up these two partial relations, then the general Szegö formula is derived for this specific setting, in which the essential range of the symbol h is Y shaped and compact.

In conclusion, despite the negative answer provided by Proposition 4.17 for satisfying the key assumption of Theorem 4.16, the arguments used for the case where the range of h is contained in a Y shaped compact tells us that the Szegö relation can be extended as long as we have a finite number of branches.

We conclude the first part of this section with the "second version" of Theorem 4.16, which overcomes the problems related to the continuity and injectivity of the function ϕ on a Y shaped compact domain. The proof of the following result is obtained using a more powerful tool: the Theorem 3.6

Theorem 4.18. *Let $f_{\alpha,\beta} \in L^\infty\left(Q^d\right)$ with $\alpha = 1, \dots, \rho$, $\beta = 1, \dots, v_\alpha$, $\rho, v_\alpha < \infty$, $d \geq 1$. Assume that the function*

$$h = \sum_{\alpha=1}^{\rho} \prod_{\beta=1}^{v_\alpha} f_{\alpha,\beta},$$

belongs to the Tilli class and consider the sequence $\{A_n\}$ with $A_n = \sum_{\alpha=1}^{\rho} \prod_{\beta=1}^{v_\alpha} T_n(f_{\alpha,\beta})$. Then $\{A_n\} \sim_\lambda \left(h, Q^d\right)$, $\mathcal{S}(h)$ is a weak cluster for $\{A_n\}$, and any $s \in \mathcal{S}(h)$ strongly attracts the spectra of $\{A_n\}$ with infinite order.

Proof. We choose to apply Theorem 3.6. Assumption **(c1)** is easily obtained by repeated applications of the triangle inequality to the infinity norm of the function h since the module of the eigenvalues is dominated by the infinity norm of the symbol. Statement **(c3)** is true for every p by Theorem 3.7, since $\{P(A_n)\} \sim_\sigma \left(P(h), Q^d\right)$ for every fixed polynomial P (see Remark 4.4); assumption **(c4)** is verified with $\theta = h$ since h belongs to the Tilli class. The only thing left is statement **(c2)** which is a consequence of Lemma 4.15, since any positive power of linear combinations of products is still a linear combination of products. Therefore $\{A_n\} \sim_\lambda \left(h, Q^d\right)$ by Theorem 3.6 and the proof is completed by invoking **a)** and **b)** from Theorem 2.16. □

4.3.1 The Tilli class in the case of matrix-valued symbols

With the same tools we can easily give the generalization of Theorem 4.18 to the case of $q \times q$ matrix-valued symbols. Lemmas 4.14 and 4.15 are easy to extend and indeed this extension can be found in [21]. The only key point is to define the Tilli class in this context. We say that f belongs to the $q \times q$ matrix-valued Tilli class if f is essentially bounded (i.e. this is true

for any entry of f) and if the union of the ranges of the eigenvalues of f has empty interior and does not disconnect the complex plane. We have to observe that the case where $f(t)$ is diagonalizable, by a constant transformation independent of t, is special in the sense that the Szegö-type distribution result holds under the milder assumption that every eigenvalue of f (now a scalar complex-valued function) belongs to the standard Tilli class. This leaves open the question whether this weaker requirement is sufficient in general.

Finally we remark that such results can be seen as a generalization of the analysis by Böttcher and coauthors in [21], with the advantage that the technical and difficult assumption of normality is dropped.

4.3.2 The role of thin spectrum in the case of Laurent polynomials

In this subsection we treat a problem suggested by Böttcher, the case where the symbol f of our Toeplitz operator is a Laurent polynomial, i.e.,

$$f(z) = \sum_{j=-r}^{s} \tilde{f}_j z^j, \qquad z \in \mathbb{T}. \tag{4.17}$$

Given a Laurent polynomial f and given a value $\rho > 0$, we denote by $f^{[\rho]}$ the function

$$f^{[\rho]}(z) = \sum_{j=-r}^{s} \tilde{f}_j \rho^j z^j. \tag{4.18}$$

Clearly $f^{[\rho]}$ is still a Laurent polynomial and, if we define the matrix $D_\rho = \underset{j=0,\ldots,n-1}{\operatorname{diag}} (\rho^j) \in M_n(\mathbb{R})$ then a straightforward computation shows that

$$D_\rho T_n(f) D_\rho^{-1} = T_n\left(f^{[\rho]}\right). \tag{4.19}$$

Now, if f is any Laurent polynomial, then, as shown in the book [18] the eigenvalues of the sequence $\{T_n(f)\}$ cluster along a certain set called the Schmidt-Spitzer set, and denoted by $\Lambda(f)$. It was shown by Hirschmann ([18, Theorems 11.16 and 11.17]) that, under certain hypotheses,

$$\{T_n(f)\} \sim_\lambda (\theta_f, G_f), \tag{4.20}$$

where θ_f is a suitable function supported on $G_f = \bigcap_{\rho>0} Area\left(\mathcal{S}\left(f^{[\rho]}\right)\right)$, $f^{[\rho]}$ is as in (4.18), and the concept of *Area* is defined as in Definition 3.4.

Suppose now the functions $f_{\alpha,\beta}$, $\alpha = 1,\ldots,\nu$, $\beta = 1,\ldots,v_\alpha$, $\nu, v_\alpha < \infty$, are all Laurent polynomials, then the function h defined by

$$h = \sum_{\alpha=1}^{\nu} \prod_{\beta=1}^{v_\alpha} f_{\alpha,\beta},$$

is also a Laurent polynomial. We want to prove that if h satisfies the hypotheses of the Hirschmann theorem so that $\{T_n(h)\} \sim_\lambda (\theta_h, G_h)$, then we can obtain the corresponding result for the sequence $\{A_n\} = \left\{ \sum_{\alpha=1}^{\nu} \prod_{\beta=1}^{v_\alpha} T_n(f_{\alpha,\beta}) \right\}$, i.e., $\{A_n\} \sim_\lambda (\theta_h, G_h)$.

In order to do this, we need to prove the results of Section 4.2 when f, g are Laurent polynomials.

Theorem 4.19. *Let f, g be two Laurent polynomials, $A_n = T_n(f) T_n(g)$ and let $h = fg$. If we put $D_\rho = \underset{j=0,\ldots,n-1}{\operatorname{diag}} (\rho^j)$, for each $\rho > 0$, then $\left\| D_\rho A_n D_\rho^{-1} - D_\rho T_n(h) D_\rho^{-1} \right\|_1 = o(n)$.*

Proof. This is a direct consequence of Lemma 4.5 applied to the functions $f^{[\rho]}$ and $g^{[\rho]}$ since (using (4.19)) we have

$$D_\rho A_n D_\rho^{-1} = T_n\left(f^{[\rho]}\right) T_n\left(g^{[\rho]}\right), \quad \text{and} \quad D_\rho T_n(h) D_\rho^{-1} = T_n\left(h^{[\rho]}\right),$$

with $f^{[\rho]} g^{[\rho]} = h^{[\rho]}$. □

Lemma 4.20. *Let $f_\alpha \in L^\infty(\mathbb{T})$ be Laurent polynomials with $\alpha = 1, \dots, \nu$, $\nu < \infty$. Let*

$$h = \prod_{\alpha=1}^{\nu} f_\alpha, \quad h^{[\rho]} = \prod_{\alpha=1}^{\nu} f_\alpha^{[\rho]},$$

be a new Laurent polynomial and let $\{A_n\}$ be defined as $A_n = \prod_{\alpha=1}^{\nu} T_n(f_\alpha)$. For each $\rho > 0$ we have

$$\left\| D_\rho A_n D_\rho^{-1} - D_\rho T_n(h) D_\rho^{-1} \right\|_1 = o(n),$$
$$\lim_{n\to\infty} \frac{\operatorname{tr}\left(D_\rho A_n D_\rho^{-1}\right)}{n} = \frac{1}{2\pi} \int_Q h^{[\rho]}\left(e^{it}\right) \mathrm{d}t.$$

Proof. The same reasoning as above shows that

$$D_\rho A_n D_\rho^{-1} = \prod_{\alpha=1}^{\nu} T_n\left(f_\alpha^{[\rho]}\right), \quad \text{and} \quad D_\rho T_n(h) D_\rho^{-1} = T_n\left(h^{[\rho]}\right),$$

so this lemma is a direct consequence of Lemma 4.14 with $d = 1$. □

Lemma 4.21. *Let $f_{\alpha,\beta} \in L^\infty(\mathbb{T})$ be Laurent polynomials with $\alpha = 1, \dots, \nu$, $\beta = 1, \dots, v_\alpha$, $\nu, v_\alpha < \infty$. Let*

$$h = \sum_{\alpha=1}^{\nu} \prod_{\beta=1}^{v_\alpha} f_{\alpha,\beta}, \quad h^{[\rho]} = \sum_{\alpha=1}^{\nu} \prod_{\beta=1}^{v_\alpha} f_{\alpha,\beta}^{[\rho]},$$

be a new Laurent polynomial and let $\{A_n\}$ be defined as $A_n = \sum_{\alpha=1}^{\nu} \prod_{\beta=1}^{v_\alpha} T_n(f_{\alpha,\beta})$. For each $\rho > 0$ we have

$$\left\| D_\rho A_n D_\rho^{-1} - D_\rho T_n(h) D_\rho^{-1} \right\|_1 = o(n),$$
$$\lim_{n\to\infty} \frac{\operatorname{tr}\left(D_\rho A_n D_\rho^{-1}\right)}{n} = \frac{1}{2\pi} \int_Q h^{[\rho]}\left(e^{it}\right) \mathrm{d}t.$$

Proof. Once again, we apply (4.19) to see that

$$D_\rho A_n D_\rho^{-1} = \sum_{\alpha=1}^{\nu} \prod_{\beta=1}^{v_\alpha} T_n\left(f_{\alpha,\beta}^{[\rho]}\right), \quad \text{and} \quad D_\rho T_n(h) D_\rho^{-1} = T_n\left(h^{[\rho]}\right), \tag{4.21}$$

so a direct application of Lemma 4.15, with $d = 1$, gives the desired result. □

Theorem 4.22. *Let $f_{\alpha,\beta} \in L^\infty(\mathbb{T})$ be Laurent polynomials with $\alpha = 1, \dots, \nu$, $\beta = 1, \dots, v_\alpha$, $\nu, v_\alpha < \infty$. Let*

$$h = \sum_{\alpha=1}^{\nu} \prod_{\beta=1}^{v_\alpha} f_{\alpha,\beta}, \quad h^{[\rho]} = \sum_{\alpha=1}^{\nu} \prod_{\beta=1}^{v_\alpha} f_{\alpha,\beta}^{[\rho]},$$

be a new Laurent polynomial and let $\{A_n\}$ *be defined as* $A_n = \sum_{\alpha=1}^{\nu} \prod_{\beta=1}^{v_\alpha} T_n(f_{\alpha,\beta})$. *Denoting by* $\mathcal{S}\left(h^{[\rho]}\right)$ *the essential range of* $h^{[\rho]}$, *for each* $\rho > 0$, *the set* $Area\left(\mathcal{S}\left(h^{[\rho]}\right)\right)$ *is a weak cluster for* $\{A_n\}$.

Proof. We apply Theorem 3.6 to the sequence $\left\{D_\rho A_n D_\rho^{-1}\right\}$ using the equations (4.21). Condition **(c1)** is obtained by repeatedly applying the triangle inequality to $\left\|\sum_{\alpha=1}^{\nu}\prod_{\beta=1}^{v_\alpha} f_{\alpha,\beta}^{[\rho]}\right\|_{L^\infty}$; **(c2)** is a consequence of Lemma 4.21, since any positive integer power of a linear combination of products is still linear combination of products; **(c3)** is true, in light of Theorem 3.7, since $\left\{P\left(D_\rho^{-1} A_n D_\rho\right)\right\} \sim_\sigma \left(P\left(h^{[\rho]}\right), \mathbb{T}\right)$ for every polynomial P as a consequence of Lemma 4.21. Therefore Theorem 3.6 implies that the sequence $\left\{D_\rho A_n D_\rho^{-1}\right\}$ is weakly clustered at $Area\left(\mathcal{S}\left(h^{[\rho]}\right)\right)$. Since A_n has the same eigenvalues as $D_\rho A_n D_\rho^{-1}$ this means that the sequence $\{A_n\}$ is also weakly clustered at $Area\left(\mathcal{S}\left(h^{[\rho]}\right)\right)$. □

Theorem 4.23. *With the same notation as in Theorem 4.22,* $\bigcap_{\rho>0} Area\left(\mathcal{S}\left(h^{[\rho]}\right)\right)$ *is a weak cluster both for* $\{A_n\}$ *and for* $\{T_n(h)\}$.

Proof. This follows from Theorem 4.22. □

Now, we use the result of Hirschmann theorem in (4.20), with $f = h$ and so that $\mathcal{S}(\theta_h) \subseteq \bigcap_{\rho>0} Area\left(\mathcal{S}\left(h^{[\rho]}\right)\right)$, in order to prove the following theorem.

Theorem 4.24. *Let* $f_{\alpha,\beta} \in L^\infty(\mathbb{T})$ *be Laurent polynomials with* $\alpha = 1, \ldots, \nu$, $\beta = 1, \ldots, v_\alpha$, $\nu, v_\alpha < \infty$ *and let*

$$h = \sum_{\alpha=1}^{\nu} \prod_{\beta=1}^{v_\alpha} f_{\alpha,\beta},$$

be a new Laurent polynomial satisfying the hypotheses of the Hirschmann theorem. Let $\{A_n\}$ *be defined as* $A_n = \sum_{\alpha=1}^{\nu} \prod_{\beta=1}^{v_\alpha} T_n(f_{\alpha,\beta})$, *and set* $G_h = \bigcap_{\rho>0} Area\left(\mathcal{S}\left(h^{[\rho]}\right)\right)$. *If* $\mathbb{C} \setminus G_h$ *is connected in the complex field and the interior of* G_h *is empty, then* $\{A_n\} \sim_\lambda (\theta_h, G_h)$ *where* θ_h *is the distribution function of* $\{T_n(h)\}$ *indicated in (4.20), see* [18].

Proof. We will use Theorem 2.15. First we see that **(a1)** holds since G_h is compact by construction and $\mathbb{C} \setminus G_h$ is connected by the hypotheses. Condition **(a2)** is a consequence of Theorem 4.23; while **(a3)** follows from a repeated application of the triangle inequality to $\left\|\sum_{\alpha=1}^{\nu}\prod_{\beta=1}^{v_\alpha} f_{\alpha,\beta}\right\|_{L^\infty}$.

Condition **(a4)** amounts in proving that

$$\lim_{n\to\infty} \frac{\operatorname{tr}\left(A_n^L\right)}{n} = \frac{1}{m\{G_h\}} \int_{G_h} \theta_h^L(t)\,\mathrm{d}t. \tag{4.22}$$

In fact, from Lemma 4.15, with $d = 1$, we find $A_n = T_n(h) + R_{n,h}$ where $\|R_{n,h}\|_1 = o(n)$ and, in addition, by assumption $\{T_n(h)\} \sim_\lambda (\theta_h, G_h)$ (this second claim is indeed the Hirschmann result).

With these ingredients, we now prove formula (4.22). Since

$$\operatorname{tr}(X) = \sum_{\lambda\in\Lambda(X)} \lambda = \sum_{k=1}^{n} [X]_{k,k},$$

and since $\mathrm{tr}(\cdot)$ is a linear functional, the assumption $A_n = T_n(h) + R_{n,h}$ implies that

$$\mathrm{tr}(A_n) - \mathrm{tr}(T_n(h)) = \mathrm{tr}(R_{n,h}).$$

Consequently

$$\begin{aligned}\left|\frac{1}{n}\mathrm{tr}(A_n) - \frac{1}{n}\mathrm{tr}(T_n(h))\right| &= \left|\frac{1}{n}\mathrm{tr}(R_{n,h})\right| \\ &\underset{(a)}{\leq} \frac{1}{n}\|R_{n,h}\|_1 \\ &\underset{(b)}{\leq} \frac{1}{n}o(n) = o(1),\end{aligned}$$

where (a) follows from (1.14) and (b) follows from Lemma 4.15 (with $d = 1$). Since $T_n(h)$ is distributed as θ_h over G_h, we infer

$$\lim_{n\to\infty}\frac{1}{n}\mathrm{tr}(A_n) = \lim_{n\to\infty}\frac{1}{n}\mathrm{tr}(T_n(h)) = \frac{1}{m\{G_h\}}\int_{G_h}\theta_h(t)\,\mathrm{d}t,$$

therefore (4.22) is satisfied in the special case where $L = 1$.

Now we consider all non-negative integers $L > 0$. For $L = 0, 1$ the result is valid, so that we focus our attention to the case where $L \geq 2$. Relation $A_n = T_n(h) + R_{n,h}$ implies

$$\begin{aligned}A_n^L &= (T_n(h) + R_{n,h})^L \\ &= T_n(h)^L + \widetilde{R}_{n,h},\end{aligned}$$

where $\widetilde{R}_{n,h}$ is a term of the form

$$\widetilde{R}_{n,h} = \sum_{X_i\in\{T_n(h),R_{n,h}\}}(X_1\cdots X_L) - T_n(h)^L. \tag{4.23}$$

In other words the error matrix $\widetilde{R}_{n,h}$ is the sum of all possible combinations of products of j matrices $T_n(h)$ and k matrices $R_{n,h}$, with $j+k = L$ and the exception of $j = L$ (obviously it is understood that all the addends are pair-wise different). By using a simple Hölder's inequality involving Schatten p-norms (see (1.13)), for every summand R in (4.23), we deduce that there exists $j \geq 1$, $k = L - j$ for which

$$\begin{aligned}\|R\|_1 &\leq \|T_n(h)\|^k \|R_{n,h}\|^{j-1}\|R_{n,h}\|_1 \\ &\underset{(a)}{\leq} C^k C^{j-1}o(n),\end{aligned} \tag{4.24}$$

where (a) follows from the assumptions:

$$\begin{aligned}&\|T_n(h)\| \leq \|h\|_{L^\infty} \leq C < \infty, \\ &\|R_{n,h}\| = \|A_n - T_n(h)\| \leq C < \infty.\end{aligned}$$

Therefore by the triangle inequality and by applying inequality (4.24) to any summand in (4.23), we find $\left\|\widetilde{R}_{n,h}\right\|_1 \leq \hat{K}o(n)$, with $\hat{K} = \hat{K}(L)$ constant independent of n. Consequently

$\mathrm{tr}\left(A_n^L\right) - \mathrm{tr}\left(T_n\left(h\right)^L\right) = \mathrm{tr}\left(\widetilde{R}_{n,h}\right)$, and, since $\lambda\left(X^L\right) = \lambda^L\left(X\right)$, we have

$$\begin{aligned}\left|\frac{1}{n}\mathrm{tr}\left(A_n^L\right) - \frac{1}{n}\mathrm{tr}\left(T_n\left(h\right)^L\right)\right| &= \left|\frac{1}{n}\sum_{\lambda\in\Lambda(A_n)}\lambda^L - \frac{1}{n}\sum_{\lambda\in\Lambda(T_n(h))}\lambda^L\right| \\ &= \left|\frac{1}{n}\sum_{\lambda\in\Lambda\left(\widetilde{R}_{n,h}\right)}\lambda\right| \\ &\leq \frac{1}{n}\left\|\widetilde{R}_{n,h}\right\|_1 \\ &\leq \frac{1}{n}\hat{K}o\left(n\right) = o\left(1\right).\end{aligned}$$

Since $T_n\left(h\right)$ is distributed as θ_h over G_h, we infer

$$\lim_{n\to\infty}\frac{1}{n}\mathrm{tr}\left(A_n^L\right) = \lim_{n\to\infty}\frac{1}{n}\mathrm{tr}\left(T_n\left(h\right)^L\right) = \frac{1}{m\left\{G_h\right\}}\int_{G_h}\theta_h\left(t\right)^L\,\mathrm{d}t.$$

The latter proves that (4.22) is satisfied for any non-negative integer L.

Condition **(a5)** is true since $\mathcal{S}\left(\theta_h\right) \subset G_h$; finally G_h has empty interior by hypothesis. Therefore we can apply Theorem 2.15 and we conclude that $\{A_n\} \sim_\lambda \left(\theta_h, G_h\right)$. □

4.3.3 A complex analysis consequence for H^∞ functions

Let us consider the space $\mathcal{H}$ given by L^∞ functions defined on $\mathbb{T}^d$, $d \geq 1$; (where $\mathbb{T}$ is the unit circle in the complex plane) such that the Fourier coefficients $\tilde{f}_j$, $j = (j_1, \ldots, j_d) \in \mathbb{Z}^d$, defined as in (4.2) are equal to zero if $j_k < 0$ for some k with $1 \leq k \leq d$.

Theorem 4.25. *If $h \in \mathcal{H}$, $\left[\mathcal{S}\left(h\right)\right]^C$ is connected, and the interior of $\mathcal{S}\left(h\right)$ is empty, then h is necessarily constant almost everywhere.*

Proof. By [112, Theorem 2] (or equivalently, by Theorem 4.18 with $\rho = 1$ and $v_1 = 1$) we know that $\{T_n\left(h\right)\} \sim_\lambda \left(h, \mathbb{T}^d\right)$. However $T_n\left(h\right)$ is lower triangular with $\tilde{h}_0$ on the main diagonal since $\tilde{h}_j = 0$ if there exists k, $1 \leq k \leq d$, with $j_k < 0$. Therefore it is also true that $\{T_n\left(h\right)\} \sim_\lambda \left(\tilde{h}_0, \mathbb{T}^d\right)$, i.e., $h \equiv \tilde{h}_0$ and the proof is concluded. □

In other words, if $f \in \mathcal{H}$ and it is not constant almost everywhere, then its essential range necessarily divides the complex field in (at least two) unconnected components or its interior is not empty. Since a function is in $\mathcal{H}$ if and only if it is equal to the boundary values of a function in H^∞ this rigidity is not surprising.

From an operator theory viewpoint the proof is as follows (A. Böttcher has suggested the following alternative proof). Since $\mathcal{H}$ is a closed subalgebra of L^∞, the spectrum of h in the subalgebra results from the spectrum of h in L^∞ by filling in holes. Thus, if the first set has no holes, then the two sets coincides and are equal to a set without interior points. As the second set is the closure of h over the polydisc, which contains interior points if h is not constant, it follows that h must be constant.

4.3.4 Some issues from statistics

Given the function $W_n : C^0\left(\Pi, \mathbb{R}\right) \to \mathbb{R}$ where $C^0\left(\Pi, \mathbb{R}\right)$ is the space of real continuous functions on the circle and W_n is defined by:

$$W_n\left(f\right) = \frac{1}{2\pi n}\int_\Pi f\left(t\right)\left|\sum_{j=0}^{n} X_j e^{(\mathrm{i}jt)}\right|^2\mathrm{d}t,$$

where (X_n) is a centered stationary real Gaussian process, we have that, if the spectral density of (X_n) is the positive bounded function g then

$$W_n(f) = \frac{1}{n} Y^{(n)} T_n(g)^{\frac{1}{2}} T_n(f) T_n(g)^{\frac{1}{2}} Y^{(n)},$$

where the vector $Y^{(n)}$ has a Gaussian $\mathcal{N}(0, I_n)$ distribution. We hope that the results of this chapter may be useful. In fact the matrix $T_n(g)^{\frac{1}{2}} T_n(f) T_n(g)^{\frac{1}{2}}$ is similar to $T_n(g) T_n(f)$ since $T_n(g)$ is Hermitian positive definite. As a consequence in view of item **c)** in Theorem 2.16 and in view of Theorem 4.7, we can claim that the eigenvalue distribution of the the sequence $\left\{T_n(g)^{\frac{1}{2}} T_n(f) T_n(g)^{\frac{1}{2}}\right\}$ is $h = fg$ and that its maximal eigenvalue has lim sup bounded from above by $\|f\|_{L^\infty} \|g\|_{L^\infty}$ and lim inf bounded from below by $\|h\|_{L^\infty}$.

Chapter 5

Spectral features and asymptotic properties for g-circulants and g-Toeplitz sequences

In this chapter we address the problem of characterizing the singular values and the eigenvalues of g-circulants and of providing an asymptotic analysis of the distribution results for the singular values of g-Toeplitz sequences in the case where the sequence of values $\{a_k\}$, defining the entries of the matrices, can be interpreted as the sequence of Fourier coefficients of an integrable function f over the domain $(-\pi, \pi)$. We generalize the analysis to the block, multi-level case, amounting to choosing the symbol f multivariate, i.e., defined on the set $(-\pi, \pi)^d$ for some $d > 1$, and matrix-valued, i.e., such that $f(x)$ is a matrix of given size $p \times q$. As a byproduct, we will see in Chapter 7 interesting relations between g-circulant matrices and the analysis of convergence of multigrid methods given, e.g., in [95, 3].

5.1 Circulant and g-circulant matrices

The circulant matrices have been deeply studied in the literature (see, e.g., [30, 95, 115]) and much is known about their algebraic and spectral characterization.

A generic circulant matrix is generated starting from n elements, those on the first column of the matrix, and the remaining columns are obtained with a circular shift down one positions of the previous column:

$$C_n = \left[a_{(r-s) \bmod n}\right]_{r,s=0}^{n-1} = \begin{bmatrix} a_0 & a_{n-1} & \cdots & a_2 & a_1 \\ a_1 & a_0 & \ddots & \vdots & a_2 \\ a_2 & a_1 & \ddots & a_{n-1} & \vdots \\ \vdots & a_2 & \ddots & a_0 & a_{n-1} \\ a_{n-1} & \cdots & a_2 & a_1 & a_0 \end{bmatrix}, \tag{5.1}$$

and is straightforward to verify that it can be written as

$$C_n = \sum_{j=0}^{n-1} a_j Z_n^j,$$

where the matrix

$$Z_n = \begin{bmatrix} 0 & \cdots & 0 & 1 \\ 1 & & & 0 \\ & \ddots & & \vdots \\ 0 & & 1 & 0 \end{bmatrix}, \tag{5.2}$$

is the cyclic permutation Toeplitz matrix. Moreover, if $F_n \in M_n(\mathbb{C})$ denotes the Fourier matrix, i.e.

$$F_n = \frac{1}{\sqrt{n}} \left[e^{-\frac{2\pi \mathrm{i} jk}{n}} \right]_{j,k=0}^{n-1}, \tag{5.3}$$

then F_n is unitary, $F_n F_n^* = I_n$, and it is well-known (see, e.g., [30]) that

$$C_n = F_n D_n F_n^*, \tag{5.4}$$

where

$$\begin{aligned} D_n &= \operatorname{diag}\left(\sqrt{n} F_n^* \underline{a}\right), \qquad \underline{a} = [a_0, a_1, \ldots, a_{n-1}]^\top, \\ &= \operatorname*{diag}_{j=0,\ldots,n-1} \left(\sum_{k=0}^{n-1} a_k e^{\frac{2\pi \mathrm{i} jk}{n}} \right), \end{aligned} \tag{5.5}$$

$\underline{a}$ being the first column of the matrix C_n. Since F_n is a unitary matrix, the diagonal elements of D_n are the eigenvalues of C_n; then the circulant matrices form a commutative algebra simultaneously diagonalized by the unitary transform F_n.

Let p be a trigonometric polynomial defined over the set $Q = (-\pi, \pi)$ and having degree $c \geq 0$, i.e., $p(t) = \sum_{k=-c}^{c} a_k e^{\mathrm{i}kt}$, $\mathrm{i}^2 = -1$. From the Fourier coefficients a_k of p (see (4.2) with $d = 1$) one can build the circulant matrix $C_n(p) = \left[a_{(r-s) \bmod n} + a_{(r-s) \bmod n - n} \right]_{r,s=0}^{n-1}$ For example, we take $p(u) = 3 - 2e^{iu} + e^{-2iu} + 4e^{3iu}$. The degree of p is $c = 3$ and we have $a_0 = 3$, $a_1 = -2$, $a_3 = 4$ and $a_{-2} = 1$; if we take $n = 5$, the circulant matrix $C_5(p)$ is given by

$$C_5(p) = \begin{bmatrix} 3 & 0 & 5 & 0 & -2 \\ -2 & 3 & 0 & 5 & 0 \\ 0 & -2 & 3 & 0 & 5 \\ 5 & 0 & -2 & 3 & 0 \\ 0 & 5 & 0 & -2 & 3 \end{bmatrix}.$$

It is clear that the Fourier coefficient a_j equals zero if the condition $|j| \leq c$ is violated. The matrix $C_n(p)$ is said to be the circulant matrix of order n generated by p, and, following (5.1) and (5.4), it can be written as $C_n(p) = \sum_{|j| \leq c} a_j Z_n^j$, or, equivalently, as $C_n(p) = F_n D_n(p) F_n^*$ where, in this case, it is immediate to observe (from (5.5) and the expression of the polynomial p) that the matrix $D_n(p)$ is given by

$$D_n(p) = \operatorname*{diag}_{j=0,\ldots,n-1} \left(p\left(x_j^{(n)} \right) \right), \qquad x_j^{(n)} = \frac{2\pi j}{n}, \tag{5.6}$$

then the eigenvalues of $C_n(p)$ are the evaluations of the polynomial p at the grid points $\frac{2\pi j}{n}$, $j = 0, \ldots, n-1$.

Under the assumption that $c \leq \left\lfloor \frac{n-1}{2} \right\rfloor$, the matrix $C_n(p)$ is the Strang or natural circulant preconditioner of the corresponding Toeplitz matrix $T_n(p) = \left[a_{(r-s)} \right]_{r,s=0}^{n-1}$ (see [24] and the references therein). We observe that the above-mentioned assumption $c \leq \left\lfloor \frac{n-1}{2} \right\rfloor$ is fulfilled for n large enough, since c is a fixed constant and n is the matrix order: in actuality, in real applications it is natural to suppose that n is large, if we assume that $C_n(p)$ comes from an approximation process of an infinite-dimensional problem. Furthermore, if the symbol p has a zero at zero (this happens in the case of approximation of differential operators), then $C_n(p)$ is singular and it is usually replaced by a rank-one correction that forces invertibility: in the relevant literature, the latter is called modified Strang preconditioner.

A matrix $C_{n,g} \in M_n(\mathbb{C})$ is called g-circulant if its entries obey the rule

$$C_{n,g} = \left[a_{(r-gs) \bmod n}\right]_{r,s=0}^{n-1}; \tag{5.7}$$

for an introduction and for the algebraic properties of such matrices, refer to the classical book by Davis [30, Section 5.1], while new additional results can be found in [113] and the references therein. For instance, if $n = 5$ and $g = 3$, then we have

$$C_{5,3} = \begin{bmatrix} a_0 & a_2 & a_4 & a_1 & a_3 \\ a_1 & a_3 & a_0 & a_2 & a_4 \\ a_2 & a_4 & a_1 & a_3 & a_0 \\ a_3 & a_0 & a_2 & a_4 & a_1 \\ a_4 & a_1 & a_3 & a_0 & a_2 \end{bmatrix}.$$

Also in this case, as in ordinary circulant setting, if the coefficients a_j, $j \in \mathbb{Z}$, arise from a given symbol p (see (4.2)), we can build the g-circulant matrix generated by p as $C_{n,g}(p) = \left[a_{(r-gs) \bmod n} + a_{(r-gs) \bmod n-n}\right]_{r,s=0}^{n-1}$. For instance, with $p(u) = 3 - 2e^{iu} + e^{-2iu} + 4e^{3iu}$, we find $a_0 = 3$, $a_1 = -2$, $a_3 = 4$ and $a_{-2} = 1$ so that

$$C_{5,3}(p) = \begin{bmatrix} 3 & 0 & 0 & -2 & 5 \\ -2 & 5 & 3 & 0 & 0 \\ 0 & 0 & -2 & 5 & 3 \\ 5 & 3 & 0 & 0 & -2 \\ 0 & -2 & 5 & 3 & 0 \end{bmatrix}.$$

It is immediate to observe that a generic g-circulant matrix is constructed starting from n elements, those on the first column of the matrix, and the remaining columns are obtained with a circular shift down g positions of the previous column, therefore, the circulant matrices are g-circulant matrices with $g = 1$.

If $C_{n,g}$ is the g-circulant matrix generated by the same elements of the circulant matrix C_n (the two matrices have the same first column), for generic n and g one verifies immediately that

$$C_{n,g} = C_n Z_{n,g}, \tag{5.8}$$

where

$$Z_{n,g} = [\delta_{r-gs}]_{r,s=0}^{n-1}, \qquad \delta_k = \begin{cases} 1 & \text{if } k \equiv 0 \pmod{n}, \\ 0 & \text{otherwise.} \end{cases} \tag{5.9}$$

Proof. (of relation (5.8).) For $j, k = 0, 1, \ldots, n-1$ one has

$$\begin{aligned} (C_{n,g})_{j,k} &= a_{(j-gk) \bmod n}, \\ (C_n)_{j,k} &= a_{(j-k) \bmod n}, \\ (C_{n,g})_{j,k} &= (C_n)_{j,gk}, \end{aligned} \tag{5.10}$$

then, from (5.9),

$$\begin{aligned} (C_n Z_{n,g})_{j,k} &= \sum_{\ell=0}^{n-1} (C_n)_{j,\ell} (Z_{n,g})_{\ell,k} \\ &= a_{(j-\ell) \bmod n} \delta_{\ell-gk} \\ &\underset{\text{(a)}}{=} a_{(j-gk) \bmod n} \\ &= (C_{n,g})_{j,k}, \end{aligned}$$

where (a) follows from the fact that there exists a unique $\ell \in \{0, 1, \ldots, n-1\}$ such that $\ell - gk \equiv 0 \pmod{n}$, that is $\ell \equiv gk \pmod{n}$, so

$$(j - \ell) \bmod n = (j - (gk) \bmod n) \bmod n = (j - gk) \bmod n.$$

□

Remark 5.1. *We can consider the parameter g only non-negative. Indeed, the case of non-positive g can be reduced to the case of a non-negative g. In fact, the role of circulants will be played by (-1)-circulant matrices (also called anticirculants or skew-circulants),* [30]*: as for the circulants, (-1)-circulants form a commutative algebra simultaneously diagonalized by another unitary transform that can be written as the product of the Fourier matrix and a diagonal unitary matrix.*

Remark 5.1 and the following lemma tell us that we can consider the parameter g only in the interval $[0, n)$.

Lemma 5.2. *If $g \geq n$, then $Z_{n,g} = Z_{n,g^\circ}$, where $g^\circ \equiv g \pmod{n}$ and $Z_{n,g}$ is defined in (5.9), so $C_{n,g} = C_{n,g^\circ}$.*

Proof. From (5.9) we know that, for $r, s = 0, 1, \ldots, n-1$, one has

$$(Z_{n,g})_{r,s} = \delta_{r-gs} = \delta_{r-(tn+g^\circ)s} = \delta_{r-g^\circ s} = (Z_{n,g^\circ})_{r,s},$$

since $tns \equiv 0 \pmod{n}$. Whence $Z_{n,g} = Z_{n,g^\circ}$.

So, from (5.8) we infer that

$$C_{n,g} = C_n Z_{n,g} = C_n Z_{n,g^\circ} = C_{n,g^\circ}.$$

□

Finally, it is worth noticing that the use of (5.4) and (5.8) implies that

$$C_{n,g} = F_n D_n F_n^* Z_{n,g}. \tag{5.11}$$

Formula (5.11) plays an important role for studying the singular values of the g-circulant matrices.

5.1.1 A characterization of $Z_{n,g}$ in terms of Fourier matrices

The relations between the specific matrix $Z_{n,g}$ and the Fourier matrices was explained for $g = 2$ in the multigrid literature (see, e.g., [42, 95]). Here we report an extension to the case of a generic g and the details of the proof which uses the same tools as in [42, 95].

First we need some preparatory straightforward results. In the following, we denote by (n, g) the greatest common divisor of n and g, i.e., $(n, g) = \gcd(n, g)$, and by $I_t \in M_t(\mathbb{R})$ the identity matrix, while the quantities n_g and $\breve{g}$ are defined, respectively, as $n_g = \frac{n}{(n,g)}$ and $\breve{g} = \frac{g}{(n,g)}$.

Lemma 5.3. *Let n be any integer greater than 2, then*

$$Z_{n,g} = \underbrace{\left[\widetilde{Z}_{n,g} | \widetilde{Z}_{n,g} | \cdots | \widetilde{Z}_{n,g}\right]}_{(n,g)\ times}, \tag{5.12}$$

where $Z_{n,g}$ is the matrix defined in (5.9) and $\widetilde{Z}_{n,g} \in M_{n,n_g}(\mathbb{R})$ is the submatrix of $Z_{n,g}$ obtained by considering only its first n_g columns, that is,

$$\widetilde{Z}_{n,g} = Z_{n,g} \begin{bmatrix} I_{n_g} \\ 0 \end{bmatrix}. \tag{5.13}$$

Therefore $\widetilde{Z}_{n,g}^{\top}\widetilde{Z}_{n,g} = I_{n_g}$. *Finally if* $\widehat{Z}_{n,g} \in \mathbb{C}^{n\times\mu_g}$, $\mu_g = \left\lceil \frac{n}{g} \right\rceil$, *denotes the matrix* $Z_{n,g}$ *by considering only the* μ_g *first columns, then* $1 \le (n,g) \le g$, $\mu_g \le n_g \le n$, *and*

$$\widetilde{Z}_{n,g}^{\top}\widehat{Z}_{n,g} = \begin{bmatrix} I_{\mu_g} \\ 0 \end{bmatrix}.$$

Proof. Setting $\widetilde{Z}_{n,g}^{(0)} = \widetilde{Z}_{n,g}$ and denoting by $\widetilde{Z}_{n,g}^{(j)} \in M_{n,n_g}(\mathbb{R})$ the $(j+1)$th block-column of the matrix $Z_{n,g}$ for $j = 0, \dots, (n,g)-1$, we find

$$Z_{n,g} = \left[\underbrace{\widetilde{Z}_{n,g}^{(0)}}_{n\times n_g} \mid \underbrace{\widetilde{Z}_{n,g}^{(1)}}_{n\times n_g} \mid \cdots \mid \underbrace{\widetilde{Z}_{n,g}^{((n,g)-1)}}_{n\times n_g} \right].$$

For $r = 0, 1, \dots, n-1$ and $s = 0, 1, \dots, n_g - 1$, we observe that

$$\left(\widetilde{Z}_{n,g}^{(j)}\right)_{r,s} = (Z_{n,g})_{r,jn_g+s},$$

and

$$\begin{aligned} (Z_{n,g})_{r,jn_g+s} &= \delta_{r-g(jn_g+s)} \\ &= \delta_{r-jgn_g-gs} \\ &\underset{\text{(a)}}{=} \delta_{r-gs} \\ &= \left(\widetilde{Z}_{n,g}^{(0)}\right)_{r,s} = \left(\widetilde{Z}_{n,g}\right)_{r,s}, \end{aligned}$$

where $n_g = \frac{n}{(n,g)}$ and (a) is a consequence of the fact that $\frac{g}{(n,g)}$ is an integer greater than zero and so $jgn_g = j\frac{g}{(n,g)}n \equiv 0 \pmod{n}$. Thus we conclude that $\widetilde{Z}_{n,g}^{(j)} = \widetilde{Z}_{n,g}^{(0)} = \widetilde{Z}_{n,g}$ for $j = 0, \dots, (n,g)-1$.

Now we consider the product $\widetilde{Z}_{n,g}^{\top}\widetilde{Z}_{n,g} = I_{n_g}$, from (5.13) we have that, for $j,k = 0, \dots, n_g - 1$

$$\begin{aligned} \left(\widetilde{Z}_{n,g}^{\top}\widetilde{Z}_{n,g}\right)_{j,k} &= \left([I_{n_g}|0]\, Z_{n,g}^{\top} Z_{n,g} \begin{bmatrix} I_{n_g} \\ \hline 0 \end{bmatrix} \right)_{j,k} \\ &= \sum_{\ell=0}^{n-1} \left(Z_{n,g}^{\top}\right)_{j,\ell} (Z_{n,g})_{\ell,k} \\ &= \sum_{\ell=0}^{n-1} \delta_{\ell-gj}\delta_{\ell-gk} \\ &\underset{\text{(a)}}{=} \delta_{(gj) \bmod n - gk} \\ &\underset{\text{(b)}}{=} \delta_{((gj) \bmod n - gk) \bmod n} \\ &\underset{\text{(c)}}{=} \delta_{(gj-gk) \bmod n} \\ &\underset{\text{(d)}}{=} \begin{cases} 1 & \text{if } j = k, \\ 0 & \text{otherwise,} \end{cases} \end{aligned}$$

where

(a) follows from the fact that there exists a unique $\ell \in \{0, 1, \dots, n-1\}$ such that $gj - \ell \equiv 0 \pmod{n}$, that is, $\ell \equiv gj \pmod{n}$;

(b) comes from the definition of δ in (5.9): $\delta_w = \delta_{(w) \bmod n}$;

(c) just remember that

$$(((gj) \bmod n) - gk) \bmod n = (gj - gk) \bmod n;$$

then

$$\widetilde{Z}_{n,g}^{\top} \widetilde{Z}_{n,g} = I_{n_g}. \tag{5.14}$$

Finally, if $\widehat{Z}_{n,g} \in \mathbb{C}^{n \times \mu_g}$, $\mu_g = \left\lceil \frac{n}{g} \right\rceil$, denotes the matrix $Z_{n,g}$ by considering only the μ_g first columns, we have that, since $\mu_g \leq n_g$

$$\begin{aligned} \widehat{Z}_{n,g} &= Z_{n,g} \left[\frac{I_{\mu_g}}{0} \right] \\ &= Z_{n,g} \left[\frac{I_{n_g}}{0} \right] \left[\frac{I_{\mu_g}}{0} \right] \\ &= \widetilde{Z}_{n,g} \left[\frac{I_{\mu_g}}{0} \right], \end{aligned} \tag{5.15}$$

where $\left[\frac{I_{\mu_g}}{0} \right] \in M_{n_g,\mu_g}(\mathbb{R})$. Using (5.14) and (5.15) we can conclude that

$$\begin{aligned} \widetilde{Z}_{n,g}^{\top} \widehat{Z}_{n,g} &= \widetilde{Z}_{n,g}^{\top} \widetilde{Z}_{n,g} \left[\begin{array}{cc} I_{\mu_g} & 0 \end{array} \right] \\ &= I_{n_g} \left[\begin{array}{cc} I_{\mu_g} & 0 \end{array} \right] \\ &= \left[\begin{array}{c} I_{\mu_g} \\ 0 \end{array} \right]. \end{aligned}$$

□

Another useful fact is represented by the following equation:

$$\widetilde{Z}_{n,g} = \widetilde{Z}_{n,(n,g)} Z_{n_g,\check{g}}, \tag{5.16}$$

where $Z_{n_g,\check{g}} \in M_{n_g}(\mathbb{R})$ is the matrix defined in (5.9). Therefore

$$Z_{n_g,\check{g}} = \left[\widehat{\delta}_{r-\check{g}s} \right]_{r,s=0,}^{n_g-1} \qquad \widehat{\delta}_k = \begin{cases} 1 & \text{if } k \equiv 0 \pmod{n_g}, \\ 0 & \text{otherwise.} \end{cases} \tag{5.17}$$

Proof. (of relation (5.16).) We show that the two matrices $\widetilde{Z}_{n,g}$ and $\widetilde{Z}_{n,(n,g)} Z_{n_g,\check{g}}$ in (5.16) have the same elements. For $r = 0, 1, \ldots, n-1$ and $s = 0, 1, \ldots, n_g - 1$, we find

$$\begin{aligned} \left(\widetilde{Z}_{n,g} \right)_{r,s} &= \delta_{r-gs} \\ &= \delta_{(r-gs) \bmod n}, \end{aligned}$$

and

$$\begin{aligned} \left(\widetilde{Z}_{n,(n,g)} Z_{n_g,\check{g}} \right)_{r,s} &= \sum_{l=0}^{n_g-1} \left(\widetilde{Z}_{n,(n,g)} \right)_{r,l} \left(Z_{n_g,\check{g}} \right)_{l,s} \\ &= \sum_{l=0}^{n_g-1} \delta_{r-(n,g)l} \widehat{\delta}_{l-\check{g}s} \\ &\underset{\text{(a)}}{=} \delta_{r-(n,g)((\check{g}s) \bmod n_g)} \\ &= \delta_{r-(n,g)\left(\left(\frac{g}{(n,g)}s\right) \bmod n_g\right)} \\ &\underset{\text{(b)}}{=} \delta_{r-(gs) \bmod n} \\ &= \delta_{(r-(gs) \bmod n) \bmod n} \\ &= \delta_{(r-gs) \bmod n}, \end{aligned}$$

where

(a) holds true since there exists a unique $l \in \{0, 1, \ldots, n_g - 1\}$ such that $l - \check{g}s \equiv 0 \pmod{n_g}$, that is, $l \equiv \check{g}s \pmod{n_g}$ and hence $\delta_{r-(n,g)l} = \delta_{r-(n,g)((\check{g}s) \bmod n_g)}$;

(b) is due to the following property: if we have three integer numbers ρ, θ, and γ, then

$$\rho\,((\theta) \bmod \gamma) = (\rho\theta) \bmod \rho\gamma.$$

□

Lemma 5.4. *Let $F_n \in M_n(\mathbb{C})$ be the Fourier matrix defined in (5.3), and let $\widetilde{Z}_{n,g} \in M_{n,n_g}(\mathbb{R})$ be the matrix represented in (5.13). Then*

$$F_n \widetilde{Z}_{n,g} = \frac{1}{\sqrt{(n,g)}} I_{n,g} F_{n_g} Z_{n_g,\check{g}}, \tag{5.18}$$

where $I_{n,g} \in M_{n,n_g}(\mathbb{R})$ and

$$I_{n,g} = \left.\begin{bmatrix} I_{n_g} \\ \hline I_{n_g} \\ \hline \vdots \\ \hline I_{n_g} \end{bmatrix}\right\} (n,g) \text{ times},$$

with $I_{n_g} \in M_{n_g}(\mathbb{R})$ being the identity matrix and $Z_{n_g,\check{g}}$ as in (5.17).

Proof. Rewrite the Fourier matrix as

$$F_n = \frac{1}{\sqrt{n}} \left[\; f_0 \;\middle|\; f_1 \;\middle|\; f_2 \;\middle|\; \cdots \;\middle|\; f_{n-1} \;\right],$$

where f_k, $k = 0, 1, 2, \ldots, n-1$, is the $(k)\,th$ column of the Fourier matrix $F_n \in M_n(\mathbb{C})$:

$$f_k = \left[e^{-\frac{2\pi i k j}{n}}\right]_{j=0}^{n-1} = \begin{bmatrix} e^{-\frac{2\pi i k \cdot 0}{n}} \\ e^{-\frac{2\pi i k \cdot 1}{n}} \\ e^{-\frac{2\pi i k \cdot 2}{n}} \\ \vdots \\ e^{-\frac{2\pi i k \cdot (n-1)}{n}} \end{bmatrix}. \tag{5.19}$$

From (5.16), we find

$$\begin{aligned} F_n \widetilde{Z}_{n,g} &= F_n \widetilde{Z}_{n,(n,g)} Z_{n_g,\check{g}} \\ &= \frac{1}{\sqrt{n}} \left[\; f_0 \;\middle|\; f_{1\cdot(n,g)} \;\middle|\; f_{2\cdot(n,g)} \;\middle|\; \cdots \;\middle|\; f_{(n_g-1)\cdot(n,g)} \;\right] Z_{n_g,\check{g}} \in M_{n,n_g}(\mathbb{C}). \end{aligned} \tag{5.20}$$

Indeed, for $k = 0, 1, \ldots, n_g - 1$, $j = 0, 1, \ldots, n-1$, one has

$$\left(F_n \widetilde{Z}_{n,(n,g)}\right)_{j,k} = \sum_{l=0}^{n-1} (F_n)_{j,l} \left(\widetilde{Z}_{n,(n,g)}\right)_{l,k} = \sum_{l=0}^{n-1} \delta_{l-(n,g)k} e^{-\frac{2\pi i j l}{n}}, \tag{5.21}$$

and, since $0 \le (n,g)\,k \le n - (n,g)$, there exists a unique $l_k \in \{0, 1, 2, \ldots, n-1\}$ such that $l_k - (n,g)\,k \equiv 0 \pmod{n}$, so $l_k = (n,g)\,k$. Consequently relation (5.21) implies

$$\left(F_n \widetilde{Z}_{n,(n,g)}\right)_{j,k} = \delta_{l_k-(n,g)k} e^{-\frac{2\pi i j l_k}{n}} = e^{-\frac{2\pi i j (n,g) k}{n}} = \left(f_{(n,g)k}\right)_j,$$

for all $0 \leq j \leq n-1$ and $0 \leq k \leq n_g - 1$, and hence

$$F_n \widetilde{Z}_{n,(n,g)} = \frac{1}{\sqrt{n}} \left[\; f_0 \; \middle| \; f_{1\cdot(n,g)} \; \middle| \; f_{2\cdot(n,g)} \; \middle| \cdots \middle| \; f_{(n_g-1)\cdot(n,g)} \; \right].$$

For $k = 0, 1, 2, \ldots, n_g - 1$, we deduce

$$f_{(n,g)k} = \left[e^{-\frac{2\pi i j (n,g) k}{n}} \right]_{j=0}^{n-1} = \left[e^{-\frac{2\pi i j k}{n_g}} \right]_{j=0}^{n-1},$$

and then, taking into account the equalities $n = (n,g)\,\frac{n}{(n,g)} = (n,g)\,n_g$, we can write

$$f_{(n,g)k} = \begin{bmatrix} \left[e^{-\frac{2\pi i k j}{n_g}} \right]_{j=0}^{n_g-1} \\ \left[e^{-\frac{2\pi i k j}{n_g}} \right]_{j=n_g}^{2n_g-1} \\ \vdots \\ \left[e^{-\frac{2\pi i k j}{n_g}} \right]_{j=((n,g)-1)n_g}^{(n,g)n_g-1} \end{bmatrix}, \tag{5.22}$$

where

$$\left[e^{-\frac{2\pi i k j}{n_g}} \right]_{j=0}^{n_g-1} = \begin{bmatrix} e^{-\frac{2\pi i k \cdot 0}{n_g}} \\ e^{-\frac{2\pi i k \cdot 1}{n_g}} \\ e^{-\frac{2\pi i k \cdot 2}{n_g}} \\ \vdots \\ e^{-\frac{2\pi i k \cdot (n_g-1)}{n_g}} \end{bmatrix}. \tag{5.23}$$

According to formula (5.19), one observes that the vector in (5.23) is the $(k)\,th$ column of the Fourier matrix F_{n_g}. Furthermore, for $l = 0, 1, 2, \ldots, (n,g) - 1$, we find

$$\left[e^{-\frac{2\pi i k j}{n_g}} \right]_{j=ln_g}^{(l+1)n_g-1} = \begin{bmatrix} e^{-\frac{2\pi i k l n_g}{n_g}} \\ e^{-\frac{2\pi i k (ln_g+1)}{n_g}} \\ e^{-\frac{2\pi i k (ln_g+2)}{n_g}} \\ \vdots \\ e^{-\frac{2\pi i k (ln_g+n_g-1)}{n_g}} \end{bmatrix} = e^{-2\pi i k l} \begin{bmatrix} e^{-\frac{2\pi i k \cdot 0}{n_g}} \\ e^{-\frac{2\pi i k \cdot 1}{n_g}} \\ e^{-\frac{2\pi i k \cdot 2}{n_g}} \\ \vdots \\ e^{-\frac{2\pi i k \cdot (n_g-1)}{n_g}} \end{bmatrix} = \left[e^{-\frac{2\pi i k j}{n_g}} \right]_{j=0}^{n_g-1}. \tag{5.24}$$

Using (5.24), the expression of the vector in (5.22) becomes

$$f_{(n,g)k} = \left\{ \begin{bmatrix} \left[e^{-\frac{2\pi i k j}{n_g}} \right]_{j=0}^{n_g-1} \\ \left[e^{-\frac{2\pi i k j}{n_g}} \right]_{j=0}^{n_g-1} \\ \vdots \\ \left[e^{-\frac{2\pi i k j}{n_g}} \right]_{j=0}^{n_g-1} \end{bmatrix} \right. \; (n,g) \text{ times.} \tag{5.25}$$

Setting $\widehat{f}_r = \left[e^{-\frac{2\pi i r j}{n_g}} \right]_{j=0}^{n_g-1}$, for $0 \leq r \leq n_g - 1$, the Fourier matrix $F_{n_g} \in M_{n_g}(\mathbb{C})$ takes the form

$$F_{n_g} = \frac{1}{\sqrt{n_g}} \left[\; \widehat{f}_0 \; \middle| \; \widehat{f}_1 \; \middle| \; \widehat{f}_2 \; \middle| \cdots \middle| \; \widehat{f}_{n_g-1} \; \right].$$

From formula (5.23), the relation (5.25) can be expressed as

$$f_{(n,g)k} = \left.\begin{bmatrix} \widehat{f}_k \\ \widehat{f}_k \\ \vdots \\ \widehat{f}_k \end{bmatrix}\right\} (n,g) \text{ times}, \qquad k = 0, \ldots, n_g - 1,$$

and, as a consequence, formula (5.20) can be rewritten as

$$\begin{aligned} F_n \widetilde{Z}_{n,g} = F_n \widetilde{Z}_{n,(n,g)} Z_{n_g,\breve{g}} &= \frac{1}{\sqrt{n}} \left[\begin{array}{c|c|c|c|c} \widehat{f}_0 & \widehat{f}_1 & \widehat{f}_2 & \cdots & \widehat{f}_{n_g-1} \\ \widehat{f}_0 & \widehat{f}_1 & \widehat{f}_2 & \cdots & \widehat{f}_{n_g-1} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \widehat{f}_0 & \widehat{f}_1 & \widehat{f}_2 & \cdots & \widehat{f}_{n_g-1} \end{array}\right] Z_{n_g,\breve{g}} \\ &= \frac{1}{\sqrt{(n,g)\, n_g}} \left[\begin{array}{c} \sqrt{n_g} F_{n_g} \\ \hline \sqrt{n_g} F_{n_g} \\ \hline \vdots \\ \hline \sqrt{n_g} F_{n_g} \end{array}\right] Z_{n_g,\breve{g}} \\ &= \frac{1}{\sqrt{(n,g)}} \left[\begin{array}{c} F_{n_g} \\ \hline F_{n_g} \\ \hline \vdots \\ \hline F_{n_g} \end{array}\right] Z_{n_g,\breve{g}} \\ &= \frac{1}{\sqrt{(n,g)}} \left[\begin{array}{c} I_{n_g} \\ \hline I_{n_g} \\ \hline \vdots \\ \hline I_{n_g} \end{array}\right] F_{n_g} Z_{n_g,\breve{g}} \\ &= \frac{1}{\sqrt{(n,g)}} I_{n,g} F_{n_g} Z_{n_g,\breve{g}}. \end{aligned}$$

□

In the subsequent section, we will exploit Lemma 5.4 in order to characterize the singular values of the g-circulant matrices $C_{n,g}$.

Remark 5.5. *In Lemma 5.4, if $(n,g) = g$, we have $n_g = \frac{n}{(n,g)} = \frac{n}{g}$ and $\breve{g} = \frac{g}{(n,g)} = 1$; so the matrix $Z_{n_g,\breve{g}} = Z_{n_g,1}$, appearing in (5.18), is the identity matrix of dimension $\frac{n}{g} \times \frac{n}{g}$. The relation (5.18) becomes*

$$F_n \widetilde{Z}_{n,g} = \frac{1}{\sqrt{g}} I_{n,g} F_{n_g}.$$

The latter equation with $g = 2$ and even n appears (and is crucial) in the multigrid literature; see [95, (3.2), p. 59] *and, in slightly different form for the sine algebra of type I, see* [42, Section 2.1].

Remark 5.6. *If $(n,g) = 1$, Lemma 5.4 is trivial, because $n_g = \frac{n}{(n,g)} = n$, $\breve{g} = \frac{g}{(n,g)} = g$, and so $\widetilde{Z}_{n,g} = Z_{n,g}$. The relation (5.18) becomes*

$$\begin{aligned} F_n \widetilde{Z}_{n,g} = F_n Z_{n,g} &= I_{n,g} F_{n_g} Z_{n_g,\breve{g}} \\ &= F_n Z_{n,g}, \end{aligned}$$

since the matrix $I_{n,g}$ reduces by its definition to the identity matrix of order n.

Remark 5.7. *Lemma 5.4 is true also if, instead of F_n and F_{n_g}, we put F_n^* and $F_{n_g}^*$, respectively, because $F_n^* = \overline{F}_n$. In fact there is no transposition, but only conjugation.*

Now we see another characterization of the matrix $Z_{n,g}$ in terms of Fourier matrices that will be useful in Section 5.3 for the study of eigenvalues of g-circulant matrices.

Lemma 5.8. *Let $Z_{n,g} \in M_n(\mathbb{R})$ be the matrix defined in (5.9), and let $F_n \in M_n(\mathbb{C})$ be the Fourier matrix defined in (5.3), then we have that*

$$Z_{n,g} = F_n S_{n,g} F_n^*, \tag{5.26}$$

where

$$S_{n,g} = [\delta_{rg-c}]_{r,c=0}^{n-1}, \qquad \delta_k = \begin{cases} 1 & \text{if } k \equiv 0 \ (mod\ n), \\ 0 & \text{otherwise.} \end{cases} \tag{5.27}$$

Proof. It suffices to show that

$$F_n^* Z_{n,g} = S_{n,g} F_n^*.$$

For $j,k = 0,1,...,n-1$, we have that

$$(F_n^* Z_{n,g})_{j,k} = \sum_{\ell=0}^{n-1} (F_n^*)_{j,\ell} (Z_{n,g})_{\ell,k} = \sum_{\ell=0}^{n-1} (F_n^*)_{j,\ell}\, \delta_{\ell-gk} \underset{\text{(a)}}{=} (F_n^*)_{j,(gk) \bmod n}, \tag{5.28}$$

where (a) follows from the fact that there exists a unique $\ell \in \{0,1,\ldots,n-1\}$ such that $\ell - gk \equiv 0 \pmod n$ that is $\ell \equiv gk \pmod n$.

For $j,k = 0,1,...,n-1$, it holds

$$(S_{n,g} F_n^*)_{j,k} = \sum_{\ell=0}^{n-1} (S_{n,g})_{j,\ell} (F_n^*)_{\ell,k} = \sum_{\ell=0}^{n-1} \delta_{gj-\ell} (F_n^*)_{\ell,k} \underset{\text{(b)}}{=} (F_n^*)_{(gj) \bmod n,k}, \tag{5.29}$$

where (b) follows from the fact that there exists a unique $\ell \in \{0,1,\ldots,n-1\}$ such that $gj - \ell \equiv 0 \pmod n$, that is $\ell \equiv gj \pmod n$.

From (5.3), since $(F_n)_{j,k} = (F_n)_{k,j}$ is clear that, for $j,k = 0,\ldots,n-1$,

$$(F_n^*)_{j,k} = e^{\frac{2\pi i jk}{n}};$$

now, if $jk = qn + c$, for some $q, c \in \mathbb{N}$, with $c = ((jk) \bmod n) < n$, then applies that

$$e^{\frac{2\pi i jk}{n}} = e^{\frac{2\pi i (qn+c)}{n}} = e^{\frac{2\pi i qn}{n}} e^{\frac{2\pi i c}{n}} = \underbrace{e^{2\pi i q}}_{=1} e^{\frac{2\pi i c}{n}} = e^{\frac{2\pi i c}{n}},$$

this means that

$$e^{\frac{2\pi i jk}{n}} = e^{\frac{2\pi i ((jk) \bmod n)}{n}}. \tag{5.30}$$

Finally, from (5.28), (5.29) and (5.30), we obtain

$$\begin{aligned} (F_n^*)_{j,(gk) \bmod n} &= e^{\frac{2\pi i j((gk) \bmod n)}{n}} \\ &= e^{\frac{2\pi i [j((gk) \bmod n) \bmod n]}{n}} \\ &\underset{\text{(a)}}{=} e^{\frac{2\pi i [((gkj) \bmod nj) \bmod n]}{n}} \\ &\underset{\text{(b)}}{=} e^{\frac{2\pi i ((gkj) \bmod n)}{n}}, \end{aligned} \tag{5.31}$$

$$\begin{aligned} (F_n^*)_{(gj) \bmod n,k} &= e^{\frac{2\pi i k((gj) \bmod n)}{n}} \\ &= e^{\frac{2\pi i [k((gj) \bmod n) \bmod n]}{n}} \\ &\underset{\text{(a)}}{=} e^{\frac{2\pi i [((gkj) \bmod nk) \bmod n]}{n}} \\ &\underset{\text{(b)}}{=} e^{\frac{2\pi i ((gkj) \bmod n)}{n}}, \end{aligned} \tag{5.32}$$

where (a) is due to this property: if we have three integer numbers ρ, θ, and γ, then

$$\rho\left((\theta) \bmod \gamma\right) = (\rho\theta) \bmod \rho\gamma;$$

while (b) follows from the fact that, since $j, k \in \mathbb{N}^+$, jn and kn are multiples of n, then, given any number $w \in \mathbb{Z}$, is true that

$$\begin{aligned} (w) \bmod n &= [(w) \bmod jn] \bmod n; \\ (w) \bmod n &= [(w) \bmod kn] \bmod n. \end{aligned}$$

The equality between the relations (5.31) and (5.32), is equivalent to equality between the expressions (5.28) and (5.29); so (5.26) is proved. □

A similar result to the first part of Lemma 5.2 also applies for matrices $S_{n,g}$.

Lemma 5.9. *If $g \geq n$ then $S_{n,g} = S_{n,g^\circ}$ with $g^\circ \equiv g \pmod n$, where $S_{n,g}$ is defined in (5.27).*

Proof. For Lemma 5.8 we have that

$$\begin{aligned} F_n S_{n,g} F_n^* &= Z_{n,g}, \\ F_n S_{n,g^\circ} F_n^* &= Z_{n,g^\circ}. \end{aligned}$$

Now, by Lemma 5.2, it holds that $Z_{n,g} = Z_{n,g^\circ}$ with $g^\circ \equiv g \pmod n$, then

$$F_n S_{n,g} F_n^* = F_n S_{n,g^\circ} F_n^* \qquad \Rightarrow \qquad S_{n,g} = S_{n,g^\circ}.$$

□

Lemma 5.10. *Let $D \in M_n(\mathbb{C})$ be a diagonal matrix,*

$$D = \operatorname*{diag}_{j=0,\ldots,n-1} (d_j),$$

and let $S_{n,g}$ be the matrix defined in (5.27), then

$$S_{n,g} D = \widetilde{D} S_{n,g}, \tag{5.33}$$

where

$$\widetilde{D} = \operatorname*{diag}_{j=0,\ldots,n-1} \left(d_{(gj) \bmod n}\right). \tag{5.34}$$

Proof. We show that the two matrices $S_{n,g}D$ and $\widetilde{D}S_{n,g}$ in (5.33) have the same elements. For $j, k = 0, \ldots, n-1$, we have that

$$\begin{aligned} \left(S_{n,g} D\right)_{j,k} &= \sum_{\ell=0}^{n-1} \left(S_{n,g}\right)_{j,\ell} (D)_{\ell,k} \\ &= \sum_{\ell=0}^{n-1} \delta_{gj-\ell} (D)_{\ell,k} \\ &\underset{\text{(a)}}{=} (D)_{(gj) \bmod n,k} \\ &\underset{\text{(b)}}{=} \begin{cases} d_{(gj) \bmod n} & \text{if } k \equiv gj \pmod n, \\ 0 & \text{otherwise,} \end{cases} \end{aligned} \tag{5.35}$$

where (a) follows from the fact that there exists a unique $\ell \in \{0, 1, \ldots, n-1\}$ such that $gj - \ell \equiv 0 \pmod n$, that is $\ell \equiv gj \pmod n$; while in (b) simply recall that D is a diagonal matrix, then $(D)_{\ell,k} = 0$ if $\ell \neq k$.

For $j,k=0,\dots,n-1$, from (5.34) we have that

$$\begin{aligned}
\left(\widetilde{D}S_{n,g}\right)_{j,k} &= \sum_{\ell=0}^{n-1}\left(\widetilde{D}\right)_{j,\ell}(S_{n,g})_{\ell,k}\\
&= \sum_{\ell=0}^{n-1}\left(\widetilde{D}\right)_{j,\ell}\delta_{g\ell-k}\\
&\underset{\text{(a)}}{=} \left(\widetilde{D}\right)_{j,j}\delta_{gj-k}\\
&= \begin{cases} d_{(gj)\bmod n} & \text{if } gj-k\equiv 0 \pmod n,\\ 0 & \text{otherwise,}\end{cases}\\
&\underset{\text{(b)}}{=} \begin{cases} d_{(gj)\bmod n} & \text{if } k\equiv gj \pmod n,\\ 0 & \text{otherwise,}\end{cases}
\end{aligned} \tag{5.36}$$

where (a) follows from the fact that, since D is a diagonal matrix, $(D)_{j,\ell}=0$ if $\ell\neq j$, then we take $\ell=j$; while in (b) we observe that there exists a unique $k\in\{0,1,\dots,n-1\}$ such that $gj-k\equiv 0 \pmod n$, that is $k\equiv gj \pmod n$.

The two expression in (5.35) and (5.36) are equivalent, and the equality (5.33) is thus proved. □

We conclude this section with a result on the product of g-circulant matrices.

Lemma 5.11. *If $C_{n,g}\in M_n(\mathbb{C})$ is a g-circulant matrix and $C_{n,h}\in M_n(\mathbb{C})$ is a h-circulant matrix, then $C_{n,g}C_{n,h}\in M_n(\mathbb{C})$ is a gh-circulant matrix.*

Proof. From (5.8) and (5.11) we have that

$$\begin{aligned}
C_{n,g} &= F_n D_n^{(1)} F_n^* Z_{n,g},\\
C_{n,h} &= F_n D_n^{(2)} F_n^* Z_{n,h};
\end{aligned}$$

hence, using (5.26) and (5.33), we obtain

$$\begin{aligned}
C_{n,g}C_{n,h} &= F_n D_n^{(1)} F_n^* Z_{n,g} F_n D_n^{(2)} F_n^* Z_{n,h}\\
&= F_n D_n^{(1)} F_n^* F_n S_{n,g} F_n^* F_n D_n^{(2)} F_n^* F_n S_{n,h} F_n^*\\
&= F_n D_n^{(1)} S_{n,g} D_n^{(2)} S_{n,h} F_n^*\\
&= F_n D_n^{(1)} \widetilde{D}_n^{(2)} S_{n,g} S_{n,h} F_n^*,
\end{aligned} \tag{5.37}$$

where $\widetilde{D}_n^{(2)}$ is a diagonal matrix. We compute separately the product $S_{n,g}S_{n,h}$:

$$\begin{aligned}
(S_{n,g}S_{n,h})_{j,k} &= \sum_{\ell=0}^{n-1}(S_{n,g})_{j,\ell}(S_{n,h})_{\ell,k}\\
&= \sum_{\ell=0}^{n-1}\delta_{gj-\ell}\delta_{h\ell-k}\\
&\underset{\text{(a)}}{=} \delta_{h((gj)\bmod n)-k}\\
&\underset{\text{(b)}}{=} \delta_{((hgj)\bmod hn)-k}\\
&\underset{\text{(c)}}{=} \delta_{[((hgj)\bmod hn)-k]\bmod n}\\
&\underset{\text{(d)}}{=} \delta_{\{[((hgj)\bmod hn)-k]\bmod hn\}\bmod n}\\
&\underset{\text{(e)}}{=} \delta_{[(hgj-k)\bmod hn]\bmod n}\\
&\underset{\text{(f)}}{=} \delta_{(hgj-k)\bmod n}\\
&\underset{\text{(m)}}{=} \delta_{hgj-k},
\end{aligned} \tag{5.38}$$

where

(a) follows from the fact that there exists a unique $\ell \in \{0, 1, \ldots, n-1\}$ such that $gj - \ell \equiv 0 \pmod{n}$, that is, $\ell \equiv gj \pmod{n}$;

(b) is due to this property: if we have three integer numbers ρ, θ, and γ, then $\rho\,((\theta) \bmod \gamma) = (\rho\theta) \bmod \rho\gamma$;

(c) comes from the definition of δ in (5.9): $\delta_w = \delta_{(w) \bmod n}$;

(d) since $h \in \mathbb{N}^+$, hn is a multiple of n, then, given any number $w \in \mathbb{Z}$ is true that

$$(w) \bmod n = [(w) \bmod hn] \bmod n;$$

(e) just remember that

$$[((hgj) \bmod hn) - k] \bmod hn = (hgj - k) \bmod hn;$$

(f) is the same argument made in (d);

(m) see (c).

From (5.38) and (5.27) we have that, for $j, k = 0, \ldots, n-1$

$$\begin{aligned} \left(S_{n,g} S_{n,h}\right)_{j,k} &= \delta_{hgj-k} \\ &= \left(S_{n,gh}\right)_{j,k}, \end{aligned} \tag{5.39}$$

and since for Lemma 5.8 it holds

$$F_n S_{n,gh} F_n^* = Z_{n,gh}, \tag{5.40}$$

then, using (5.39) and (5.40), relation (5.37) becomes

$$\begin{aligned} C_{n,g} C_{n,h} &= F_n D_n^{(1)} \widetilde{D}_n^{(2)} S_{n,g} S_{n,h} F_n^* \\ &= F_n D_n^{(1)} \widetilde{D}_n^{(2)} S_{n,gh} F_n^* \\ &= F_n D_n^{(1)} \widetilde{D}_n^{(2)} F_n^* F_n S_{n,gh} F_n^* \\ &= F_n D_n^{(1)} \widetilde{D}_n^{(2)} F_n^* Z_{n,gh}, \end{aligned} \tag{5.41}$$

and since $D_n^{(1)} \widetilde{D}_n^{(2)}$ is a diagonal matrix, (5.41) is exactly the expression of a gh-circulant matrix. □

Remark 5.12. *We recall that for Lemma 5.2, $Z_{n,g} = Z_{n,(g) \bmod n}$, then Lemma 5.11 can also be stated by saying that $C_{n,g} C_{n,h}$ is a $\widehat{gh}$-circulant matrix, where $\widehat{gh} \equiv gh \pmod{n}$.*

Remark 5.13. *In Lemma 5.11, $g, h \in \{1, 2, \ldots, n\}$, i.e., instead of taking $g = 0$ and/or $h = 0$, we choose, under Remark 5.12, $g = n$ and/or $h = n$, this is solely due to the fact that in some passages in the proof (see (5.38)) is required the rest of the division for hn, and if $h = 0$, this term does not make sense.*

5.2 Singular values of g-circulant matrices

Now we link the singular values of g-circulant matrices with the eigenvalues of its circulant counterpart C_n. This is non-trivial given the multiplicative relation $C_{n,g} = C_n Z_{n,g}$.

Theorem 5.14. *Let $C_{n,g} \in M_n(\mathbb{C})$ be a g-circulant matrix and let $C_n = F_n D_n F_n^*$ be the circulant matrix generated by the same elements (the two matrices $C_{n,g}$ and C_n have the same first column); then the singular values of $C_{n,g}$ are given by*

$$\begin{aligned} \sigma_j(C_{n,g}) &= \sqrt{\sum_{l=1}^{(n,g)} d_{(l-1)n_g+j}}, \quad j = 0, 1, \ldots, n_g - 1, \\ \sigma_j(C_{n,g}) &= 0, \qquad j = n_g, \ldots, n-1, \end{aligned} \tag{5.42}$$

where the values d_k, $k = 0, \ldots, n-1$, are the diagonal elements of $D_n^* D_n$.

Proof. Having in mind the definition of the diagonal matrix D_n given in (5.5), we start by setting

$$D_n^* D_n = \underset{s=0,\ldots,n-1}{\operatorname{diag}} \left(|D_n|_{s,s}^2 \right) = \underset{s=0,\ldots,n-1}{\operatorname{diag}} (d_s) = \bigoplus_{l=1}^{(n,g)} \Delta_l,$$

$$J_{(n,g)} \otimes I_{n_g} = \underbrace{[I_{n,g} | I_{n,g} | \cdots | I_{n,g}]}_{(n,g)\ times} = \left.\begin{bmatrix} I_{n_g} & I_{n_g} & \cdots & I_{n_g} \\ I_{n_g} & I_{n_g} & \cdots & I_{n_g} \\ \vdots & \vdots & \vdots & \vdots \\ I_{n_g} & I_{n_g} & \cdots & I_{n_g} \end{bmatrix}\right\} (n,g) \text{ times}, \tag{5.43}$$

where

$$d_s = |D_n|_{s,s}^2 = (D_n)_{s,s} \cdot \overline{(D_n)_{s,s}}, \quad s = 0, 1, \ldots, n-1, \tag{5.44}$$

$$\Delta_l = \begin{bmatrix} d_{(l-1)n_g} & & & \\ & d_{(l-1)n_g+1} & & \\ & & \ddots & \\ & & & d_{(l-1)n_g+n_g-1} \end{bmatrix} \in M_{n_g}(\mathbb{R});\ l = 1, 2, \ldots, (n,g),$$

$$J_{(n,g)} = \left.\begin{bmatrix} 1 & 1 & \cdots & 1 \\ 1 & 1 & \cdots & 1 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & \cdots & 1 \end{bmatrix}\right\} (n,g) \text{ times}. \tag{5.45}$$

We now exploit relation (5.12) and Lemma 5.4, and we obtain that

$$\begin{aligned} F_n Z_{n,g} &= F_n \left[\widetilde{Z}_{n,g} | \widetilde{Z}_{n,g} | \cdots | \widetilde{Z}_{n,g} \right] \\ &= \left[F_n \widetilde{Z}_{n,g} | F_n \widetilde{Z}_{n,g} | \cdots | F_n \widetilde{Z}_{n,g} \right] \\ &= \frac{1}{\sqrt{(n,g)}} \left[I_{n,g} F_{n_g} Z_{n_g,\breve{g}} | I_{n,g} F_{n_g} Z_{n_g,\breve{g}} | \cdots | I_{n,g} F_{n_g} Z_{n_g,\breve{g}} \right] \\ &= \frac{1}{\sqrt{(n,g)}} [I_{n,g} | I_{n,g} | \cdots | I_{n,g}] \underbrace{\begin{bmatrix} F_{n_g} Z_{n_g,\breve{g}} & & & \\ & F_{n_g} Z_{n_g,\breve{g}} & & \\ & & \ddots & \\ & & & F_{n_g} Z_{n_g,\breve{g}} \end{bmatrix}}_{(n,g) \text{ times}} \\ &= \frac{1}{\sqrt{(n,g)}} [I_{n,g} | I_{n,g} | \cdots | I_{n,g}] \left(I_{(n,g)} \otimes F_{n_g} Z_{n_g,\breve{g}} \right), \end{aligned} \tag{5.46}$$

where $I_{(n,g)} \in M_{(n,g)}(\mathbb{R})$ is the identity matrix. Furthermore,

$$\begin{aligned} C_{n,g}^* C_{n,g} &= (F_n D_n F_n^* Z_{n,g})^* (F_n D_n F_n^* Z_{n,g}) \\ &= Z_{n,g}^\top F_n D_n^* F_n^* F_n D_n F_n^* Z_{n,g} \\ &= Z_{n,g}^\top F_n D_n^* D_n F_n^* Z_{n,g} \\ &= (F_n^* Z_{n,g})^* D_n^* D_n F_n^* Z_{n,g}. \end{aligned} \tag{5.47}$$

From (5.46) and (5.43), we plainly infer the following relations:

$$\begin{aligned} (F_n^* Z_{n,g})^* &= \left(\frac{1}{\sqrt{(n,g)}} [I_{n,g}|I_{n,g}|\cdots|I_{n,g}] \left(I_{(n,g)} \otimes F_{n_g}^* Z_{n_g,\check{g}} \right) \right)^* \\ &= \frac{1}{\sqrt{(n,g)}} \left(I_{(n,g)} \otimes F_{n_g}^* Z_{n_g,\check{g}} \right)^* \left(J_{(n,g)} \otimes I_{n_g} \right) \\ &= \frac{1}{\sqrt{(n,g)}} \left(I_{(n,g)} \otimes Z_{n_g,\check{g}}^\top F_{n_g} \right) \left(J_{(n,g)} \otimes I_{n_g} \right), \\ F_n^* Z_{n,g} &= \frac{1}{\sqrt{(n,g)}} [I_{n,g}|I_{n,g}|\cdots|I_{n,g}] \left(I_{(n,g)} \otimes F_{n_g}^* Z_{n_g,\check{g}} \right) \\ &= \frac{1}{\sqrt{(n,g)}} \left(J_{(n,g)} \otimes I_{n_g} \right) \left(I_{(n,g)} \otimes F_{n_g}^* Z_{n_g,\check{g}} \right). \end{aligned}$$

Hence

$$\begin{aligned} &C_{n,g}^* C_{n,g} \\ &= \left(I_{(n,g)} \otimes Z_{n_g,\check{g}}^\top F_{n_g} \right) \left(J_{(n,g)} \otimes I_{n_g} \right) \frac{1}{(n,g)} D_n^* D_n \left(J_{(n,g)} \otimes I_{n_g} \right) \left(I_{(n,g)} \otimes F_{n_g}^* Z_{n_g,\check{g}} \right). \end{aligned}$$

Now using the properties of the tensorial product

$$\begin{aligned} \left(I_{(n,g)} \otimes Z_{n_g,\check{g}}^\top F_{n_g} \right) \left(I_{(n,g)} \otimes F_{n_g}^* Z_{n_g,\check{g}} \right) &= I_{(n,g)} I_{(n,g)} \otimes Z_{n_g,\check{g}}^\top F_{n_g} F_{n_g}^* Z_{n_g,\check{g}} \\ &= I_{(n,g)} I_{(n,g)} \otimes Z_{n_g,\check{g}}^\top Z_{n_g,\check{g}} \\ &= I_{(n,g)} I_{(n,g)} \otimes I_{n_g} = I_n, \end{aligned}$$

and from a similarity argument, one deduces that the eigenvalues of $C_{n,g}^* C_{n,g}$ are the eigenvalues

of the matrix

$$
\begin{aligned}
&\left(J_{(n,g)}\otimes I_{n_g}\right)\frac{1}{(n,g)}D_n^*D_n\left(J_{(n,g)}\otimes I_{n_g}\right)\\
&= \frac{1}{(n,g)}\left[\begin{array}{c|c|c|c} I_{n_g} & I_{n_g} & \cdots & I_{n_g}\\ \hline I_{n_g} & I_{n_g} & \cdots & I_{n_g}\\ \hline \vdots & \vdots & \vdots & \vdots\\ \hline I_{n_g} & I_{n_g} & \cdots & I_{n_g}\end{array}\right]\left[\begin{array}{cccc} \Delta_1 & & & \\ & \Delta_2 & & \\ & & \ddots & \\ & & & \Delta_{(n,g)}\end{array}\right]\left[\begin{array}{c|c|c|c} I_{n_g} & I_{n_g} & \cdots & I_{n_g}\\ \hline I_{n_g} & I_{n_g} & \cdots & I_{n_g}\\ \hline \vdots & \vdots & \vdots & \vdots\\ \hline I_{n_g} & I_{n_g} & \cdots & I_{n_g}\end{array}\right]\\
&= \frac{1}{(n,g)}\left[\begin{array}{c|c|c|c} I_{n_g} & I_{n_g} & \cdots & I_{n_g}\\ \hline I_{n_g} & I_{n_g} & \cdots & I_{n_g}\\ \hline \vdots & \vdots & \vdots & \vdots\\ \hline I_{n_g} & I_{n_g} & \cdots & I_{n_g}\end{array}\right]\left[\begin{array}{c|c|c|c} \Delta_1 & \Delta_1 & \cdots & \Delta_1\\ \hline \Delta_2 & \Delta_2 & \cdots & \Delta_2\\ \hline \vdots & \vdots & \vdots & \vdots\\ \hline \Delta_{(n,g)} & \Delta_{(n,g)} & \cdots & \Delta_{(n,g)}\end{array}\right]\\
&= \frac{1}{(n,g)}\left[\begin{array}{c|c|c|c} \sum_{l=1}^{(n,g)}\Delta_l & \sum_{l=1}^{(n,g)}\Delta_l & \cdots & \sum_{l=1}^{(n,g)}\Delta_l\\ \hline \sum_{l=1}^{(n,g)}\Delta_l & \sum_{l=1}^{(n,g)}\Delta_l & \cdots & \sum_{l=1}^{(n,g)}\Delta_l\\ \hline \vdots & \vdots & \vdots & \vdots\\ \hline \sum_{l=1}^{(n,g)}\Delta_l & \sum_{l=1}^{(n,g)}\Delta_l & \cdots & \sum_{l=1}^{(n,g)}\Delta_l\end{array}\right]\\
&= \frac{1}{(n,g)}\underbrace{\left[\begin{array}{cccc} 1 & 1 & \cdots & 1\\ 1 & 1 & \cdots & 1\\ \vdots & \vdots & \vdots & \vdots\\ 1 & 1 & \cdots & 1\end{array}\right]}_{(n,g)\text{ times}}\otimes\left(\sum_{l=1}^{(n,g)}\Delta_l\right).
\end{aligned}
$$

Therefore, from (5.45), we infer that

$$\Lambda\left(C_{n,g}^*C_{n,g}\right)=\frac{1}{(n,g)}\Lambda\left(J_{(n,g)}\otimes\sum_{l=1}^{(n,g)}\Delta_l\right),\tag{5.48}$$

where

$$\frac{1}{(n,g)}\Lambda\left(J_{(n,g)}\right)=\{0,1\}.\tag{5.49}$$

Here we must observe that $\frac{1}{(n,g)}J_{(n,g)}$ is a matrix of rank 1 with trace $(n,g)\cdot\frac{1}{(n,g)}=1$, so it has all eigenvalues equal to zero except one eigenvalue equal to 1. Moreover,

$$
\begin{aligned}
\sum_{l=1}^{(n,g)}\Delta_l &= \sum_{l=1}^{(n,g)}\operatorname*{diag}_{j=0,\ldots,n_g-1}\left(d_{(l-1)n_g+j}\right)\\
&= \operatorname*{diag}_{j=0,\ldots,n_g-1}\left(\sum_{l=1}^{(n,g)}d_{(l-1)n_g+j}\right).
\end{aligned}
$$

Consequently, since $\sum_{l=1}^{(n,g)}\Delta_l$ is a diagonal matrix, we have

$$\Lambda\left(\sum_{l=1}^{(n,g)}\Delta_l\right)=\left\{\sum_{l=1}^{(n,g)}d_{(l-1)n_g+j};\ j=0,1,\ldots,n_g-1\right\},\tag{5.50}$$

where d_k are defined in (5.44).

Finally, by exploiting basic properties of the tensor product, we know that the eigenvalues of a tensor product of two square matrices $A \otimes B$ are given by all possible products of eigenvalues of A of order p and of eigenvalues of B of order q, that is, $\lambda(A \otimes B) = \lambda_j(A)\lambda_k(B)$ for $j = 1, \ldots, p$ and $k = 1, \ldots, q$. Therefore, by taking into consideration (5.48), (5.49), and (5.50), we find

$$\lambda_j\left(C_{n,g}^* C_{n,g}\right) = \sum_{l=1}^{(n,g)} d_{(l-1)n_g+j}, \quad j = 0, 1, \ldots, n_g - 1, \tag{5.51}$$

$$\lambda_j\left(C_{n,g}^* C_{n,g}\right) = 0, \qquad j = n_g, \ldots, n-1. \tag{5.52}$$

From (5.51), (5.52), and (1.5), one obtains that the singular values of a g-circulant matrix $C_{n,g}$ are given by

$$\begin{aligned} \sigma_j(C_{n,g}) &= \sqrt{\sum_{l=1}^{(n,g)} d_{(l-1)n_g+j}}, \quad j = 0, 1, \ldots, n_g - 1, \\ \sigma_j(C_{n,g}) &= 0, \qquad j = n_g, \ldots, n-1, \end{aligned}$$

where the values d_k, $k = 0, \ldots, n-1$, are defined in (5.44). □

5.2.1 Special cases and observations

In this subsection we consider some special cases and furnish a further link between the eigenvalues of circulant matrices and the singular values of g-circulants.

Case $g = 0$.

If $g = 0$, from (5.7) we have that, for $j, k = 0, \ldots, n-1$,

$$(C_{n,g})_{j,k} = (C_{n,0})_{j,k} = a_{(j-0\cdot k) \bmod n} = a_j.$$

This means that $C_{n,0}$ is a matrix that has constant elements along all the rows and, therefore, it has rank 1. Then the matrix $C_{n,0}$ has only one singular value different from zero, and using the formula (5.42), since $(n, g) = n$ and $n_g = \frac{n}{(n,g)} = 1$, we get

$$\begin{aligned} \sigma_0(C_{n,g}) &= \sqrt{\sum_{l=0}^{n-1} d_l}, \\ \sigma_j(C_{n,g}) &= 0, \qquad j = 1, \ldots, n-1. \end{aligned}$$

Case $(n, g) = 1$.

In the case where $(n, g) = 1$ (for example when $g = 1$) we have $n_g = \frac{n}{(n,g)} = n$. Hence the formula (5.42) becomes

$$\sigma_j(C_{n,g}) = \sqrt{d_j}, \quad j = 0, 1, \ldots, n-1.$$

In other words the singular values of $C_{n,g}$ coincide with those of C_n (this is expected since $Z_{n,g}$ is a permutation matrix) and, in particular, with the moduli of the eigenvalues of C_n.

Distribution in the singular value sense for g-circulant matrices

In Section 5.1 we have seen that the eigenvalues of a circulant matrix $C_n(p)$ generated by a polynomial p are given by

$$\lambda_j\left(C_n(p)\right) = p\left(\frac{2\pi j}{n}\right), \qquad j = 0, \ldots, n-1.$$

The question that naturally arises is how to connect the expression in (5.42) of the non-trivial singular values of $C_{n,g}(p)$ (the g-circulant matrix generated by p) with the polynomial p. The answer is somehow intriguing and can be resumed in the following formula which could be of interest in the multigrid community (see Chapter 7):

$$\sigma_j\left(C_{n,g}(p)\right) = \sqrt{\sum_{l=0}^{(n,g)-1} |p|^2\left(\frac{x_j + 2\pi l}{(n,g)}\right)}, \quad x_j = \frac{2\pi j}{n_g}, \qquad j = 0, 1, \ldots, n_g - 1. \tag{5.53}$$

If g is fixed the sequence n is chosen so that $\gamma = (n,g)$ is a fixed number, by Definition 1.7 and using (5.53) we have

$$\begin{aligned}
\lim_{n\to\infty} \frac{1}{n}\sum_{j=1}^{n} F\left(\sigma_j\left(C_{n,g}(p)\right)\right) &= \lim_{n\to\infty} \frac{1}{n}\sum_{j=0}^{n_g-1} F\left(\sigma_j\left(C_{n,g}(p)\right)\right) + \lim_{n\to\infty} \frac{1}{n}\sum_{j=n_g}^{n-1} F(0) \\
&= \lim_{n\to\infty} \frac{n_g}{n}\sum_{j=0}^{n_g-1} \frac{F\left(\sigma_j\left(C_{n,g}(p)\right)\right)}{n_g} + \lim_{n\to\infty} \frac{n-n_g}{n} F(0) \\
&= \frac{1}{\gamma}\frac{1}{2\pi}\int_Q F\left(\sqrt{\sum_{l=0}^{\gamma-1} |p|^2\left(\frac{x+2\pi l}{\gamma}\right)}\right) \mathrm{d}x + \left(1 - \frac{1}{\gamma}\right) F(0),
\end{aligned}$$

which results to being equivalent to the following distribution formula:

$$\{C_{n,g}(p)\} \sim_\sigma \left(\eta_p, Q \times [0,1]\right),$$

where

$$\eta_p(x,t) = \begin{cases} \sqrt{\widehat{|p|^{(2)}}(x)} & \text{for } t \in \left[0, \frac{1}{\gamma}\right], \\ 0 & \text{for } t \in \left(\frac{1}{\gamma}, 1\right], \end{cases}$$

with

$$\widehat{|p|^{(2)}}(x) = \sum_{j=0}^{\gamma-1} |p|^2\left(\frac{x+2\pi j}{\gamma}\right).$$

If $g = 1$ that is we are in standard circulant context, then $C_{n,g}(p) = C_n(p)$, $\sqrt{\widehat{|p|^{(2)}}(x)}$ reduces to $|p(x)|$, and the variable $t \in [0,1]$ becomes useless so that

$$\{C_n(p)\} \sim_\sigma \left(p, Q \times [0,1]\right),$$

which is the same as the classical result

$$\{C_n(p)\} \sim_\sigma (p, Q).$$

If $g = 0$, since from Subsection 5.2.1 most of the singular values are identically zero, we infer that

$$\{C_{n,0}\} \sim_\sigma (0, Q).$$

In addition if g is fixed and a sequence of integers n is chosen so that $(n, g) > 1$ for n large enough, then

$$\{C_{n,g}\} \sim_\sigma (0, G),$$

for a suitable set G.

From the above reasoning it is clear that if n is allowed to vary among all the positive integer numbers, then $\{C_{n,g}\}$ does not possess a joint singular value distribution.

5.3 Eigenvalues of g-circulant matrices

In this section we show how to calculate the eigenvalues of a g-circulant matrix; first we consider two special cases: $g = 0$ and $g = 1$, then we will give an explicit formula for the eigenvalues of $C_{n,g}$ where $(n, g) = 1$ and then, using Theorem 1.4, we will see a recursive way to calculate the eigenvalues of a g-circulant matrix when $(n, g) \neq 1$.

5.3.1 Case $g = 1$.

If $g = 1$, then $C_{n,g} = C_{n,1} = C_n$ is the "classical" circulant matrix and the eigenvalues are given by the formula (5.5), i.e.

$$\lambda_j (C_n) = \sum_{k=0}^{n-1} e^{\frac{2\pi i j k}{n}} a_k, \qquad j = 0, \ldots, n-1.$$

5.3.2 Case $g = 0$.

If $g = 0$, from Subsection 5.2.1, $C_{n,g}$ has rank 1; then, remembering that the trace ($\mathrm{tr}(\cdot)$) of a matrix is the sum of its eigenvalues, we can conclude that $C_{n,0}$ has $n - 1$ zero eigenvalues and one eigenvalue λ different from zero given by

$$\lambda = \mathrm{tr}\,(C_{n,0}) = \sum_{r=0}^{n-1} (C_{n,0})_{r,r} = \sum_{r=0}^{n-1} a_r.$$

5.3.3 Case $(n, g) = 1$ and $g \notin \{0, 1\}$.

If n and g are coprime the following lemma gives us a direct formula for calculating the eigenvalues of a g-circulant matrix $C_{n,g}$.

Lemma 5.15. *Let $C_{n,g} \in M_n(\mathbb{C})$ be a g-circulant matrix such that $(n, g) = 1$. So if*

$$C_{n,g} = F_n D_n F_n^* Z_{n,g},$$

with

$$D_n = \operatorname*{diag}_{j=0,\ldots,n-1} (d_j),$$

the eigenvalues of $C_{n,g}$ are given by

$$|\lambda_j (C_{n,g})| = \left| \sqrt[s]{\prod_{k=0}^{s-1} d_{(g^k j) \bmod n}} \right|, \qquad j = 0, \ldots, n-1,$$

where $s \in \mathbb{N}^+$ is such that $g^s \equiv 1$ (mod n).

Proof. From Lemma 5.11 it holds that, if $C_{n,g}$ is a g-circulant matrix, then $C_{n,g}^2$ is a g^2-circulant matrix and, more generally, $C_{n,g}^r$ is a g^r-circulant matrix ($r \in \mathbb{N}^+$) or, equivalently, for Remark 5.12, $C_{n,g}^r$ is a $\widehat{g^r}$-circulant matrix, with $\widehat{g^r} \equiv g^r \pmod n$; this means that if s is such that $g^s \pmod n \equiv 1$, then $C_{n,g}^s$ is a circulant matrix and we are able to calculate the eigenvalues of this matrix (see Subsection 5.3.1) and, consequently, the modulo of the eigenvalues of $C_{n,g}$ are the modulo of the roots of index s of the eigenvalues of the circulant matrix $C_{n,g}^s$.

So we calculate the eigenvalues of $C_{n,g}^s$. From (5.11) and (5.26) we have that

$$\begin{aligned} C_{n,g}^s &= (F_n D_n F_n^* Z_{n,g})^s \\ &= (F_n D_n F_n^* F_n S_{n,g} F_n^*)^s \\ &= (F_n D_n S_{n,g} F_n^*)^s \\ &= F_n \left(D_n S_{n,g}\right)^s F_n^*, \end{aligned}$$

then, since $F_n^* F_n = I_n$, we obtain

$$\Lambda\left(C_{n,g}^s\right) = \Lambda\left(\left(D_n S_{n,g}\right)^s\right). \tag{5.54}$$

Before continuing the proof, we introduce the following symbology:

$$D_{n,g^r} = \underset{j=0,\ldots,n-1}{\operatorname{diag}} \left(d_{(g^r j) \bmod n}\right); \tag{5.55}$$

from this we have that $D_n = D_{n,g^0}$. We work now on the matrix $(D_n S_{n,g})^s$ and, repeatedly using the formulae (5.34) and (5.55) we get:

$$\begin{aligned} \left(D_n S_{n,g}\right)^s &= D_n \left(S_{n,g} D_n\right)^{s-1} S_{n,g} \\ &= D_{n,g^0} \left(\underset{j=0,\ldots,n-1}{\operatorname{diag}} \left(d_{(gj) \bmod n}\right) S_{n,g}\right)^{s-1} S_{n,g} \\ &= D_{n,g^0} \underset{j=0,\ldots,n-1}{\operatorname{diag}} \left(d_{(gj) \bmod n}\right) \left(S_{n,g} \underset{j=0,\ldots,n-1}{\operatorname{diag}} \left(d_{(gj) \bmod n}\right)\right)^{s-2} S_{n,g} S_{n,g} \\ &= D_{n,g^0} D_{n,g^1} \left(S_{n,g} \underset{j=0,\ldots,n-1}{\operatorname{diag}} \left(d_{(gj) \bmod n}\right)\right)^{s-2} S_{n,g}^2 \\ &= D_{n,g^0} D_{n,g^1} \left(\underset{j=0,\ldots,n-1}{\operatorname{diag}} \left(d_{[g((gj) \bmod n)] \bmod n}\right) S_{n,g}\right)^{s-2} S_{n,g}^2 \\ &\underset{\text{(a)}}{=} D_{n,g^0} D_{n,g^1} \left(\underset{j=0,\ldots,n-1}{\operatorname{diag}} \left(d_{(g^2 j) \bmod n}\right) S_{n,g}\right)^{s-2} S_{n,g}^2 \\ &= D_{n,g^0} D_{n,g^1} \underset{j=0,\ldots,n-1}{\operatorname{diag}} \left(d_{(g^2 j) \bmod n}\right) \left(S_{n,g} \underset{j=0,\ldots,n-1}{\operatorname{diag}} \left(d_{(g^2 j) \bmod n}\right)\right)^{s-3} S_{n,g} S_{n,g}^2 \\ &= D_{n,g^0} D_{n,g^1} D_{n,g^2} \left(S_{n,g} \underset{j=0,\ldots,n-1}{\operatorname{diag}} \left(d_{(g^2 j) \bmod n}\right)\right)^{s-3} S_{n,g}^3 \\ &= \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ &= D_{n,g^0} D_{n,g^1} D_{n,g^2} \cdots D_{n,g^{s-1}} S_{n,g}^s, \end{aligned}$$

where (a) is due to this property: if we have three integer numbers ρ, θ, and γ, then

$$\rho\left((\theta) \bmod \gamma\right) = (\rho\theta) \bmod \rho\gamma,$$

so

$$g\left((gj) \bmod n\right) = (ggj) \bmod gn = \left(g^2 j\right) \bmod gn,$$

moreover, since $g \in \mathbb{N}^+$, gn is a multiple of n, we have that

$$\left[\left(g^2 j\right) \bmod gn\right] \bmod n = \left(g^2 j\right) \bmod n.$$

Now, since $g^s \pmod n \equiv 1$, by (5.39) and by Lemma 5.9 it holds that

$$\begin{aligned} S_{n,g}^s &= \underbrace{S_{n,g} S_{n,g} \cdots S_{n,g}}_{s \text{ times}} \\ &= S_{n,g^s} \\ &= S_{n,(g^s) \bmod n} = S_{n,1} = [\delta_{j-k}]_{j,k=0}^{n-1} = I_n, \end{aligned}$$

so $(D_n S_{n,g})^s$ is a diagonal matrix and its eigenvalues are given by the diagonals elements of $D_{n,g^0} D_{n,g^1} D_{n,g^2} \cdots D_{n,g^{s-1}}$, i.e., from (5.55),

$$\begin{aligned} \lambda_j \left((D_n S_{n,g})^s\right) &= d_{(g^0 j) \bmod n} d_{(g^1 j) \bmod n} d_{(g^2 j) \bmod n} \cdots d_{(g^{s-1} j) \bmod n} \\ &= \prod_{k=0}^{s-1} d_{\left(g^k j\right) \bmod n}, \qquad j = 0, \ldots, n-1. \end{aligned} \tag{5.56}$$

Finally, from (5.54) and from the fact that, as mentioned above, the modulo of the eigenvalues of $C_{n,g}$ are the modulo of the roots of index s of the eigenvalues of the circulant matrix $C_{n,g}^s$, by (5.56) it follows that

$$\left|\lambda_j \left(C_{n,g}\right)\right| = \left| \sqrt[s]{\prod_{k=0}^{s-1} d_{\left(g^k j\right) \bmod n}} \right|, \qquad j = 0, \ldots, n-1.$$

□

Remark 5.16. *In Lemma 5.15, the existence of a number $s \in \mathbb{N}^+$ such that $g^s \equiv 1$ (mod n) is guaranteed by Euler's Theorem (or Fermat-Euler Theorem), which states that if n is a positive integer and g is a positive integer coprime to n, i.e. $(n, g) = 1$ as in the hypothesis of Lemma 5.15, then*

$$g^{\varphi(n)} \equiv 1 \ (mod\ n),$$

where $\varphi(n)$ is the Euler function defined in this way: if n can be factored as

$$n = p_1^{k_1} p_2^{k_2} \cdots p_r^{k_r}, \qquad p_1, p_2, \ldots p_r \text{ prime numbers,}$$

then

$$\varphi(n) = n \left[\left(1 - \frac{1}{p_1}\right)\left(1 - \frac{1}{p_2}\right) \cdots \left(1 - \frac{1}{p_r}\right)\right].$$

5.3.4 Case $(n, g) \neq 1$ and $g \notin \{0, 1\}$.

In the case where n and g are not coprime, Lemma 5.15 is no longer valid, then we proceed in another way. The idea is to exploit the fact that if $(n, g) \neq 1$, then the matrix $C_{n,g}$ is singular, and then apply Theorem 1.4.

By Lemma 5.3 we have that

$$\begin{aligned} Z_{n,g} &= \underbrace{\left[\widetilde{Z}_{n,g} | \widetilde{Z}_{n,g} | \cdots | \widetilde{Z}_{n,g}\right]}_{(n,g) \text{ times}} \\ &= \widetilde{Z}_{n,g} \underbrace{\left[\, I_{n_g} \mid I_{n_g} \mid \cdots \mid I_{n_g} \,\right]}_{(n,g) \text{ times}}, \end{aligned}$$

where $\widetilde{Z}_{n,g} \in M_{n,n_g}(\mathbb{R})$ and $I_{n_g} \in M_{n_g}(\mathbb{R})$ is the identity matrix; now, is immediate to verify that we can rewrite (5.8) as

$$\begin{aligned} C_{n,g} &= C_n Z_{n,g} \\ &= C_n \widetilde{Z}_{n,g} \underbrace{[I_{n_g}|I_{n_g}|\cdots|I_{n_g}]}_{(n,g) \text{ times}} \\ &= \widetilde{C}_{n,g} I_{n,n_g}, \end{aligned} \tag{5.57}$$

with $\widetilde{C}_{n,g} = C_n \widetilde{Z}_{n,g} \in M_{n,n_g}(\mathbb{C})$, $I_{n,n_g} \in M_{n_g,n}(\mathbb{R})$ and, for $j = 0,\ldots,n$, $k = 0,\ldots,n_g$ it holds $\left(\widetilde{C}_{n,g}\right)_{j,k} = (C_{n,g})_{j,k}$.

Using Theorem 1.4, we have that the eigenvalues of $C_{n,g}$ are the same as those of $I_{n,n_g}\widetilde{C}_{n,g} \in M_{n_g}(\mathbb{C})$, plus $n - n_g$ null eigenvalues:

$$\Lambda(C_{n,g}) = \Lambda\left(I_{n,n_g}\widetilde{C}_{n,g}\right) \cup \{0 \text{ with geometric multiplicity } n - n_g\}.$$

Theorem 5.17. *Let $C_{n,g} = \widetilde{C}_{n,g} I_{n,n_g} = \left[a_{(j-gk) \bmod n}\right]_{j,k=0}^{n-1} \in M_n(\mathbb{C})$ be the g-circulant matrix defined in (5.57); then the matrix $\widehat{C}_{n_g,\widehat{g}} = I_{n,n_g}\widetilde{C}_{n,g} \in M_{n_g}(\mathbb{C})$ is a $\widehat{g}$-circulant matrix of dimension $n_g = \frac{n}{(n,g)}$, with $\widehat{g} \equiv g \pmod{n_g}$, whose elements are given by*

$$\left(\widehat{C}_{n_g,\widehat{g}}\right)_{j,k} = \sum_{t=0}^{(n,g)-1} a_{(j+tn_g-gk) \bmod n}, \qquad j,k = 0,\ldots,n_g - 1. \tag{5.58}$$

Proof. For $j,k = 0,\ldots,n_g - 1$ we have that

$$\left(\widehat{C}_{n_g,\widehat{g}}\right)_{j,k} = \sum_{\ell=0}^{n-1} \left(I_{n,n_g}\right)_{j,\ell} \left(\widetilde{C}_{n,g}\right)_{\ell,k}, \tag{5.59}$$

where, for $r = 0,\ldots,n_g$ and $s = 0,\ldots,n$,

$$\left(I_{n,n_g}\right)_{r,s} = \begin{cases} 1 & \text{if } s - r \equiv 0 \pmod{n_g}, \\ 0 & \text{otherwise,} \end{cases}$$

then, since $s = 0,\ldots,n-1$, $r = 0,\ldots,n_g$ and $\frac{n}{n_g} = (n,g)$, there are precisely (n,g) values of s such that

$$s - r \equiv 0 \pmod{n_g},$$

that is

$$s_t = r + tn_g, \qquad t = 0,\ldots,(n,g) - 1.$$

Now, in (5.59), since $\left(I_{n,n_g}\right)_{j,\ell} = 1$ if and only if $\ell = j + tn_g$ for $t = 0,\ldots,(n,g)-1$, using (5.57) and (5.10), we obtain

$$\begin{aligned} \left(\widehat{C}_{n_g,\widehat{g}}\right)_{j,k} &= \sum_{\ell=0}^{n-1} \left(I_{n,n_g}\right)_{j,\ell} \left(\widetilde{C}_{n,g}\right)_{\ell,k} \\ &= \sum_{t=0}^{(n,g)-1} \left(\widetilde{C}_{n,g}\right)_{j+tn_g,k} \\ &= \sum_{t=0}^{(n,g)-1} (C_{n,g})_{j+tn_g,k} \\ &= \sum_{t=0}^{(n,g)-1} (C_n)_{j+tn_g,gk} \\ &= \sum_{t=0}^{(n,g)-1} a_{(j+tn_g-gk) \bmod n}, \end{aligned} \tag{5.60}$$

and this proves (5.58); remains to prove that $\widehat{C}_{n_g,\widehat{g}}$ is a $\widehat{g}$-circulant matrix. The first column of the matrix $\widehat{C}_{n_g,\widehat{g}}$ is formed by the elements

$$\begin{aligned}\left(\widehat{C}_{n_g,\widehat{g}}\right)_{j,0} &= \sum_{t=0}^{(n,g)-1} a_{(j+tn_g-g\cdot 0)\bmod n} \\ &= \sum_{t=0}^{(n,g)-1} a_{(j+tn_g)\bmod n} \\ &= \widehat{a}_{(j-g\cdot 0)\bmod n_g} \\ &= \widehat{a}_{(j)\bmod n_g} \\ &= \widehat{a}_j \qquad j=0,\ldots,n_g-1; \end{aligned} \tag{5.61}$$

if we denote by C_{n_g} the classical circulant matrix of dimension n_g whose first column is given by $[\widehat{a}_0,\widehat{a}_1,\ldots,\widehat{a}_{n_g-1}]^\top$ ($\widehat{a}_j$ defined as in (5.61)), and if $Z_{n_g,g}$ is the matrix defined in (5.9) (with n_g instead of n), using (5.60) we have that

$$\begin{aligned}\left(C_{n_g}Z_{n_g,g}\right)_{j,k} &= \sum_{\ell=0}^{n_g-1} \left(C_{n_g}\right)_{j,\ell}\left(Z_{n_g,g}\right)_{\ell,k} \\ &\underset{\text{(a)}}{=} \widehat{a}_{(j-\ell)\bmod n_g}\delta_{\ell-gk} \\ &\underset{\text{(b)}}{=} \widehat{a}_{(j-gk)\bmod n_g} \\ &= \sum_{t=0}^{(n,g)-1} a_{(j+tn_g-gk)\bmod n} \\ &= \left(\widehat{C}_{n_g,\widehat{g}}\right)_{j,k}; \end{aligned}$$

where

(a) remember that, for $j,k=0,\ldots,n_g-1$, $\left(C_{n_g}\right)_{j,k}=\widehat{a}_{(j-k)\bmod n_g}$;

(b) follows from the fact that there is a unique $\ell\in\{0,1,\ldots,n_g-1\}$ such that $\ell-gk\equiv 0 \pmod{n_g}$, that is $\ell\equiv gk \pmod{n_g}$, so

$$(j-\ell)\bmod n_g=(j-(gk)\bmod n_g)\bmod n_g=(j-gk)\bmod n_g.$$

In conclusion we can write

$$\widehat{C}_{n_g,\widehat{g}}=C_{n_g}Z_{n_g,g},$$

and, by Lemma 5.2, since $Z_{n_g,g}=Z_{n_g,\widehat{g}}$ with $\widehat{g}\equiv g \pmod{n_g}$, we have that

$$\widehat{C}_{n_g,\widehat{g}}=C_{n_g}Z_{n_g,\widehat{g}},$$

then $\widehat{C}_{n_g,\widehat{g}}$ is a $\widehat{g}$-circulant matrix. □

The result obtained above is useful because we can reduce the problem of computing the eigenvalues of a g-circulant matrix of dimension $n\times n$, to the calculation of eigenvalues of a smaller $\widehat{g}$-circulant matrix of size $n_g\times n_g$ with $n_g=\frac{n}{(n,g)}$.

So, if we have a g-circulant matrix $C_{n,g}=\left[a_{(r-gs)\bmod n}\right]_{r,s=0}^{n-1}$, whose first column is given by $[a_0,a_1,\ldots,a_{n-1}]^\top$, we can calculate the eigenvalues by following these steps:

I step: if $g=0$ see Subsection 5.3.2;

II step: if $g = 1$, see Subsection 5.3.1;

III step: if $(n, g) = 1$ with $g \notin \{0, 1\}$, see Subsection 5.3.3;

IV step: if $g \notin 0, 1$ and $(n, g) \neq 1$, from Theorem 5.17 we have that $C_{n,g}$ has $n - n_g$ eigenvalues equal to zero and the remaining $n_g = \frac{n}{(n,g)}$ are the eigenvalues of $\widehat{C}_{n_g,\widehat{g}}$ (defined in (5.58)); then we set

$$\begin{aligned} a_j &= \sum_{t=0}^{(n,g)-1} a_{(j+tn_g) \bmod n}, \qquad j = 0, \ldots, n_g - 1, \\ g &= \widehat{g} \equiv g \pmod{n_g}, \\ n &= n_g = \frac{n}{(n,g)}, \end{aligned}$$

and restart from I step.

Particular case: $n = g^r$.

If $n = g^r$ ($r \in \mathbb{N}^+$, $g \geq 2$), that is, if n is a power of g, then we can find explicitly the eigenvalues of $C_{n,g}$ by following the recursive algorithm proposed above. Indeed, since $g \notin \{0, 1\}$ and $(n, g) = g \neq 1$, the algorithm proceeds by performing repeatedly the step IV:

$$\begin{aligned} \Lambda(C_{n,g}) &= \Lambda(C_{g^r,g}) \\ &= \Lambda\left(C_{g^{r-1},g}\right) \cup \left\{0 \text{ with geometric multiplicity } n - g^{r-1}\right\} \\ &= \Lambda\left(C_{g^{r-2},g}\right) \cup \left\{0 \text{ with geometric multiplicity } g^{r-1} - g^{r-2}\right\} \cup \\ &\quad \cup \left\{0 \text{ with geometric multiplicity } n - g^{r-1}\right\} \\ &= \Lambda\left(C_{g^{r-3},g}\right) \cup \left\{0 \text{ with geometric multiplicity } g^{r-2} - g^{r-3}\right\} \cup \\ &\quad \cup \left\{0 \text{ with geometric multiplicity } g^{r-1} - g^{r-2}\right\} \cup \\ &\quad \cup \left\{0 \text{ with geometric multiplicity } n - g^{r-1}\right\} \\ &= \cdots\cdots\cdots\cdots\cdots\cdots \\ &= \Lambda(C_{g,0}) \cup \left\{0 \text{ with geometric multiplicity } g^2 - g\right\} \cup \cdots \\ &\quad \cdots \cup \left\{0 \text{ with geometric multiplicity } g^{r-2} - g^{r-3}\right\} \cup \\ &\quad \cup \left\{0 \text{ with geometric multiplicity } g^{r-1} - g^{r-2}\right\} \cup \\ &\quad \cup \left\{0 \text{ with geometric multiplicity } n - g^{r-1}\right\}, \end{aligned}$$

at the end we arrive at the matrix $C_{g,0}$ that, for the step I of the algorithm, has $g-1$ eigenvalues equal to zero and only one eigenvalue different from zero. If we come back in the equalities we can conclude that $C_{n,g}$ has $n - 1$ eigenvalues equal to zero and one eigenvalue different zero; now, since the trace of a matrix is the sum of its eigenvalues, the only eigenvalue λ different from zero of $C_{n,g} = \left[a_{(r-gs) \bmod n}\right]_{r,s=0}^{n-1}$, is given by

$$\lambda = \operatorname{tr}(C_{n,g}) = \sum_{j=0}^{n-1} (C_{n,g})_{j,j} = \sum_{j=0}^{n-1} a_j,$$

so the matrix $C_{g^r,g}$ has the same eigenvalues of the matrix $C_{g^r,0}$.

5.4 Toeplitz and g-Toeplitz matrices

In Chapter 4 we have introduced the Toeplitz matrices and we have seen that, given a function $f \in L^1(Q)$, $Q = (-\pi, \pi)$, with Fourier coefficients a_j ($a_j = \tilde{f}_j$ as in (4.2) with $d = 1$), if we denote by $T_n := T_n(f)$ the classical Toeplitz matrix generated by the function f (see Definition 4.1), $T_n = [a_{r-c}]_{r,c=0}^{n-1}$, then the sequence of Toeplitz matrices $\{T_n\}$ is distributed in the singular value sense as the function f: $\{T_n\} \sim_\sigma (f, Q)$ (see Proposition 4.2).

We want to prove a similar distribution result for the g-Toeplitz matrices, where, like in the g-circulant case, a generic g-Toeplitz matrix of dimension $n \times n$ is defined as

$$T_{n,g} = [a_{r-gc}]_{r,c=0}^{n-1}, \tag{5.62}$$

where the quantities $r - gs$ are not reduced modulus n, for example, if $n = 5$ and $g = 3$, then

$$T_{5,3} = \begin{bmatrix} a_0 & a_{-3} & a_{-6} & a_{-9} & a_{-12} \\ a_1 & a_{-2} & a_{-5} & a_{-8} & a_{-11} \\ a_2 & a_{-1} & a_{-4} & a_{-7} & a_{-10} \\ a_3 & a_0 & a_{-3} & a_{-6} & a_{-9} \\ a_4 & a_1 & a_{-2} & a_{-5} & a_{-8} \end{bmatrix}.$$

In analogy with the case of $g = 1$ (the "classical" Toeplitz matrix), we consider a_j as the Fourier coefficients of some function f in $L^1(Q)$.

If we denote by T_n the classical Toeplitz matrix generated by the function $f \in L^1(Q)$, and by $T_{n,g}$ the g-Toeplitz matrix generated by the same function (in this case the two matrices T_n and $T_{n,g}$ have the same first column), one verifies immediately for n and g generic that

$$T_{n,g} = \left[\widehat{T}_{n,g} | \widetilde{T}_{n,g}\right] = \left[T_n \widehat{Z}_{n,g} | \widetilde{T}_{n,g}\right], \tag{5.63}$$

where $\widehat{T}_{n,g} \in M_{n,\mu_g}(\mathbb{C})$, $\mu_g = \left\lceil \frac{n}{g} \right\rceil$, is the matrix $T_{n,g}$ defined in (5.62) by considering only the μ_g first columns, $\widetilde{T}_{n,g} \in M_{n,(n-\mu_g)}(\mathbb{C})$ is the matrix $T_{n,g}$ defined in (5.62) by considering only the $n - \mu_g$ last columns, and $\widehat{Z}_{n,g} \in M_{n,\mu_g}(\mathbb{R})$ is the matrix defined in (5.9) by considering only the μ_g first columns.

Proof. (of relation (5.63)). For $r = 0, 1, \ldots, n-1$, and $s = 0, 1, \ldots, \mu_g - 1$, one has

$$\begin{aligned} \left(\widehat{T}_{n,g}\right)_{r,s} &= (T_n)_{r,gs}, \\ \left(\widehat{Z}_{n,g}\right)_{r,s} &= \delta_{r-gs}, \end{aligned}$$

and

$$\begin{aligned} \left(T_n \widehat{Z}_{n,g}\right)_{r,s} &= \sum_{l=0}^{n-1} (T_n)_{r,l} \left(\widehat{Z}_{n,g}\right)_{l,s} \\ &= \sum_{l=0}^{n-1} \delta_{l-gs} (T_n)_{r,l} \\ &\underset{\text{(a)}}{=} (T_n)_{r,gs} \\ &= \left(\widehat{T}_{n,g}\right)_{r,s}, \end{aligned}$$

where (a) follows because there exists a unique $l \in \{0, 1, \ldots, n-1\}$ such that $l - gs \equiv 0 \pmod{n}$, that is, $l \equiv gs \pmod{n}$, and, since $0 \le gs \le n - 1$, we obtain $l = gs$. □

If we take the matrix $\widehat{T}_{n,g} \in M_{n,(\mu_g+1)}(\mathbb{C})$, then relation (5.63) is no longer true. In reality, looking at the (μ_g+1)th column of the g-Toeplitz we observe Fourier coefficients with indices which are not present (less or equal to $-n$) in the Toeplitz matrix T_n. More precisely,

$$(T_{n,g})_{0,\mu_g} = a_{0-g\mu_g} = a_{-g\mu_g}, \qquad \text{and } -g\mu_g \leq -n.$$

It follows that μ_g is the maximum number of columns for which relation (5.63) is true.

5.5 Singular value distribution for the g-Toeplitz sequences

As stated in formula (5.63), the matrix $T_{n,g}$ can be written as

$$\begin{aligned} T_{n,g} &= \left[T_n\widehat{Z}_{n,g}|\widetilde{T}_{n,g}\right] \\ &= \left[\begin{array}{c|c} T_n\widehat{Z}_{n,g} & 0 \end{array}\right] + \left[\begin{array}{c|c} 0 & \widetilde{T}_{n,g} \end{array}\right]. \end{aligned} \tag{5.64}$$

To find the distribution in the singular value sense of the sequence $\{T_{n,g}\}$, the idea is to study separately the distribution of the two sequences $\left\{\left[T_n\widehat{Z}_{n,g}|0\right]\right\}$ and $\left\{\left[0|\widetilde{T}_{n,g}\right]\right\}$, to prove $\left\{\left[0|\widetilde{T}_{n,g}\right]\right\} \sim_\sigma (0, G)$, and then apply Proposition 2.4.

Proposition 5.18. *Let $T_{n,g}$ be the g-Toeplitz matrix generated by the Fourier coefficients of a function $f \in L^1(Q)$. If we consider the matrix $\left[T_n\widehat{Z}_{n,g}|0\right]$ defined in (5.64), it holds that*

$$\left\{\left[T_n\widehat{Z}_{n,g}|0\right]\right\} \sim_\sigma (\theta, Q \times [0,1]),$$

where

$$\theta(x,t) = \begin{cases} \sqrt{\widehat{|f|^{(2)}}(x)} & \text{for } t \in \left[0, \frac{1}{g}\right], \\ 0 & \text{for } t \in \left(\frac{1}{g}, 1\right], \end{cases} \tag{5.65}$$

and

$$\widehat{|f|^{(2)}}(x) = \frac{1}{g}\sum_{j=0}^{g-1} |f|^2\left(\frac{x+2\pi j}{g}\right). \tag{5.66}$$

Proof. Since $T_n\widehat{Z}_{n,g} \in M_{n,\mu_g}(\mathbb{C})$ and $\left[T_n\widehat{Z}_{n,g}|0\right] \in M_n(\mathbb{C})$, the matrix $\left[T_n\widehat{Z}_{n,g}|0\right]$ has $n-\mu_g$ singular values equal to zero and the remaining μ_g equal to those of $T_n\widehat{Z}_{n,g}$; to study the distribution in the singular value sense of this sequence of non-square matrices, we use Lemma 2.2: consider the g-Toeplitz matrix "truncated" $\widehat{T}_{n,g} = T_n(f)\widehat{Z}_{n,g}$, where the elements of the Toeplitz matrix $T_n(f) = [a_{r-c}]_{r,c=0}^{n-1}$ are the Fourier coefficients of a function f in $L^1(Q)$, $Q = (-\pi, \pi)$, then we have

$$\begin{aligned} \widehat{T}_{n,g}^*\widehat{T}_{n,g} &= \left(T_n(f)\widehat{Z}_{n,g}\right)^* T_n(f)\widehat{Z}_{n,g} = \widehat{Z}_{n,g}^\top T_n(f)^* T_n(f)\widehat{Z}_{n,g} \\ &= \widehat{Z}_{n,g}^\top T_n\left(\overline{f}\right) T_n(f)\widehat{Z}_{n,g}. \end{aligned} \tag{5.67}$$

We provide in detail the analysis in the case where $f \in L^2(Q)$. The general setting in which $f \in L^1(Q)$ can be obtained by approximation and density arguments as done in [83]. From Proposition 4.2 if $f \in L^2(Q) \subset L^1(Q)$ (that is, $|f|^2 \in L^1(Q)$), then $\left\{T_n\left(\overline{f}\right)T_n(f)\right\} \sim_\sigma \left(|f|^2, Q\right)$. Consequently, for every m sufficiently large, $m \in \mathbb{N}$, the use of Theorem 2.1 implies

$$T_n\left(\overline{f}\right)T_n(f) = T_n\left(|f|^2\right) + R_{n,m} + N_{n,m}, \qquad \forall n > n_m,$$

with

$$\text{rank}\,(R_{n,m}) \leq nc\,(m)\,, \qquad \|N_{n,m}\| \leq \omega\,(m)\,,$$

where $n_m \geq 0$, $c\,(m)$ and $\omega\,(m)$ depend only on m, and, moreover,

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

Therefore (5.67) becomes

$$\begin{aligned}
\widehat{T}^*_{n,g}\widehat{T}_{n,g} &= \widehat{Z}^\top_{n,g}\left(T_n\left(|f|^2\right) + R_{n,m} + N_{n,m}\right)\widehat{Z}_{n,g} \\
&= \widehat{Z}^\top_{n,g}T_n\left(|f|^2\right)\widehat{Z}_{n,g} + \widehat{Z}^\top_{n,g}R_{n,m}\widehat{Z}_{n,g} + \widehat{Z}^\top_{n,g}N_{n,m}\widehat{Z}_{n,g} \\
&= \widehat{Z}^\top_{n,g}T_n\left(|f|^2\right)\widehat{Z}_{n,g} + \widehat{R}_{n,m,g} + \widehat{N}_{n,m,g}, \qquad (5.68)
\end{aligned}$$

with

$$\text{rank}\left(\widehat{R}_{n,m,g}\right) \leq \min\left\{\text{rank}\left(\breve{Z}_{n,g}\right), \text{rank}\,(R_{n,m})\right\} \leq \text{rank}\,(R_{n,m}) \leq nc\,(m)\,, \quad (5.69)$$

$$\left\|\widehat{N}_{n,m,g}\right\| \leq 2\left\|\breve{Z}_{n,g}\right\|\|N_{n,m}\| \leq 2\omega\,(m)\,, \qquad (5.70)$$

and

$$\lim_{m\to\infty} c\,(m) = 0, \qquad \lim_{m\to\infty} 2\omega\,(m) = 0,$$

where in (5.69) and (5.70), $\breve{Z}_{n,g} = \left[\widehat{Z}_{n,g}|0\right] \in M_n\,(\mathbb{R})$. In other words $\breve{Z}_{n,g}$ is the matrix $\widehat{Z}_{n,g}$ supplemented by an appropriate number of zero columns in order to make it square. Furthermore, it is worth noting that $\left\|\widehat{Z}_{n,g}\right\| = \left\|\widehat{Z}^\top_{n,g}\right\| = 1$, because $\widehat{Z}_{n,g}$ is a submatrix of the identity; we have used the latter relations in (5.70).

Now, consider the matrix $\widehat{Z}^\top_{n,g}T_n\left(|f|^2\right)\widehat{Z}_{n,g} \in M_{\mu_g}\,(\mathbb{C})$, with $\mu_g = \left\lceil\frac{n}{g}\right\rceil$, $f \in L^2\,(Q) \subset L^1\,(Q)$ (so $|f|^2 \in L^1\,(Q)$). From (5.63), setting $T_n = T_n\left(|f|^2\right) = [\tilde{a}_{r-c}]^{n-1}_{r,c=0}$, with $\tilde{a}_j$ being the Fourier coefficients of $|f|^2$, and setting $T_{n,g}$ the g-Toeplitz generated by the same function $|f|^2$, it is immediate to observe

$$T_n\widehat{Z}_{n,g} = \widehat{T}_{n,g} \in M_{n,\mu_g}\,(\mathbb{C})\,, \qquad \text{with} \quad \left(\widehat{T}_{n,g}\right)_{r,c} = \tilde{a}_{r-gc}, \qquad (5.71)$$

for $r = 0, \ldots, n-1$ and $c = 0, \ldots, \mu_g - 1$. If we compute $\widehat{Z}^\top_{n,g}\widehat{T}_{n,g} \in M_{\mu_g}\,(\mathbb{C})$, where $Z^\top_{n,g} = [\delta_{c-gr}]^{n-1}_{r,c=0}$ (δ_k defined as in (5.9)) and $\widehat{Z}^\top_{n,g} \in M_{\mu_g,n}\,(\mathbb{R})$ is the submatrix of $Z^\top_{n,g}$ obtained by considering only the μ_g first rows, for $r, c = 0, \ldots, \mu_g - 1$, using (5.71) we obtain

$$\begin{aligned}
\left(\widehat{Z}^\top_{n,g}T_n\left(|f|^2\right)\widehat{Z}_{n,g}\right)_{r,c} &= \left(\widehat{Z}^\top_{n,g}\widehat{T}_{n,g}\right)_{r,c} \\
&= \sum_{\ell=0}^{n-1}\left(\widehat{Z}^\top_{n,g}\right)_{r,\ell}\left(\widehat{T}_{n,g}\right)_{\ell,c} \\
&\underset{\text{(a)}}{=} \left(\widehat{T}_{n,g}\right)_{gr,c} \\
&= \widehat{a}_{gr-gc},
\end{aligned}$$

where (a) follows from the existence of a unique $\ell \in \{0, 1, \ldots, n-1\}$ such that $\ell - gr \equiv 0 \pmod{n}$, that is, $\ell \equiv gr \pmod{n}$, and, since $0 \leq gr \leq n-1$, we find $\ell = gr$.

Therefore $\widehat{Z}_{n,g}^{\top} T_n\left(|f|^2\right) \widehat{Z}_{n,g} = [\tilde{a}_{gr-gc}]_{r,c=0}^{\mu_g-1} = T_{\mu_g}\left(\widehat{|f|^{(2)}}\right)$, where $\widehat{|f|^{(2)}} \in L^1(Q)$ is given by

$$\begin{aligned} \widehat{|f|^{(2)}}(x) &= \frac{1}{g}\sum_{j=0}^{g-1} |f|^2\left(\frac{x+2\pi j}{g}\right), \\ |f|^2(x) &= \sum_{k=-\infty}^{+\infty} \tilde{a}_k e^{\mathrm{i}kx}. \end{aligned} \tag{5.72}$$

Indeed, if we denote by a_j the Fourier coefficients of $\widehat{|f|^{(2)}}$, for $r, c = 0, \ldots, \mu_g - 1$, we have $\mathrm{a}_{r-c} = \tilde{a}_{gr-gc}$, where $\tilde{a}_k$ are the Fourier coefficients of $|f|^2$. This can be demonstrated by observing that, from (4.2), (5.66), and (5.72), we have

$$\begin{aligned} \mathrm{a}_{r-c} &= \frac{1}{2\pi}\int_Q \frac{1}{g}\sum_{j=0}^{g-1}\sum_{k=-\infty}^{+\infty} \tilde{a}_k e^{\mathrm{i}k\left(\frac{x+2\pi j}{g}\right)} e^{-\mathrm{i}(r-c)x}\mathrm{d}x \\ &= \frac{1}{2\pi g}\int_Q \sum_{k=-\infty}^{+\infty} \tilde{a}_k \left(\sum_{j=0}^{g-1} e^{\frac{\mathrm{i}2\pi kj}{g}}\right) e^{\frac{\mathrm{i}kx}{g}} e^{-\mathrm{i}(r-c)x}\mathrm{d}x. \end{aligned}$$

The following remarks are in order:

- if k is a multiple of g, $k = gt$ for some value of t, then we have that

$$\sum_{j=0}^{g-1} e^{\frac{\mathrm{i}2\pi kj}{g}} = \sum_{j=0}^{g-1} e^{\frac{\mathrm{i}2\pi gtj}{g}} = \sum_{j=0}^{g-1} e^{\mathrm{i}2\pi tj} = \sum_{j=0}^{g-1} 1 = g;$$

- if k is not a multiple of g, then $e^{\frac{\mathrm{i}2\pi k}{g}} \neq 1$ and therefore

$$\sum_{j=0}^{g-1} e^{\frac{\mathrm{i}2\pi kj}{g}} = \sum_{j=0}^{g-1} \left(e^{\frac{\mathrm{i}2\pi k}{g}}\right)^j,$$

  is a finite geometric series whose sum is given by

$$\sum_{j=0}^{g-1} \left(e^{\frac{\mathrm{i}2\pi k}{g}}\right)^j = \frac{1-e^{\frac{\mathrm{i}2\pi kg}{g}}}{1-e^{\frac{\mathrm{i}2\pi k}{g}}} = \frac{1-e^{\mathrm{i}2\pi k}}{1-e^{\frac{\mathrm{i}2\pi k}{g}}} = \frac{1-1}{1-e^{\frac{\mathrm{i}2\pi k}{g}}} = 0.$$

Finally, taking into account the latter statements and recalling that

$$\frac{1}{2\pi}\int_Q e^{\mathrm{i}\ell x}\mathrm{d}x = \begin{cases} 1 & \text{if } \ell = 0 \\ 0 & \text{otherwise} \end{cases},$$

we find

$$\begin{aligned} \mathrm{a}_{r-c} &= \frac{1}{2\pi g}\int_Q \sum_{t=-\infty}^{+\infty} \tilde{a}_{gt} g e^{\frac{\mathrm{i}gtx}{g}} e^{-\mathrm{i}(r-c)x}\mathrm{d}x \\ &= \sum_{t=-\infty}^{+\infty} \tilde{a}_{gt} \frac{1}{2\pi}\int_Q e^{\mathrm{i}x(t-(r-c))}\mathrm{d}x \\ &= \tilde{a}_{g(r-c)}, \end{aligned}$$

then if $\tilde{a}_k$ are the Fourier coefficients of $|f|^2(x)$, $\tilde{a}_{gk}$ are the Fourier coefficients of $\widehat{|f|^{(2)}}$.

In summary, from (5.68) we have

$$\widehat{T}_{n,g}^{*}\widehat{T}_{n,g} = T_{\mu_g}\left(\widehat{|f|^{(2)}}\right) + \widehat{R}_{n,m,g} + \widehat{N}_{n,m,g},$$

with $\left\{T_{\mu_g}\left(\widehat{|f|^{(2)}}\right)\right\} \sim_\sigma \left(\widehat{|f|^{(2)}}, Q\right)$. We recall that, owing to (5.66), the relation $|f|^2 \in L^1(Q)$ implies $\widehat{|f|^{(2)}} \in L^1(Q)$. Consequently Theorem 2.1 implies that $\left\{\widehat{T}_{n,g}^{*}\widehat{T}_{n,g}\right\} \sim_\sigma \left(\widehat{|f|^{(2)}}, Q\right)$. Clearly $\widehat{|f|^{(2)}} \in L^1(Q)$ is equivalent to write $\sqrt{\widehat{|f|^{(2)}}} \in L^2(Q)$: therefore, from Lemma 2.2, we infer $\left\{\widehat{T}_{n,g}\right\} \sim_\sigma \left(\sqrt{\widehat{|f|^{(2)}}}, Q\right)$.

Now, as mentioned at the beginning of this proof, by Definition 1.7, we have

$$\begin{aligned}
\lim_{n\to\infty}\frac{1}{n}\sum_{j=1}^{n} F\left(\sigma_j\left(\left[\widehat{T}_{n,g}|0\right]\right)\right) &= \lim_{n\to\infty}\frac{1}{n}\sum_{j=1}^{\mu_g} F\left(\sigma_j\left(\left[\widehat{T}_{n,g}|0\right]\right)\right) + \lim_{n\to\infty}\frac{1}{n}\sum_{j=\mu_g+1}^{n} F(0) \\
&= \lim_{n\to\infty}\frac{\mu_g}{n}\sum_{j=1}^{\mu_g}\frac{F\left(\sigma_j\left(\left[\widehat{T}_{n,g}|0\right]\right)\right)}{\mu_g} + \lim_{n\to\infty}\frac{n-\mu_g}{n}F(0) \\
&= \frac{1}{g}\frac{1}{2\pi}\int_Q F\left(\sqrt{\widehat{|f|^{(2)}}(x)}\right)\mathrm{d}x + \left(1-\frac{1}{g}\right)F(0),
\end{aligned}$$

which results to being equivalent to the following distribution formula:

$$\left\{\left[T_n\widehat{Z}_{n,g}|0\right]\right\} \sim_\sigma (\theta, Q\times[0,1]),$$

where

$$\theta(x,t) = \begin{cases} \sqrt{\widehat{|f|^{(2)}}(x)} & \text{for } t\in\left[0,\frac{1}{g}\right], \\ 0 & \text{for } t\in\left(\frac{1}{g},1\right]. \end{cases}$$

□

Remark 5.19. *We observe that the requirement that the symbol f is square integrable can be removed. In* [83] *it is proven that the singular value distribution of $\{T_n(f)T_n(g)\}$ is given by $h = fg$ with f, g being just Lebesgue integrable and with h that is only measurable and, therefore, may fail to be Lebesgue integrable. This fact is sufficient for extending the proof to the case where $\theta(x,t)$ is defined as in (5.65) with the original symbol $f \in L^1(Q)$.*

Proposition 5.20. *Let $T_{n,g}$ be the g-Toeplitz matrix generated generated by the Fourier coefficients of a function $f \in L^1(Q)$. If we consider the matrix $\left[0|\widetilde{T}_{n,g}\right]$ defined in (5.64), it holds that*

$$\left\{\left[0|\widetilde{T}_{n,g}\right]\right\} \sim_\sigma (0, Q). \tag{5.73}$$

Proof. In perfect analogy with the case of the matrix $\left[T_n\widehat{Z}_{n,g}|0\right]$, we can observe that $\widetilde{T}_{n,g} \in M_{n,(n-\mu_g)}(\mathbb{C})$ and $\left[0|\widetilde{T}_{n,g}\right] \in M_n(\mathbb{C})$. Therefore the matrix $\left[0|\widetilde{T}_{n,g}\right]$ has μ_g singular values equal to zero and the remaining $n-\mu_g$ equal to those of $\widetilde{T}_{n,g}$. However, in this case we have additional difficulties with respect to the matrix $\widehat{T}_{n,g} = T_n\widehat{Z}_{n,g}$, because it is not always true that $\widetilde{T}_{n,g}$ can be written as $T_n\check{Z}_{n,g}$, where $\check{Z}_{n,g}$ is the matrix obtained by considering the $n-\mu_g$ last columns of $Z_{n,g}$. Indeed, in $\widetilde{T}_{n,g}$ there are Fourier coefficients with index, in modulus,

greater than n: the Toeplitz matrix $T_n = [a_{r-c}]_{r,c=0}^{n-1}$ has coefficients a_j with j ranging from $1-n$ to $n-1$, while the g-Toeplitz matrix $T_{n,g} = [a_{r-gc}]_{r,c=0}^{n-1}$ has a_{n-1} as the coefficient of maximum index and $a_{-g(n-1)}$ as the coefficient of minimum index, and, if $g \geq 2$, we have $-g\,(n-1) < -\,(n-1)$.

Even if we take the Toeplitz matrix T_n, which has as its first column the first column of $\widetilde{T}_{n,g}$ and the other generated according to the rule $(T_n)_{j,k} = a_{j-k}$, it is not always true that we can write $\widetilde{T}_{n,g} = T_n P$ for a suitable submatrix P of a permutation matrix; indeed, if the matrix $T_n = [\beta_{r-c}]_{r,c=0}^{n-1}$ has as the first column the first column of $\widetilde{T}_{n,g}$, we find that $\beta_0 = \left(\widetilde{T}_{n,g}\right)_{0,0} = (T_{n,g})_{0,\mu_g} = a_{-g\mu_g}$. As a consequence, T_n has $\beta_{-(n-1)} = a_{-(n-1)-g\mu_g}$ as coefficient of minimum index, while $\widetilde{T}_{n,g}$ has $a_{-g(n-1)}$ as coefficient of minimum index. Therefore,

$$\begin{aligned}
n \leq g\mu_g = g\left\lceil \frac{n}{g} \right\rceil &\leq (n+g-1) \\
&\Downarrow \\
-\,(n-1)\,g - (-\,(n-1) - g\mu_g) &= (1-g)\,(n-1) + g\mu_g \\
&\leq (1-g)\,(n-1) + (n+g-1) \\
&= (1-g)\,(n-1) + (n-1) + g \\
&= (n-1)\,(1-g+1) + g \\
&= (2-g)\,(n-1) + g < 0 \qquad \text{for } g > 2 \text{ and } n > 4.
\end{aligned}$$

Thus, if $g > 2$ and $n > 4$, we have $-\,(n-1)\,g < -\,(n-1) - g\mu_g$ and the coefficient of minimum index $a_{-g(n-1)}$ of $\widetilde{T}_{n,g}$ is not contained in the matrix T_n that has $a_{-(n-1)-g\mu_g}$ as the coefficient of minimum index.

Then we proceed in another way: in the first column of $\widetilde{T}_{n,g} \in M_{n,(n-\mu_g)}(\mathbb{C})$ (and, consequently, throughout the matrix) there are only coefficients with index < 0; indeed the coefficient with the largest index of $\widetilde{T}_{n,g}$ is $\left(\widetilde{T}_{n,g}\right)_{n-1,0} = (T_{n,g})_{n-1,\mu_g} = a_{n-1-g\mu_g}$ and $n-1-g\mu_g \leq n-1-n < 0$, and the coefficient with smallest index is $\left(\widetilde{T}_{n,g}\right)_{0,n-\mu_g-1} = (T_{n,g})_{0,n-\mu_g-1+\mu_g} = (T_{n,g})_{0,n-1} = a_{-g(n-1)}$. Consider, therefore, a Toeplitz matrix $T_{d_{n,g}} \in M_{d_{n,g}}(\mathbb{C})$ with $d_{n,g} > \frac{g(n-1)}{2} + 1$ defined in this way:

$$\begin{aligned}
T_{d_{n,g}} &= \begin{bmatrix}
a_{-d_{n,g}+1} & a_{-d_{n,g}} & a_{-d_{n,g}-1} & \cdots & a_{-2d_{n,g}+2} \\
a_{-d_{n,g}+2} & a_{-d_{n,g}+1} & \ddots & \ddots & a_{-2d_{n,g}+3} \\
\vdots & \ddots & \ddots & \ddots & \vdots \\
a_{-1} & a_{-2} & \ddots & \ddots & a_{-d_{n,g}} \\
a_0 & a_{-1} & a_{-2} & \cdots & a_{-d_{n,g}+1}
\end{bmatrix} \\
&= \left[a_{r-c-d_{n,g}+1}\right]_{r,c=0}^{d_{n,g}-1}.
\end{aligned} \tag{5.74}$$

Since the coefficient with the smallest index is $a_{-2d_{n,g}+2}$, we find

$$-2d_{n,g} + 2 < -2\left(\frac{g\,(n-1)}{2} + 1\right) + 2 = -g\,(n-1) - 2 + 2 = -g\,(n-1)\,.$$

As a consequence, we obtain that all the coefficients of $\widetilde{T}_{n,g}$ are "contained" in the matrix $T_{d_{n,g}}$. In particular, if $d_{n,g} > (g-1)\,(n-1) + 2$ (this condition ensures $d_{n,g} > \frac{g(n-1)}{2} + 1$, that all the subsequent inequalities are correct, and that the size of all the matrices involved are non-negative), then it can be shown that

$$\widetilde{T}_{n,g} = [0_1|I_n|0_2]\,T_{d_{n,g}}\check{Z}_{d_{n,g},g}, \tag{5.75}$$

where $\check{Z}_{d_{n,g},g} \in M_{d_{n,g},(n-\mu_g)}(\mathbb{R})$ is the matrix defined in (5.9), of dimension $d_{n,g} \times d_{n,g}$, by considering only the $n-\mu_g$ first columns and $[0_1|I_n|0_2] \in M_{n,d_{n,g}}(\mathbb{R})$ is a block matrix with $0_1 \in M_{n,(d_{n,g}-g\mu_g-1)}(\mathbb{R})$ and $0_2 \in M_{n,(g\mu_g-n+1)}(\mathbb{R})$.

Indeed, first we observe the following:

- for $r = 0, 1, \ldots, n-1$ and $s = 0, 1, \ldots, n-\mu_g-1$, we have
$$\left(\widetilde{T}_{n,g}\right)_{r,s} = (T_{n,g})_{r,s+\mu_g} = a_{r-gs-g\mu_g}; \tag{5.76}$$
- for $r = 0, 1, \ldots, n-1$ and $s = 0, 1, \ldots, d_{n,g}-1$, we have
$$([0_1|I_n|0_2])_{r,s} = \begin{cases} 1 & \text{if } s = r + d_{n,g} - g\mu_g - 1, \\ 0 & \text{otherwise;} \end{cases} \tag{5.77}$$
- for $r, s = 0, 1, \ldots, d_{n,g}-1$ we have $\left(T_{d_{n,g}}\right)_{r,s} = a_{r-s-d_{n,g}+1}$;
- for $r = 0, 1, \ldots, d_{n,g}-1$ and $s = 0, 1, \ldots, n-\mu_g-1$, we have $\left(\check{Z}_{d_{n,g},g}\right)_{r,s} = \delta_{r-gs}$.

Since $T_{d_{n,g}}\check{Z}_{d_{n,g},g} \in M_{d_{n,g},(n-\mu_g)}(\mathbb{C})$, for $r = 0, 1, \ldots, d_{n,g}-1$ and $s = 0, 1, \ldots, n-\mu_g-1$, it holds that

$$\begin{aligned} \left(T_{d_{n,g}}\check{Z}_{d_{n,g},g}\right)_{r,s} &= \sum_{l=0}^{d_{n,g}-1} \left(T_{d_{n,g}}\right)_{r,l} \left(\check{Z}_{d_{n,g},g}\right)_{l,s} \\ &= \sum_{l=0}^{d_{n,g}-1} \delta_{l-gs} a_{r-l-d_{n,g}+1} \\ &\underset{\text{(a)}}{=} a_{r-gs-d_{n,g}+1}, \end{aligned} \tag{5.78}$$

where (a) follows from the existence of a unique $l \in \{0, 1, \ldots, d_{n,g}-1\}$ such that $l - gs \equiv 0 \pmod{d_{n,g}}$, that is, $l \equiv gs \pmod{d_{n,g}}$, and, since $0 \le gs \le d_{n,g}-1$, we have $l = gs$. Since $[0_1|I_n|0_2]\, T_{d_{n,g}}\check{Z}_{d_{n,g},g} \in M_{n,(n-\mu_g)}(\mathbb{C})$, for $r = 0, 1, \ldots, n-1$ and $s = 0, 1, \ldots, n-\mu_g-1$, using (5.76) we find

$$\begin{aligned} \left([0_1|I_n|0_2]\, T_{d_{n,g}}\check{Z}_{d_{n,g},g}\right)_{r,s} &= \sum_{l=0}^{d_{n,g}-1} ([0_1|I_n|0_2])_{r,l} \left(T_{d_{n,g}}\check{Z}_{d_{n,g},g}\right)_{l,s} \\ &\underset{\text{(d)}}{=} a_{r+d_{n,g}-g\mu_g-1-gs-d_{n,g}+1} \\ &= a_{r-g\mu_g-gs} \\ &= \left(\widetilde{T}_{n,g}\right)_{r,s}, \end{aligned}$$

where (d) follows from (5.78), $\left(T_{d_{n,g}}\check{Z}_{d_{n,g},g}\right)_{l,s} = a_{l-gs-d_{n,g}+1}$, and the following fact: using (5.77), we find $([0_1|I_n|0_2])_{r,l} = 1$ if and only if $l = r + d_{n,g} - g\mu_g - 1$.

We can now observe immediately that the matrix $T_{d_{n,g}}$ defined in (5.74) can be written as

$$T_{d_{n,g}} = JH_{d_{n,g}}, \tag{5.79}$$

where $J \in M_{d_{n,g}}(\mathbb{R})$ is the "flip" permutation matrix, that is, $(J)_{s,t} = 1$ if and only if $s + t = d_{n,g} + 1$, and $H_{d_{n,g}} \in M_{d_{n,g}}(\mathbb{C})$ is the Hankel matrix, that is,

$$H_{d_{n,g}} = \begin{bmatrix} a_0 & a_{-1} & a_{-2} & \cdots & a_{-d_{n,g}+1} \\ a_{-1} & a_{-2} & \iddots & \iddots & a_{-d_{n,g}} \\ \vdots & \iddots & \iddots & \iddots & \vdots \\ a_{-d_{n,g}+2} & a_{-d_{n,g}+1} & \iddots & \iddots & a_{-2d_{n,g}+3} \\ a_{-d_{n,g}+1} & a_{-d_{n,g}} & a_{-d_{n,g}-1} & \cdots & a_{-2d_{n,g}+2} \end{bmatrix}.$$

If $f(x) \in L^1(Q)$, $Q = (-\pi, \pi)$, is the generating function of the Toeplitz matrix $T_n = T_n(f) = [a_{r-c}]_{r,c=0}^{n-1}$ in (5.63), where the (k)th Fourier coefficient of f is a_k, then $f(-x) \in L^1(Q)$ is the generating function of the Hankel matrix $H_{d_{n,g}} = [a_{-r-c}]_{r,c=0}^{d_{n,g}-1}$; by invoking [41, Theorem 6], the sequence of matrices $\left\{H_{d_{n,g}}\right\}$ is distributed in the singular value sense as the zero function: $\left\{H_{d_{n,g}}\right\} \sim_\sigma (0, Q)$. From Lemma 2.5, by (5.79), since J is a unitary matrix, we have $\left\{T_{d_{n,g}}\right\} \sim_\sigma (0, Q)$ as well.

Consider the decomposition in (5.75), that is,

$$\widetilde{T}_{n,g} = [0_1 | I_n | 0_2]\, T_{d_{n,g}} \check{Z}_{d_{n,g},g} = G_{d_{n,g}} T_{d_{n,g}} \check{Z}_{d_{n,g},g}.$$

If we complete the matrix $G_{d_{n,g}} \in M_{n,d_{n,g}}(\mathbb{R})$ and the matrix $\check{Z}_{d_{n,g},g} \in M_{d_{n,g},(n-\mu_g)}(\mathbb{R})$ by adding an appropriate number of zero rows and columns, respectively, in order to make it square, then

$$\begin{aligned} \mathbf{G}_{d_{n,g}} &= \left[\begin{array}{c} G_{d_{n,g}} \\ \hline 0 \end{array}\right] \in M_{d_{n,g}}(\mathbb{R}), \\ \mathbf{Z}_{d_{n,g},g} &= \left[\begin{array}{c|c} \check{Z}_{d_{n,g},g} & 0 \end{array}\right] \in M_{d_{n,g}}(\mathbb{R}), \end{aligned}$$

and it is immediate to note that

$$\mathbf{G}_{d_{n,g}} T_{d_{n,g}} \mathbf{Z}_{d_{n,g},g} = \left[\begin{array}{c|c} \widetilde{T}_{n,g} & 0 \\ \hline 0 & 0 \end{array}\right] = \mathbf{T}_{n,g} \in M_{d_{n,g}}(\mathbb{C}).$$

From Lemma 2.6, since $\left\|\mathbf{G}_{d_{n,g}}\right\| = \left\|\mathbf{Z}_{d_{n,g},g}\right\| = 1$ (indeed they are both "incomplete" permutation matrices), and since $\left\{T_{d_{n,g}}\right\} \sim_\sigma (0, Q)$, we infer that $\{\mathbf{T}_{n,g}\} \sim_\sigma (0, Q)$.

Recall that $\mathbf{T}_{n,g} \in M_{d_{n,g}}(\mathbb{C})$ with $d_{n,g} > (g-1)(n-1)+2$; then we can always choose $d_{n,g}$ such that $gn = d_{n,g} > (g-1)(n-1)+2$ (if $n, g \geq 2$). Now, since $\{\mathbf{T}_{n,g}\} \sim_\sigma (0, Q)$, it holds that the sequence $\{\mathbf{T}_{n,g}\}$ is weakly clustered at zero in the singular value sense, i.e., $\forall \epsilon > 0$,

$$\#\{j : \sigma_j(\mathbf{T}_{n,g}) > \epsilon\} = o(d_{n,g}) = o(gn) = o(n). \tag{5.80}$$

The matrix $\mathbf{T}_{n,g}$ is a block matrix that can be written as

$$\mathbf{T}_{n,g} = \left[\begin{array}{c|c} \widetilde{T}_{n,g} & 0 \\ \hline 0 & 0 \end{array}\right] = \left[\begin{array}{c|c} \left[\widetilde{T}_{n,g}|0\right] & 0 \\ \hline 0 & 0 \end{array}\right],$$

where $\widetilde{T}_{n,g} \in M_{n,(n-\mu_g)}(\mathbb{C})$ and $\left[\widetilde{T}_{n,g}|0\right] \in M_n(\mathbb{C})$. By the SVD we obtain

$$\begin{aligned} \mathbf{T}_{n,g} = \left[\begin{array}{c|c} \left[\widetilde{T}_{n,g}|0\right] & 0 \\ \hline 0 & 0 \end{array}\right] &= \left[\begin{array}{c|c} U_1 \Sigma_1 V_1^* & 0 \\ \hline 0 & U_2 0 V_2^* \end{array}\right] \\ &= \left[\begin{array}{c|c} U_1 & 0 \\ \hline 0 & U_2 \end{array}\right] \left[\begin{array}{c|c} \Sigma_1 & 0 \\ \hline 0 & 0 \end{array}\right] \left[\begin{array}{c|c} V_1 & 0 \\ \hline 0 & V_2 \end{array}\right]^*, \end{aligned}$$

that is, the singular values of $\mathbf{T}_{n,g}$ that are different from zero are the singular values of $\left[\widetilde{T}_{n,g}|0\right] \in M_n(\mathbb{C})$. Thus (5.80) can be written as follows: $\forall \epsilon > 0$,

$$\#\left\{j : \sigma_j\left(\left[\widetilde{T}_{n,g}|0\right]\right) > \epsilon\right\} = o(d_{n,g}) = o(gn) = o(n).$$

The latter relation means that the sequence $\left\{\left[\widetilde{T}_{n,g}|0\right]\right\}$ is weakly clustered at zero in the singular value sense, and hence $\left\{\left[\widetilde{T}_{n,g}|0\right]\right\} \sim_\sigma (0, Q)$. If we now consider the matrix $\hat{G} =$

$\left[\begin{array}{c|c} 0 & I_{n-\mu_g} \\ \hline 0 & 0 \end{array}\right] \in M_n(\mathbb{R})$, where $I_{n-\mu_g} \in M_{n-\mu_g}(\mathbb{R})$ is the identity matrix, then $\left[\widetilde{T}_{n,g}|0\right] \hat{G} = \left[0|\widetilde{T}_{n,g}\right]$, and since $\left\|\hat{G}\right\| = 1$ and $\left\{\left[\widetilde{T}_{n,g}|0\right]\right\} \sim_\sigma (0, Q)$, from Lemma 2.6 we find

$$\left\{\left[0|\widetilde{T}_{n,g}\right]\right\} \sim_\sigma (0, Q).$$

□

Theorem 5.21. *Let $T_{n,g}$ be the g-Toeplitz matrix generated generated by the Fourier coefficients of a function $f \in L^1(Q)$, then it holds that*

$$\{T_{n,g}\} \sim_\sigma (\theta, Q \times [0,1]), \tag{5.81}$$

where θ is defined in (5.65).

Proof. The relation (5.81), using (5.63) and Propositions 5.18 and 5.20, is a direct consequence of Proposition 2.4, with $G = Q \times [0,1]$, □

Notice that for $g = 1$ the symbol $\theta(x,t)$ coincides with $|f|(x)$ on the extended domain $Q \times [0,1]$. Hence the Avram–Parter theorem is found as a particular case. Indeed $\theta(x,t) = |f|(x)$ does not depend on t; therefore this additional variable can be suppressed, i.e., $\{T_{n,g}\} \sim_\sigma (f, Q)$ with $T_{n,g} = T_n(f)$. The fact that the distribution formula is not unique should not surprise since this phenomenon is inherent to the measure theory. In fact, any measure-preserving exchange function is a distribution function if one representative of the class is.

It is worthwhile to refer briefly to the case of sequences of g-Toeplitz matrices with negative g. First we observe that, for $g < 0$, all the coefficients of the g-Toeplitz matrix $T_{n,g} = [a_{r-gc}]_{r,c=0}^{n-1}$, generated by the function f, have a non-negative index, i.e., $r - gc \geq 0$. If we let $H_{d_{n,g}} = [a_{r+c}]_{r,c=0}^{d_{n,g}-1}$ be the Hankel matrix of dimension $d_{n,g} = -gn$, generated by the same symbol f of the g-Toeplitz matrix, it is immediate to verify that $T_{n,g}$ is a submatrix of $H_{d_{n,g}}$. Since $\left\{H_{d_{n,g}}\right\} \sim_\sigma (0, Q)$, following the same reasoning proposed in the proof of Proposition 5.20, we obtain that $\{T_{n,g}\} \sim_\sigma (0, Q)$, $Q = (-\pi, \pi)$.

5.6 Generalizations: the multi-level setting

All the distribution results presented in the previous sections for g-circulant and g-Toeplitz matrices can be extended to the multi-level case in which g denotes a d-dimensional vector of non-negative integers, that is, $g = (g_1, \ldots, g_d)$, and n a d-dimensional vector of positive integers, that is, $n = (n_1, \ldots, n_d)$.

In the following we use the symbol $\circ$ to denote the component-wise Hadamard product between vectors or matrices of the same size, that is, for example, if r and s are d-dimensional vectors we have

$$\begin{aligned} &(r - g \circ s) = ((r_1 - g_1 s_1), (r_2 - g_2 s_2), \ldots, (r_d - g_d s_d)), \\ &(r - g \circ s) \bmod n = ((r_1 - g_1 s_1) \bmod n_1, \ldots, (r_d - g_d s_d) \bmod n_d). \end{aligned}$$

5.6.1 Multi-level circulant and g-circulant matrices

According to the multi-index block notation introduced in Definition 4.1, if p is a d-variate trigonometric polynomial defined over Q^d, where $Q = (-\pi, \pi)$, and taking values in $M_{q_1,q_2}(\mathbb{C})$, with Fourier coefficients a_j, $j = (j_1, \ldots, j_d)$, a multi-level circulant matrix of size $q_1\hat{n} \times q_2\hat{n}$, where $\hat{n} = n_1 n_2 \cdots n_d$, generated by p is defined as

$$C_n(p) = \left[a_{(r-g\circ s) \bmod n} + a_{(r-g\circ s) \bmod n - n}\right]_{r,s=\underline{0}}^{n-e},$$

with $\underline{0}, e \in \mathbb{R}^d$, $\underline{0} = (0, \ldots, 0)$ and $e = (1, \ldots, 1)$, and is straightforward to verify, as in the one-level case, that can be written as

$$C_n(p) = \sum_{j=\underline{0}}^{n-e} Z_{n_1}^{j_1} \otimes Z_{n_2}^{j_2} \otimes \cdots \otimes Z_{n_d}^{j_d} \otimes a_j,$$

where $Z_{n_j} \in M_{n_j}(\mathbb{R})$ is the matrix defined in (5.2) of dimension n_j ($\otimes$ denotes the tensor or Kronecker product of matrices) and a_j is considered to be the matrix of size $q_1 \times q_2$ whose (u, v)th entry is the $(k_1, \ldots, k_d)$th Fourier coefficient of the polynomial $(p(t_1, \ldots, t_d))_{u,v}$. Moreover, if $F_n \in M_{\hat{n}}(\mathbb{C})$ denotes the multi-level Fourier matrix

$$F_n = F_{n_1} \otimes F_{n_2} \otimes \cdots \otimes F_{n_d},$$

where $F_{n_j} \in M_{n_j}(\mathbb{C})$ is the Fourier matrix defined in (5.3) of dimension n_j, then F_n is unitary, $F_n F_n^* = I_n$, and it holds that

$$C_n(p) = (F_n \otimes I_{q_1}) D_n(p) (F_n^* \otimes I_{q_2}),$$

where

$$\begin{aligned} D_n(p) &= \operatorname{diag}\left(\sqrt{\hat{n}}\,(F_n^* \otimes I_{q_1})\,\underline{a}\right) \\ &= \operatorname*{diag}_{k=\underline{0},\ldots,n-e}\left(\sum_{j=\underline{0}}^{n-e} a_j e^{\mathrm{i}2\pi\left(\frac{j_1 k_1}{n_1} + \cdots + \frac{j_d k_d}{n_d}\right)}\right), \end{aligned}$$

$\underline{a}$ being the first column of the matrix $C_n(p)$ whose entries a_j, $j = (j_1, \ldots, j_d)$, are ordered lexicographically. If $q_1 = q_2 = 1$ the singular values of $C_n(p)$, i.e., the diagonal entries of $D_n(p)$, are those of the $\frac{n}{2}$th Fourier sum of p evaluated at the grid points $\frac{2\pi k}{n} = 2\pi\left(\frac{k_1}{n_1}, \ldots, \frac{k_d}{n_d}\right)$, $0 \le k_j \le n_j - 1$, $j = 1, \ldots, d$.

It should be mentioned now that when $q_1 \neq q_2$ the matrices are not square and so it makes no sense to speak of eigenvalues. On the other hand, the singular values are given by the collection of those of the diagonal blocks of $D_n(p)$, while, when $q_1 = q_2$, the matrix $C_n(p)$ is square and its eigenvalues are expressible as the collection of those of the diagonal blocks of $D_n(p)$.

Using the multi-level notation introduced at the beginning of this section, a g-circulant matrix $C_{n,g} \in M_{q_1\hat{n} \times q_2\hat{n}}(\mathbb{C})$ generated by a given d-variate polynomial p, is given by

$$C_{n,g}(p) = \left[a_{(r-g\circ s) \bmod n} + a_{(r-g\circ s) \bmod n - n}\right]_{r,s=\underline{0}}^{n-e},$$

where, as in the circulant case, a_j are the Fourier coefficients of p.

Following the analysis in Subsection 5.2.1, for g fixed vector and n increasing sequence of vectors we do not find a joint distribution. Assuming $\{C_n(p)\} \sim_\sigma \left(p, Q^d\right)$ with $\{C_n(p)\}$ standard sequence of multi-level circulants (that is g-circulants where g is the vectors of all ones), and assuming that the sequence n is chosen so that $\gamma_i = (n_i, g_i)$, $i = 1, \ldots, d$, are d fixed numbers, we find

$$\{C_{n,g}(p)\} \sim_\sigma \left(\eta_p, Q^d \times [0,1]^d\right), \tag{5.82}$$

where

$$\eta_p(x,t) = \begin{cases} \sqrt{\widehat{|p|^{(2)}}\,(x)} & \text{for } t \in \left[\underline{0}, \frac{1}{\gamma}\right], \\ 0 & \text{for } t \in \left(\frac{1}{\gamma}, e\right], \end{cases} \tag{5.83}$$

with

$$\widehat{|p|^{(2)}}(x) = \sum_{j=\underline{0}}^{\gamma-e} |p|^2 \left(\frac{x+2\pi j}{\gamma}\right), \tag{5.84}$$

where all the arguments are modulus 2π and all the operations are intended component-wise; that is, $t \in \left[\underline{0}, \frac{1}{\gamma}\right]$ means that $t_k \in \left[0, \frac{1}{\gamma_k}\right]$, $k = 1, \ldots, d$, $t \in \left(\frac{1}{\gamma}, e\right]$ means that $t_k \in \left(\frac{1}{\gamma_k}, 1\right]$, $k = 1, \ldots, d$, the writing $\frac{x+2\pi j}{\gamma}$ defines the d-dimensional vector whose (k)th component is $\frac{(x_k+2\pi j_k)}{\gamma_k}$, $k = 1, \ldots, d$.

When some of the entries of g vanish.

The content of this subsection reduces to the following remark: the case of a non-negative g can be reduced to the case of a positive vector. Let g be a d-dimensional vector of non-negative integers, and let $\mathcal{N} \subset \{1, \ldots, d\}$ be the set of indices such that $j \in \mathcal{N}$ if and only if $g_j = 0$. Assume that $\mathcal{N}$ is non-empty, let $t \geq 1$ be its cardinality and let $d^+ = d - t$. Then a simple calculation shows that the singular values of the corresponding g-circulant matrix $C_{n,g} = \left[a_{(r-g\circ s) \bmod n}\right]_{r,s=\underline{0}}^{n-e}$ are zero except for few of them given by $\sqrt{\hat{n}\,[0]}\sigma$ where

$$\hat{n}\,[0] = \prod_{j\in\mathcal{N}} n_j, \qquad n\,[0] = (n_{j_1}, \ldots, n_{j_t}), \qquad \mathcal{N} = \{j_1, \ldots, j_t\},$$

and σ is any singular value of the matrix

$$\left(\sum_{j=\underline{0}}^{n[0]-e} \mathcal{C}_j^* \mathcal{C}_j\right)^{\frac{1}{2}}. \tag{5.85}$$

Here $\mathcal{C}_j$ is a d^+-level g^+-circulant matrix with $g^+ = \left(g_{k_1}, \ldots, g_{k_{d^+}}\right)$, and of partial sizes $n\,[>0] = \left(n_{k_1}, \ldots, n_{k_{d^+}}\right)$, $\mathcal{N}^C = \{k_1, \ldots, k_{d^+}\}$, and whose expression is

$$\mathcal{C}_j = \left[a_{(r-g\circ s) \bmod n}\right]_{r',s'=\underline{0}}^{n[>0]-e},$$

where $(r - g \circ s)_k = j_k$ for $g_k = 0$ and $r'_i = r_{k_i}$, $s'_i = s_{k_i}$, $i = 1, \ldots, d^+$. Since most of the singular values are identically zero, we infer that

$$\{C_{n,g}\} \sim_\sigma (0, G),$$

for any domain G satisfying the requirements of Definition 1.7. Specific examples are worked out explicitly in Subsection 5.6.3.

Case $g = \underline{0}$. When $g = \underline{0}$ the multi-level block g-circulant is given by

$$C_{n,0} = \left[a_{(r-\underline{0}\circ s) \bmod n}\right]_{r,s=\underline{0}}^{n-e} = \left[a_{(r) \bmod n}\right]_{r,s=\underline{0}}^{n-e} = [a_r]_{r,s=\underline{0}}^{n-e} = \begin{bmatrix} a_{\underline{0}} & \cdots & a_{\underline{0}} \\ \vdots & & \vdots \\ a_{n-e} & \cdots & a_{n-e} \end{bmatrix}.$$

A simple computation shows that all the singular values are zero except for a few of them given by $\sqrt{\hat{n}}\sigma$, where $\hat{n} = n_1 n_2 \cdots n_d$ and σ is any singular value of the matrix $\left(\sum_{j=\underline{0}}^{n-e} a_j^* a_j\right)^{\frac{1}{2}}$. Of course in the scalar case where $p = q = 1$ the choice of σ is unique and by the above formula it coincides with the Euclidean norm of the first column $\underline{a}$ of the original matrix. In that case it is evident that $\{C_{n,0}\} \sim_\sigma (0, G)$, for any domain G satisfying the requirements of Definition 1.7.

5.6.2 Multi-level g-Toeplitz matrices

For multi-level block Toeplitz sequences $\{T_n\}$ generated by an integrable d variate and matrix-valued symbol f, $T_n := T_n(f)$, the singular values are not explicitly known but we know the distribution in the sense of Definition 1.7; see Chapter 4. More precisely we have

$$\{T_n\} \sim_\sigma \left(f, Q^d\right), \qquad Q = (-\pi, \pi)\,.$$

When g is a positive vector, the g-Toeplitz matrix generated by f is defined as

$$T_{n,g} = \left[a_{(r-g\circ s)}\right]_{r,s=\underline{0}}^{n-e},$$

where a_j are the Fourier coefficients of f, and we have that

$$\{T_{n,g}\} \sim_\sigma \left(\theta, Q^d \times [0,1]^d\right), \tag{5.86}$$

where

$$\theta\,(x,t) = \begin{cases} \sqrt{\widehat{|f|^{(2)}}\,(x)} & \text{for } t \in \left[\underline{0}, \frac{1}{g}\right], \\ 0 & \text{for } t \in \left(\frac{1}{g}, e\right], \end{cases} \tag{5.87}$$

with

$$\widehat{|f|^{(2)}}\,(x) = \frac{1}{\hat{g}} \sum_{j=\underline{0}}^{g-e} |f|^2 \left(\frac{x+2\pi j}{g}\right), \tag{5.88}$$

and where all the arguments are modulus 2π and all the operations are intended componentwise; that is, $t \in \left[\underline{0}, \frac{1}{g}\right]$ means that $t_k \in \left[0, \frac{1}{g_k}\right]$, $k = 1, \ldots, d$, $t \in \left(\frac{1}{g}, e\right]$ means that $t_k \in \left(\frac{1}{g_k}, 1\right]$, $k = 1, \ldots, d$, the writing $\frac{x+2\pi j}{g}$ defines the d-dimensional vector whose (k)th component is $\frac{(x_k+2\pi j_k)}{g_k}$, $k = 1, \ldots, d$, and $\hat{g} = g_1 g_2 \cdots g_d$.

When some of the entries of g vanish.

Taking into account the notations of Subsection 5.6.1, for the g-Toeplitz $T_{n,g} = [a_{r-g\circ s}]_{r,s=\underline{0}}^{n-e}$ the same computation shows that all the singular values are zero except for a few of them given by $\sqrt{\hat{n}\,[0]}\sigma$ where σ is any singular value of the matrix

$$\left(\sum_{j=\underline{0}}^{n[0]-e} \mathcal{T}_j^* \mathcal{T}_j\right)^{\frac{1}{2}}. \tag{5.89}$$

Here $\mathcal{T}_j$ is a d^+-level g^+-Toeplitz matrix with $g^+ = \left(g_{k_1}, \ldots, g_{k_{d^+}}\right)$, and of partial sizes $n\,[>0] = \left(n_{k_1}, \ldots, n_{k_{d^+}}\right)$, $\mathcal{N}^C = \{k_1, \ldots, k_{d^+}\}$, and whose expression is

$$\mathcal{T}_j = \left[a_{(r-g\circ s)}\right]_{r',s'=\underline{0}}^{n[>0]-e},$$

where $(r - g \circ s)_k = j_k$ for $g_k = 0$ and $r'_i = r_{k_i}$, $s'_i = s_{k_i}$, $i = 1, \ldots, d^+$. Also in this case, since most of the singular values are identically zero, we infer that

$$\{T_{n,g}\} \sim_\sigma (0, G)\,,$$

for any domain G satisfying the requirements of Definition 1.7.

Concrete examples of g-Toeplitz sequences, where some of the entries of g vanish, are treated in detail in the next subsection together with their spectra.

Case $g = \underline{0}$. If $g = 0$ the g-Toeplitz matrix coincides with the g-circulant matrix, so the result is the same seen in Subsection 5.6.1:

$$\{T_{n,g}\} \sim_\sigma (0, G),$$

for any domain G satisfying the requirements of Definition 1.7.

5.6.3 Examples of g-circulant and g-Toeplitz matrices when some of the entries of g vanish

We start this subsection with a brief digression on multi-level matrices. A d-level matrix $A \in M_{\hat{n}}(\mathbb{C})$ with $n = (n_1, n_2, \ldots, n_d)$ and $\hat{n} = n_1 n_2 \cdots n_d$ can be viewed as a matrix of dimension $n_1 \times n_1$ in which each element is a block of dimension $n_2 n_3 \cdots n_d \times n_2 n_3 \cdots n_d$; in turn, each block of dimension $n_2 n_3 \cdots n_d \times n_2 n_3 \cdots n_d$ can be viewed as a matrix of dimension $n_2 \times n_2$ in which each element is a block of dimension $n_3 n_4 \cdots n_d \times n_3 n_4 \cdots n_d$, and so on. So we can say that n_1 is the most "outer" dimension of the matrix A and n_d is the most "inner" dimension. If we multiply by an appropriate permutation matrix P the d-level matrix A, we can exchange the "order of dimensions" of A, namely $P^\top AP$ becomes a matrix again of dimension $\hat{n} \times \hat{n}$ but with $n = \left(n_{p(1)}, n_{p(2)}, \ldots, n_{p(d)}\right)$ and $\hat{n} = n_{p(1)} n_{p(2)} \cdots n_{p(d)} = n_1 n_2 \cdots n_d$ (where p is a permutation of d elements) and $n_{p(1)}$ is the most "outer" dimension of the matrix A and $n_{p(d)}$ is the most "inner" dimension.

This trick helps us to understand what happens to the singular values of g-circulant and g-Toeplitz d-level matrices, especially when some of the entries of the vector g are zero; indeed, as we observed in Subsection 5.6.1, if $g = \underline{0}$, the d-level g-circulant (or g-Toeplitz) matrix $C_{n,0}$ is a block matrix with constant blocks on each row, so if we order the vector g (which has some components equal to zero) so that the components equal to zero are in the top positions, $g = (0, \ldots, 0, g_k, \ldots, g_d)$, the matrix $P^\top C_{n,0} P$ (where P is the permutation matrix associated with p) becomes a block matrix with constant blocks on each row and with blocks of dimension $n_k \cdots n_d \times n_k \cdots n_d$; with this "new" structure, formulae (5.85) and (5.89) are even more intuitively understandable, as we shall see later in the examples.

Lemma 5.22. *Let $T_n \in M_{\hat{n}}(\mathbb{C})$ be a 2-level Toeplitz matrix with $n = (n_1, n_2)$ and $\hat{n} = n_1 n_2$,*

$$T_n = \left[\left[a_{(j_1-k_1, j_2-k_2)}\right]_{j_2,k_2=0}^{n_2-1}\right]_{j_1,k_1=0}^{n_1-1}.$$

There exists a permutation matrix P such that

$$P^\top T_n P = \left[\left[a_{(j_1-k_1, j_2-k_2)}\right]_{j_1,k_1=0}^{n_1-1}\right]_{j_2,k_2=0}^{n_2-1}.$$

Example 5.23. *Let $n = (n_1, n_2) = (2, 3)$ and consider the 2-level Toeplitz matrix T_n of dimension 6×6*

$$T_n = \left[\begin{array}{ccc|ccc} a_{(0,0)} & a_{(0,-1)} & a_{(0,-2)} & a_{(-1,0)} & a_{(-1,-1)} & a_{(-1,-2)} \\ a_{(0,1)} & a_{(0,0)} & a_{(0,-1)} & a_{(-1,1)} & a_{(-1,0)} & a_{(-1,-1)} \\ a_{(0,2)} & a_{(0,1)} & a_{(0,0)} & a_{(-1,2)} & a_{(-1,1)} & a_{(-1,0)} \\ \hline a_{(1,0)} & a_{(1,-1)} & a_{(1,-2)} & a_{(0,0)} & a_{(0,-1)} & a_{(0,-2)} \\ a_{(1,1)} & a_{(1,0)} & a_{(1,-1)} & a_{(0,1)} & a_{(0,0)} & a_{(0,-1)} \\ a_{(1,2)} & a_{(1,1)} & a_{(1,0)} & a_{(0,2)} & a_{(0,1)} & a_{(0,0)} \end{array}\right].$$

This matrix can be viewed as a matrix of dimension 2×2 in which each element is a block

of dimension 3×3. *If we take the permutation matrix*

$$P = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix},$$

then it is plain to see that

$$P^{\top} T_n P = \left[\begin{array}{cc|cc|cc} a_{(0,0)} & a_{(-1,0)} & a_{(0,-1)} & a_{(-1,-1)} & a_{(0,-2)} & a_{(-1,-2)} \\ a_{(1,0)} & a_{(0,0)} & a_{(1,-1)} & a_{(0,-1)} & a_{(1,-2)} & a_{(0,-2)} \\ \hline a_{(0,1)} & a_{(-1,1)} & a_{(0,0)} & a_{(-1,0)} & a_{(0,-1)} & a_{(-1,-1)} \\ a_{(1,1)} & a_{(0,1)} & a_{(1,0)} & a_{(0,0)} & a_{(1,-1)} & a_{(0,-1)} \\ \hline a_{(0,2)} & a_{(-1,2)} & a_{(0,1)} & a_{(-1,1)} & a_{(0,0)} & a_{(-1,0)} \\ a_{(1,2)} & a_{(0,2)} & a_{(1,1)} & a_{(0,1)} & a_{(1,0)} & a_{(0,0)} \end{array}\right],$$

and now $P^{\top} T_n P$ *can be naturally viewed as a matrix of dimension* 3×3 *in which each element is a block of dimension* 2×2.

Corollary 5.24. *Let* $T_n \in M_{\hat{n}}(\mathbb{C})$ *be a d-level Toeplitz matrix with* $n = (n_1, n_2, \ldots, n_d)$ *and* $\hat{n} = n_1 n_2 \cdots n_d$,

$$T_n = \left[\left[\cdots \left[a_{(j_1 - k_1, j_2 - k_2, \ldots, j_d - k_d)} \right]_{j_d, k_d = 0}^{n_d - 1} \cdots \right]_{j_2, k_2 = 0}^{n_2 - 1} \right]_{j_1, k_1 = 0}^{n_1 - 1}.$$

For every permutation p *of* d *elements, there exists a permutation matrix* P *such that*

$$P^{\top} T_n P = \left[\left[\cdots \left[a_{(j_1 - k_1, j_2 - k_2, \ldots, j_d - k_d)} \right]_{j_{p(d)}, k_{p(d)} = 0}^{n_{p(d)} - 1} \cdots \right]_{j_{p(2)}, k_{p(2)} = 0}^{n_{p(2)} - 1} \right]_{j_{p(1)}, k_{p(1)} = 0}^{n_{p(1)} - 1}.$$

Remark 5.25. *Lemma 5.22 and Corollary 5.24 also apply to d-level g-circulant and g-Toeplitz matrices.*

Now, let $g = (g_1, g_2, \ldots, g_d)$ be a d-dimensional vector of non-negative integers and $t = \sharp \{ j : g_j = 0 \}$ be the number of zero entries of g. If we take a permutation p of d elements such that $g_{p(1)} = g_{p(2)} = \cdots = g_{p(t)} = 0$, (that is, p is a permutation that moves all the zero components of the vector g in the top positions), then it is easy to prove that formulae (5.85) and (5.89) remain the same for the matrices $P^{\top} C_{n,g} P$ and $P^{\top} T_{n,g} P$, respectively (where P is the permutation matrix associated with p), but with $n[0] = \left(n_{p(1)}, n_{p(2)}, \ldots, n_{p(t)} \right)$ and where $\mathcal{C}_j$ and $\mathcal{T}_j$ are a d^+-level g^+-circulant and g^+-Toeplitz matrix, respectively, with $g^+ = \left(g_{p(t+1)}, g_{p(t+2)}, \ldots, g_{p(d)} \right)$, of partial sizes $n[>0] = \left(n_{p(t+1)}, n_{p(t+2)}, \ldots, n_{p(d)} \right)$, and whose expressions are

$$\mathcal{C}_j = \left[\left[\cdots \left[a_{(r - g \circ s) \bmod n} \right]_{r_{p(d)}, s_{p(d)} = 0}^{n_{p(d)} - 1} \cdots \right]_{r_{p(t+2)}, s_{p(t+2)} = 0}^{n_{p(t+2)} - 1} \right]_{r_{p(t+1)}, s_{p(t+1)} = 0}^{n_{p(t+1)} - 1},$$

$$\mathcal{T}_j = \left[\left[\cdots \left[a_{(r - g \circ s)} \right]_{r_{p(d)}, s_{p(d)} = 0}^{n_{p(d)} - 1} \cdots \right]_{r_{p(t+2)}, s_{p(t+2)} = 0}^{n_{p(t+2)} - 1} \right]_{r_{p(t+1)}, s_{p(t+1)} = 0}^{n_{p(t+1)} - 1},$$

with $\left(r_{p(1)}, r_{p(2)}, \ldots, r_{p(t)} \right) = j$. Obviously

$$\Omega(C_{n,g}) = \Omega\left(P^{\top} C_{n,g} P \right),$$
$$\Omega(T_{n,g}) = \Omega\left(P^{\top} T_{n,g} P \right).$$

We proceed with two detailed examples: a 3-level g-circulant matrix with $g = (g_1, g_2, g_3) = (1, 2, 0)$, and a 3-level g-Toeplitz with $g = (g_1, g_2, g_3) = (0, 1, 2)$, which helps us to understand what happens if the vector g is not strictly positive. Finally we will propose the explicit calculation of the singular values of a d-level g-circulant matrix in the particular case where the vector g has only one component different from zero.

Example 5.26. *Consider a 3-level g-circulant matrix $C_{n,g}$ where $g = (g_1, g_2, g_3) = (1, 2, 0)$*

$$\begin{aligned} C_{n,g} &= \left[\left[\left[a_{((r_1-1\cdot s_1)\bmod n_1,(r_2-2\cdot s_2)\bmod n_2,(r_3-0\cdot s_3)\bmod n_3)}\right]_{r_3,s_3=0}^{n_3-1}\right]_{r_2,s_2=0}^{n_2-1}\right]_{r_1,s_1=0}^{n_1-1} \\ &= \left[\left[\left[a_{((r_1-s_1)\bmod n_1,(r_2-2s_2)\bmod n_2,r_3)}\right]_{r_3=0}^{n_3-1}\right]_{r_2,s_2=0}^{n_2-1}\right]_{r_1,s_1=0}^{n_1-1}. \end{aligned}$$

If we choose a permutation p of 3 elements such that

$$\begin{aligned} &(p(1), p(2), p(3)) = (3, 2, 1), \\ &\left(g_{p(1)}, g_{p(2)}, g_{p(3)}\right) = (0, 2, 1), \\ &\left(n_{p(1)}, n_{p(2)}, n_{p(3)}\right) = (n_3, n_2, n_1), \end{aligned}$$

and if we take the permutation matrix P related to p, then

$$P^\top C_{n,g} P \equiv \hat{C}_{n,g} = \left[\left[\left[a_{((r_1-s_1)\bmod n_1,(r_2-2s_2)\bmod n_2,r_3)}\right]_{r_1,s_1=0}^{n_1-1}\right]_{r_2,s_2=0}^{n_2-1}\right]_{r_3=0}^{n_3-1}.$$

Now, for $r_3 = 0, 1, ..., n_3 - 1$, let us set

$$\mathcal{C}_{r_3} = \left[\left[a_{((r_1-s_1)\bmod n_1,(r_2-2s_2)\bmod n_2,r_3)}\right]_{r_1,s_1=0}^{n_1-1}\right]_{r_2,s_2=0}^{n_2-1}.$$

As a consequence, $\mathcal{C}_{r_3}$ is a 2-level g^+-circulant matrix with $g^+ = (2, 1)$ and of partial sizes $n\,[>0] = (n_2, n_1)$ and the matrix $\hat{C}_{n,g}$ can be rewritten as

$$\hat{C}_{n,g} = \begin{bmatrix} \mathcal{C}_0 & \mathcal{C}_0 & \cdots & \mathcal{C}_0 \\ \mathcal{C}_1 & \mathcal{C}_1 & \cdots & \mathcal{C}_1 \\ \vdots & \vdots & \vdots & \vdots \\ \mathcal{C}_{n_3-1} & \mathcal{C}_{n_3-1} & \cdots & \mathcal{C}_{n_3-1} \end{bmatrix},$$

and this is a block matrix with constant blocks on each row. From formula (1.5), the singular

values of $\hat{C}_{n,g}$ *are the square root of the eigenvalues of* $\hat{C}^*_{n,g}\hat{C}_{n,g}$*:*

$$\begin{aligned}
\hat{C}^*_{n,g}\hat{C}_{n,g} &= \begin{bmatrix} \mathcal{C}_0^* & \mathcal{C}_1^* & \cdots & \mathcal{C}_{n_3-1}^* \\ \mathcal{C}_0^* & \mathcal{C}_1^* & \cdots & \mathcal{C}_{n_3-1}^* \\ \vdots & \vdots & \vdots & \vdots \\ \mathcal{C}_0^* & \mathcal{C}_1^* & \cdots & \mathcal{C}_{n_3-1}^* \end{bmatrix} \begin{bmatrix} \mathcal{C}_0 & \mathcal{C}_0 & \cdots & \mathcal{C}_0 \\ \mathcal{C}_1 & \mathcal{C}_1 & \cdots & \mathcal{C}_1 \\ \vdots & \vdots & \vdots & \vdots \\ \mathcal{C}_{n_3-1} & \mathcal{C}_{n_3-1} & \cdots & \mathcal{C}_{n_3-1} \end{bmatrix} \\
&= \begin{bmatrix} \sum_{j=0}^{n_3-1} \mathcal{C}_j^*\mathcal{C}_j & \sum_{j=0}^{n_3-1} \mathcal{C}_j^*\mathcal{C}_j & \cdots & \sum_{j=0}^{n_3-1} \mathcal{C}_j^*\mathcal{C}_j \\ \sum_{j=0}^{n_3-1} \mathcal{C}_j^*\mathcal{C}_j & \sum_{j=0}^{n_3-1} \mathcal{C}_j^*\mathcal{C}_j & \cdots & \sum_{j=0}^{n_3-1} \mathcal{C}_j^*\mathcal{C}_j \\ \vdots & \vdots & \vdots & \vdots \\ \sum_{j=0}^{n_3-1} \mathcal{C}_j^*\mathcal{C}_j & \sum_{j=0}^{n_3-1} \mathcal{C}_j^*\mathcal{C}_j & \cdots & \sum_{j=0}^{n_3-1} \mathcal{C}_j^*\mathcal{C}_j \end{bmatrix} \\
&= \underbrace{\begin{bmatrix} 1 & 1 & \cdots & 1 \\ 1 & 1 & \cdots & 1 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & \cdots & 1 \end{bmatrix}}_{n_3 \text{ times}} \otimes \sum_{j=0}^{n_3-1} \mathcal{C}_j^*\mathcal{C}_j \\
&= J_{n_3} \otimes \sum_{j=0}^{n_3-1} \mathcal{C}_j^*\mathcal{C}_j.
\end{aligned}$$

Therefore

$$\Lambda\left(\hat{C}^*_{n,g}\hat{C}_{n,g}\right) = \Lambda\left(J_{n_3} \otimes \sum_{j=0}^{n_3-1} \mathcal{C}_j^*\mathcal{C}_j\right), \tag{5.90}$$

where

$$\Lambda\left(J_{n_3}\right) = \{0, n_3\}, \tag{5.91}$$

because J_{n_3} *is a matrix of rank 1, so it has all eigenvalues equal to zero except one eigenvalue equal to* $\mathrm{tr}\left(J_{n_3}\right) = n_3$ *(*tr *is the trace of a matrix). If we put*

$$\lambda_k = \lambda_k\left(\sum_{j=0}^{n_3-1} \mathcal{C}_j^*\mathcal{C}_j\right), \qquad k = 0, \ldots, n_1n_2 - 1,$$

by exploiting basic properties of the tensor product and taking into consideration (5.90) and (5.91) we find

$$\lambda_k\left(\hat{C}^*_{n,g}\hat{C}_{n,g}\right) = n_3\lambda_k, \qquad k = 0, \ldots, n_1n_2 - 1, \tag{5.92}$$

$$\lambda_k\left(\hat{C}^*_{n,g}\hat{C}_{n,g}\right) = 0, \qquad k = n_1n_2, \ldots, n_1n_2n_3 - 1. \tag{5.93}$$

From (5.92), (5.93) and (1.5), and recalling that $\Omega\left(\hat{C}_{n,g}\right) = \Omega\left(C_{n,g}\right)$*, one obtains that the singular values of* $C_{n,g}$ *are given by*

$$\begin{aligned}
\sigma_k\left(C_{n,g}\right) &= \sqrt{n_3\lambda_k}, \qquad k = 0, \ldots, n_1n_2 - 1, \\
\sigma_k\left(C_{n,g}\right) &= 0, \qquad k = n_1n_2, \ldots, n_1n_2n_3 - 1,
\end{aligned}$$

and, since $\sum_{j=0}^{n_3-1} \mathcal{C}_j^*\mathcal{C}_j$ *is a positive semidefinite matrix, from (1.5) we can write*

$$\begin{aligned}
\sigma_k\left(C_{n,g}\right) &= \sqrt{n_3}\widetilde{\sigma}_k, \qquad k = 0, \ldots, n_1n_2 - 1, \\
\sigma_k\left(C_{n,g}\right) &= 0, \qquad k = n_1n_2, \ldots, n_1n_2n_3 - 1,
\end{aligned}$$

where $\widetilde{\sigma}_k$ *are the singular values of* $\left(\sum_{j=0}^{n_3-1} \mathcal{C}_j^* \mathcal{C}_j\right)^{\frac{1}{2}}$.

Regarding the distribution in the sense of singular values, let $F \in \mathcal{C}_0\left(\mathbb{R}_0^+\right)$, *continuous function over* $\mathbb{R}_0^+$ *with bounded support, then there exists* $a \in \mathbb{R}^+$ *such that*

$$|F(x)| \leq a \ \ \forall x \in \mathbb{R}_0^+. \tag{5.94}$$

From formula (1.19) we have

$$\begin{aligned}
\Sigma_\sigma(F, C_{n,g}) &= \frac{1}{n_1 n_2 n_3} \sum_{k=0}^{n_1 n_2 n_3 - 1} F\left(\sqrt{n_3}\widetilde{\sigma}_k\right) \\
&= \frac{n_1 n_2 (n_3 - 1) F(0)}{n_1 n_2 n_3} + \frac{1}{n_1 n_2 n_3} \sum_{k=0}^{n_1 n_2 - 1} F\left(\sqrt{n_3}\widetilde{\sigma}_k\right) \\
&= \left(1 - \frac{1}{n_3}\right) F(0) + \frac{1}{n_1 n_2 n_3} \sum_{k=0}^{n_1 n_2 - 1} F\left(\sqrt{n_3}\widetilde{\sigma}_k\right).
\end{aligned}$$

According to (5.94), we find

$$-a n_1 n_2 \leq \sum_{k=0}^{n_1 n_2 - 1} F\left(\sqrt{n_3}\widetilde{\sigma}_k\right) \leq a n_1 n_2.$$

Therefore

$$-\frac{a}{n_3} \leq \frac{1}{n_1 n_2 n_3} \sum_{k=0}^{n_1 n_2 - 1} F\left(\sqrt{n_3}\widetilde{\sigma}_k\right) \leq \frac{a}{n_3},$$

so that

$$\left(1 - \frac{1}{n_3}\right) F(0) - \frac{a}{n_3} \leq \Sigma_\sigma(F, C_{n,g}) \leq \left(1 - \frac{1}{n_3}\right) F(0) + \frac{a}{n_3}.$$

Now, recalling that the writing $n \to \infty$ *means* $\min_{j=1,\ldots,3} n_j \to \infty$, *we obtain*

$$F(0) \leq \lim_{n \to \infty} \Sigma_\sigma(F, C_{n,g}) \leq F(0),$$

which implies

$$\lim_{n \to \infty} \Sigma_\sigma(F, C_{n,g}) = F(0).$$

Whence

$$\{C_{n,g}\} \sim_\sigma (0, G),$$

for any domain G *satisfying the requirements of Definition 1.7.*

Example 5.27. *Consider a 3-level g-Toeplitz matrix* $T_{n,g}$ *where* $g = (g_1, g_2, g_3) = (0, 1, 2)$

$$\begin{aligned}
T_{n,g} &= \left[\left[\left[a_{(r_1 - 0\cdot s_1, r_2 - 1\cdot s_2, r_3 - 2\cdot s_3)}\right]_{r_3, s_3 = 0}^{n_3 - 1}\right]_{r_2, s_2 = 0}^{n_2 - 1}\right]_{r_1, s_1 = 0}^{n_1 - 1} \\
&= \left[\left[\left[a_{(r_1, r_2 - s_2, r_3 - 2s_3)}\right]_{r_3, s_3 = 0}^{n_3 - 1}\right]_{r_2, s_2 = 0}^{n_2 - 1}\right]_{r_1 = 0}^{n_1 - 1}.
\end{aligned}$$

The procedure is the same as in the previous example of a g-circulant matrix, but in this case we do not need to permute the vector g since the only component equal to zero is already in first position. For $r_1 = 0, 1, ..., n_1 - 1$, *let us set*

$$\mathcal{T}_{r_1} = \left[\left[a_{(r_1, r_2 - s_2, r_3 - 2s_3)} \right]_{r_3, s_3 = 0}^{n_3 - 1} \right]_{r_2, s_2 = 0}^{n_2 - 1},$$

then $\mathcal{T}_{r_1}$ *is a 2-level* g^+*-Toeplitz matrix with* $g^+ = (1, 2)$ *and of partial sizes* $n\,[> 0] = (n_2, n_3)$ *and*

$$T_{n,g} = \begin{bmatrix} \mathcal{T}_0 & \mathcal{T}_0 & \cdots & \mathcal{T}_0 \\ \mathcal{T}_1 & \mathcal{T}_1 & \cdots & \mathcal{T}_1 \\ \vdots & \vdots & \vdots & \vdots \\ \mathcal{T}_{n_1-1} & \mathcal{T}_{n_1-1} & \cdots & \mathcal{T}_{n_1-1} \end{bmatrix}.$$

The latter is a block matrix with constant blocks on each row. From formula (1.5), the singular values of $T_{n,g}$ *are the square root of the eigenvalues of* $T_{n,g}^* T_{n,g}$*:*

$$\begin{aligned} T_{n,g}^* T_{n,g} &= \begin{bmatrix} \mathcal{T}_0^* & \mathcal{T}_1^* & \cdots & \mathcal{T}_{n_1-1}^* \\ \mathcal{T}_0^* & \mathcal{T}_1^* & \cdots & \mathcal{T}_{n_1-1}^* \\ \vdots & \vdots & \vdots & \vdots \\ \mathcal{T}_0^* & \mathcal{T}_1^* & \cdots & \mathcal{T}_{n_1-1}^* \end{bmatrix} \begin{bmatrix} \mathcal{T}_0 & \mathcal{T}_0 & \cdots & \mathcal{T}_0 \\ \mathcal{T}_1 & \mathcal{T}_1 & \cdots & \mathcal{T}_1 \\ \vdots & \vdots & \vdots & \vdots \\ \mathcal{T}_{n_1-1} & \mathcal{T}_{n_1-1} & \cdots & \mathcal{T}_{n_1-1} \end{bmatrix} \\ &= \begin{bmatrix} \sum_{j=0}^{n_1-1} \mathcal{T}_j^* \mathcal{T}_j & \sum_{j=0}^{n_1-1} \mathcal{T}_j^* \mathcal{T}_j & \cdots & \sum_{j=0}^{n_1-1} \mathcal{T}_j^* \mathcal{T}_j \\ \sum_{j=0}^{n_1-1} \mathcal{T}_j^* \mathcal{T}_j & \sum_{j=0}^{n_1-1} \mathcal{T}_j^* \mathcal{T}_j & \cdots & \sum_{j=0}^{n_1-1} \mathcal{T}_j^* \mathcal{T}_j \\ \vdots & \vdots & \vdots & \vdots \\ \sum_{j=0}^{n_1-1} \mathcal{T}_j^* \mathcal{T}_j & \sum_{j=0}^{n_1-1} \mathcal{T}_j^* \mathcal{T}_j & \cdots & \sum_{j=0}^{n_1-1} \mathcal{T}_j^* \mathcal{T}_j \end{bmatrix} \\ &= \underbrace{\begin{bmatrix} 1 & 1 & \cdots & 1 \\ 1 & 1 & \cdots & 1 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & \cdots & 1 \end{bmatrix}}_{n_1 \ times} \otimes \sum_{j=0}^{n_1-1} \mathcal{T}_j^* \mathcal{T}_j \\ &= J_{n_1} \otimes \sum_{j=0}^{n_1-1} \mathcal{T}_j^* \mathcal{T}_j. \end{aligned}$$

Therefore

$$\Lambda\left(T_{n,g}^* T_{n,g}\right) = \Lambda\left(J_{n_1} \otimes \sum_{j=0}^{n_1-1} \mathcal{T}_j^* \mathcal{T}_j\right), \tag{5.95}$$

where

$$\Lambda\left(J_{n_1}\right) = \{0, n_1\}, \tag{5.96}$$

because J_{n_1} *is a matrix of rank 1, so it has all eigenvalues equal to zero except one eigenvalue equal to* $\mathrm{tr}\left(J_{n_1}\right) = n_1$ *(*tr *is the trace of a matrix). If we put*

$$\lambda_k = \lambda_k\left(\sum_{j=0}^{n_1-1} \mathcal{T}_j^* \mathcal{T}_j\right), \qquad k = 0, \ldots, n_3 n_2 - 1,$$

by exploiting basic properties of the tensor product and taking into consideration (5.95) and (5.96) we find

$$\begin{aligned} \lambda_k\left(T_{n,g}^* T_{n,g}\right) &= n_1\lambda_k, \qquad k=0,\ldots,n_3n_2-1, \quad (5.97)\\ \lambda_k\left(T_{n,g}^* T_{n,g}\right) &= 0, \qquad k=n_3n_2,\ldots,n_3n_2n_1-1. \quad (5.98)\end{aligned}$$

From (5.97), (5.98) and (1.5), one obtains that the singular values of $T_{n,g}$ are given by

$$\begin{aligned} \sigma_k\left(T_{n,g}\right) &= \sqrt{n_1\lambda_k}, \qquad k=0,\ldots,n_3n_2-1,\\ \sigma_k\left(T_{n,g}\right) &= 0, \qquad k=n_3n_2,\ldots,n_3n_2n_1-1,\end{aligned}$$

and, since $\sum_{j=0}^{n_1-1} \mathcal{T}_j^*\mathcal{T}_j$ is a positive semidefinite matrix, from (1.5) we can write

$$\begin{aligned} \sigma_k\left(T_{n,g}\right) &= \sqrt{n_1}\widetilde{\sigma}_k, \qquad k=0,\ldots,n_3n_2-1,\\ \sigma_k\left(T_{n,g}\right) &= 0, \qquad k=n_3n_2,\ldots,n_3n_2n_1-1,\end{aligned}$$

where $\widetilde{\sigma}_k$ denotes the generic singular value of $\left(\sum_{j=0}^{n_1-1} \mathcal{T}_j^*\mathcal{T}_j\right)^{\frac{1}{2}}$.

Regarding the distribution in the sense of singular values, by invoking exactly the same argument as in the above example for g-circulant matrix, we deduce that

$$\{T_{n,g}\} \sim_\sigma (0, G),$$

for any domain G satisfying the requirements of Definition 1.7.

Example 5.28. *Let us see what happens when the vector g has only one component different from zero. Let $n=(n_1,n_2,\ldots,n_d)$ and $g=(0,\ldots,0,g_k,0,\ldots,0)$, $g_k>0$; in this case we can give an explicit formula for the singular values of the d-level g-circulant matrix. For convenience and without loss of generality we take $g=(0,\ldots,0,g_d)$ (with all zero components in top positions, otherwise we use a permutation). From Subsection 5.6.1, the singular values of $C_{n,g}=\left[a_{(r-g\circ s)\bmod n}\right]_{r,s=\underline{0}}^{n-e}$ are zero except for few of them given by $\sqrt{\hat{n}\,[0]}\sigma$ where, in our case, $\hat{n}\,[0]=n_1n_2\cdots n_{d-1}$, $n\,[0]=(n_1,n_2,\ldots,n_{d-1})$, and σ is any singular value of the matrix*

$$\left(\sum_{j=\underline{0}}^{n[0]-e} \mathcal{C}_j^*\mathcal{C}_j\right)^{\frac{1}{2}},$$

where $\mathcal{C}_j$ is a g_d-circulant matrix of dimension $n_d\times n_d$ whose expression is

$$\begin{aligned} \mathcal{C}_j=\left[a_{(r-g\circ s)\bmod n}\right]_{r_d,s_d=0}^{n_d-1} &= \left[a_{(r_1,r_2,\ldots,r_{d-1},(r_d-g_ds_d)\bmod n_d)}\right]_{r_d,s_d=0}^{n_d-1}\\ &= \left[a_{(j,(r_d-g_ds_d)\bmod n_d)}\right]_{r_d,s_d=0}^{n_d-1},\end{aligned}$$

with $(r_1,r_2,\ldots,r_{d-1})=j$. For $j=\underline{0},\ldots,n\,[0]-e$, if $C_{n_d}^{(j)}$ is the circulant matrix which has as its first column the vector $a^{(j)}=\left[a_{(j,0)},a_{(j,1)},\ldots,a_{(j,n_d-1)}\right]^\top$ (which is the first column of the matrix $\mathcal{C}_j$), $C_{n_d}^{(j)}=\left[a_{(j,(r-s)\bmod n_d)}\right]_{r,s=0}^{n_d-1}=F_{n_d}D_{n_d}^{(j)}F_{n_d}^*$, with $D_{n_d}^{(j)}=\operatorname{diag}\left(\sqrt{n_d}F_{n_d}^*a^{(j)}\right)$, then,

from (5.47), (5.8), and (5.11), it is immediate to verify that

$$\begin{aligned}\sum_{j=\underline{0}}^{n[0]-e}\mathcal{C}_j^*\mathcal{C}_j &= \sum_{j=\underline{0}}^{n[0]-e}\left(F_{n_d}D_{n_d}^{(j)}F_{n_d}^*Z_{n_d,g_d}\right)^*\left(F_{n_d}D_{n_d}^{(j)}F_{n_d}^*Z_{n_d,g_d}\right)\\ &= \sum_{j=\underline{0}}^{n[0]-e}\left(F_{n_d}^*Z_{n_d,g_d}\right)^*\left(D_{n_d}^{(j)}\right)^*D_{n_d}^{(j)}\left(F_{n_d}^*Z_{n_d,g_d}\right)\\ &= \left(F_{n_d}^*Z_{n_d,g_d}\right)^*\left(\sum_{j=\underline{0}}^{n[0]-e}\left(D_{n_d}^{(j)}\right)^*D_{n_d}^{(j)}\right)\left(F_{n_d}^*Z_{n_d,g_d}\right).\end{aligned}$$

Now, if we put $n_{d,g}=\frac{n_d}{(n_d,g_d)}$ *and*

$$q_s^{(j)} = \left|D_{n_d}^{(j)}\right|_{s,s}^2=\left(D_{n_d}^{(j)}\right)_{s,s}\cdot\overline{\left(D_{n_d}^{(j)}\right)_{s,s}},\quad s=0,1,\ldots,n_d-1,$$

$$\Delta_l = \begin{bmatrix}\sum\limits_{j=\underline{0}}^{n[0]-e}q_{(l-1)n_{d,g}}^{(j)} & & & \\ & \sum\limits_{j=\underline{0}}^{n[0]-e}q_{(l-1)n_{d,g}+1}^{(j)} & & \\ & & \ddots & \\ & & & \sum\limits_{j=\underline{0}}^{n[0]-e}q_{(l-1)n_{d,g}+n_{d,g}-1}^{(j)}\end{bmatrix}\in M_{n_{d,g}}(\mathbb{R}),$$

for $l=1,2,\ldots,(n_d,g_d)$*, then, following the same reasoning employed for proving formula (5.48), we infer*

$$\Lambda\left(\sum_{j=\underline{0}}^{n[0]-e}\mathcal{C}_j^*\mathcal{C}_j\right)=\frac{1}{(n_d,g_d)}\Lambda\left(J_{(n_d,g_d)}\otimes\sum_{l=1}^{(n_d,g_d)}\Delta_l\right),$$

where

$$J_{(n_d,g_d)} = \underbrace{\begin{bmatrix}1&1&\cdots&1\\1&1&\cdots&1\\\vdots&\vdots&\vdots&\vdots\\1&1&\cdots&1\end{bmatrix}}_{(n_d,\,g_d)\ times},$$

$$\frac{1}{(n_d,g_d)}\Lambda\left(J_{(n_d,g_d)}\right) = \{0,1\},$$

and

$$\begin{aligned}\sum_{l=1}^{(n_d,g_d)}\Delta_l &= \sum_{l=1}^{(n_d,g_d)}\operatorname*{diag}_{k=0,\ldots,n_{d,g}-1}\left(\sum_{j=\underline{0}}^{n[0]-e}q_{(l-1)n_{d,g}+k}^{(j)}\right)\\ &= \operatorname*{diag}_{k=0,\ldots,n_{d,g}-1}\left(\sum_{l=1}^{(n_d,g_d)}\sum_{j=\underline{0}}^{n[0]-e}q_{(l-1)n_{d,g}+k}^{(j)}\right).\end{aligned}$$

Consequently, since $\sum\limits_{l=1}^{(n_d,g_d)}\Delta_l$ *is a diagonal matrix, and by exploiting basic properties of the*

tensor product, we find

$$\lambda_k\left(\sum_{j=\underline{0}}^{n[0]-e}\mathcal{C}_j^*\mathcal{C}_j\right) = \sum_{l=1}^{(n_d,g_d)}\sum_{j=\underline{0}}^{n[0]-e} q^{(j)}_{(l-1)n_{d,g}+k}, \quad k=0,1,\ldots,n_{d,g}-1,$$
$$\lambda_k\left(\sum_{j=\underline{0}}^{n[0]-e}\mathcal{C}_j^*\mathcal{C}_j\right) = 0, \qquad k=n_{d,g},\ldots,n_d-1.$$

Now, since $\sum_{j=\underline{0}}^{n[0]-e}\mathcal{C}_j^*\mathcal{C}_j$ *is a positive semidefinite matrix, from (1.5) we finally have*

$$\sigma_k\left(\left(\sum_{j=\underline{0}}^{n[0]-e}\mathcal{C}_j^*\mathcal{C}_j\right)^{\frac{1}{2}}\right) = \sqrt{\sum_{l=1}^{(n_d,g_d)}\sum_{j=\underline{0}}^{n[0]-e} q^{(j)}_{(l-1)n_{d,g}+k}}, \quad k=0,1,\ldots,n_{d,g}-1,$$
$$\sigma_k\left(\left(\sum_{j=\underline{0}}^{n[0]-e}\mathcal{C}_j^*\mathcal{C}_j\right)^{\frac{1}{2}}\right) = 0, \qquad k=n_{d,g},\ldots,n_d-1.$$

Part II

Spectral Distributions of Structured Matrix-Sequences: Applications

Chapter 6

A note on the (regularizing) preconditioning of g-Toeplitz sequences via g-circulants

In Chapter 5 we addressed the problem of characterizing the singular values and the eigenvalues of g-circulants and of providing an asymptotic analysis of the distribution results for the singular values of g-Toeplitz sequences, in the case where the sequence of values $\{a_k\}$, defining the entries of the matrices, can be interpreted as the sequence of Fourier coefficients of an integrable function f over the domain $(-\pi, \pi)$. Such results were plainly generalized the block, multi-level case, amounting to choose the symbol f multivariate, i.e., defined on the set $(-\pi, \pi)^d$ for some $d > 1$, and matrix-valued, i.e., such that $f(x)$ is a matrix of given size $p \times q$.
Here we consider the preconditioning problem. In particular, we consider the general case with $g \geq 2$ and the interesting result is that the preconditioned sequence $\{\mathcal{P}_n\} = \{P_n^{-1} A_n\}$, where $\{P_n\}$ is the sequence of preconditioner, cannot be clustered at 1 so that the case of $g = 1$, widely studied in the literature, is exceptional (see, e.g., [24, 26] for the one-level case, [80] for the multi-level case, and [81] for the multi-level block case). However, in this chapter we will see that while the optimal preconditioning cannot be achieved, the result has a positive implication since there exist choices of g-circulant sequences which are regularizing preconditioning sequence for the corresponding g-Toeplitz structures.

6.1 General tools from preconditioning theory

When preconditioning a spectrally bounded sequence it is compulsory to use a spectrally bounded sequence of preconditioners; otherwise the preconditioned sequence will have necessarily the minimal singular value tending to zero with the size and this is known to spoil the convergence speed of any Krylov like technique (see for instance the classical result of Axelsson, Lindkog [6] in the context of the conjugate gradient). Therefore, if we look at a preconditioned sequence $\{\mathcal{P}_n\} = \{P_n^{-1} A_n\}$, where $\{P_n\}$ is the sequence of preconditioners, such that $\{\mathcal{P}_n - I_n\}$ is clustered at 0, then the difference between the original sequence and the sequence of preconditioners, that is $\{A_n - P_n\}$, should be clustered at zero too. The latter tells us that if the original sequence has a given distribution then, necessarily, the preconditioning sequence has to be chosen with the same distribution. Such key statements and other theoretical tools are given and proven below.

Lemma 6.1. *Consider a sequence $\{A_n\}$, where $A_n \in M_{d_n}(\mathbb{C})$ ($d_k < d_{k+1}$ for each k). Then the following are equivalent.*

- *There exists a sequence $\{D_n\}$ so that $\|A_n - D_n\|_2^2 = o(d_n)$ and $\operatorname{rank}(D_n) = o(d_n)$.*

- *There exists a sequence* $\{D_n\}$ *so that* $\forall p \in [1, +\infty)$ *it holds* $\|A_n - D_n\|_p^p = o(d_n)$, $\operatorname{rank}(D_n) = o(d_n)$.
- *There exist a function* $x(s)$ *such that* $\lim_{s\to 0} x(s) = 0$ *so that* $\forall \varepsilon > 0\ \exists n_\varepsilon \in \mathbb{N}$ *such that* $\forall n \geq n_\varepsilon$ *it holds* $A_n = N_n + R_n$, *with* $\|N_n\| \leq \varepsilon$ *and* $\operatorname{rank}(R_n) \leq x(\varepsilon)\, d_n$.
- *The sequence* $\{A_n\}$ *is clustered at zero (refer to Definition 1.8).*
- *The sequence* $\{A_n\}$ *is spectrally distributed as the identically null function (refer to Definition 1.7).*

Proof. It is a direct check by making a clever use of the singular value decomposition [46]. □

Lemma 6.2. *Consider two sequences* $\{A_n\}$ *and* $\{B_n\}$, *where* $A_n, B_n \in M_{d_n}(\mathbb{C})$ *($d_k < d_{k+1}$ for each k). If there exists a sequence* $\{D_n\}$ *so that* $\|A_n - B_n - D_n\|_2^2 = o(d_n)$ *and* $\operatorname{rank}(D_n) = o(d_n)$, *then the sequence* $\{A_n - B_n\}$ *is spectrally distributed as the identically null function (in the sense of Definition 1.7) and the sequences* $\{A_n\}$ *and* $\{B_n\}$ *are* equally distributed *(in the sense of the last part of Definition 1.7). In addition, if one of the sequences is spectrally distributed as a function then the other sequence possesses the same distribution.*

Proof. By the equivalence Lemma 6.1 we get that $\{A_n - B_n\} \sim_\sigma 0$. The *equal distribution* of the sequences $\{A_n\}$ and $\{B_n\}$ was proved by Tyrtyshnikov [116]. Lastly, if one of the sequences is spectrally distributed as a function then, by definition of *equal distribution*, it is easy to recognize that the other sequence possesses the same distribution. □

Theorem 6.3. *Let* $\{X_n\}$ *and* $\{P_n\}$ *be two sequences of matrices, with* $X_n, P_n \in M_{d_n}(\mathbb{C})$ *($d_k < d_{k+1}$ for each k). Let* $\{I_n\}$, $I_n \in M_{d_n}(\mathbb{R})$, *be the sequence of identity matrices. Suppose that the sequence* $\{X_n\}$ *is* sparsely unbounded, *the matrices* P_n *are all invertible and the sequence* $\{P_n^{-1}X_n - I_n\}$ *is clustered at* 0. *Then* $\{X_n - P_n\} \sim_\sigma 0$ *and the sequences* $\{P_n\}$ *and* $\{X_n\}$ *are* equally distributed. *In addition, if the sequence* $\{X_n\}$ *is distributed as a function then the sequence* $\{P_n\}$ *has the same distribution.*

Finally, if $\{X_n - P_n\} \sim_\sigma 0$ *that is* $\{X_n - P_n\}$ *is clustered at* 0, *then* $\{P_n^{-1}X_n - I_n\}$ *is clustered at* 0, *under the condition that* $\{P_n^{-1}\}$ *is* sparsely unbounded *that is* $\{P_n\}$ *is* sparsely vanishing.

Proof. From the third assumption, by putting $Y_n = X_n - P_n$, we have $\{P_n^{-1}X_n - I_n\} = \{P_n^{-1}Y_n\} \sim_\sigma 0$ (by Lemma 6.1). Therefore, again by invoking Lemma 6.1, there exists a function $\tilde{x}(s)$ such that $\lim_{s\to 0} \tilde{x}(s) = 0$ so that $\forall \varepsilon > 0\ \exists n_\varepsilon \in \mathbb{N}$ such that $\forall n \geq n_\varepsilon$ we have $P_n^{-1}Y_n = \tilde{N}_n + \tilde{R}_n$, with

$$\left\|\tilde{N}_n\right\| \leq \frac{\varepsilon}{2}, \tag{6.1}$$

$$\operatorname{rank}\left(\tilde{R}_n\right) \leq x(\varepsilon)\, d_n. \tag{6.2}$$

Consequently an explicit computation implies

$$P_n^{-1}X_n = I_n + \tilde{N}_n + \tilde{R}_n,$$

that is

$$X_n = P_n\left(I_n + \tilde{N}_n\right) + P_n\tilde{R}_n,$$

and finally

$$P_n - X_n = X_nN_n + R_n,$$

with

$$N_n = \left(I_n + \tilde{N}_n\right)^{-1} - I_n, \tag{6.3}$$

$$R_n = -P_n \tilde{R}_n \left(I_n + \tilde{N}_n\right)^{-1}. \tag{6.4}$$

Now relations (6.3) and (6.1), and $\varepsilon < 1$ imply

$$\|N_n\| \leq \varepsilon,$$

while relations (6.4) and (6.2) lead to

$$\operatorname{rank}(R_n) \leq x(\varepsilon)\, d_n.$$

Since the sequence $\{X_n\}$ is *sparsely unbounded* we deduce that $\{X_n N_n\} \sim_\sigma 0$ by virtue of Lemma 2.14 and therefore, by using the third part of Lemma 6.1, we deduce $\{Y_n\} = \{X_n - P_n\} = \{-X_n N_n - R_n\} \sim_\sigma 0$. Furthermore, from the last part of Lemma 6.2, we infer that the sequences $\{X_n\}$ and $\{P_n\}$ are *equally distributed*. Now, if the sequence $\{X_n\}$ is distributed as a function, then the definition of *equally distributed* implies that the sequence $\{P_n\}$ has the same distribution.

For the last part we just observe that $P_n^{-1} X_n - I_n = P_n^{-1}(X_n - P_n)$ so that Lemma 2.14 implies $\{P_n^{-1} X_n - I_n\} \sim_\sigma 0$ if $\{X_n - P_n\} \sim_\sigma 0$ and $\{P_n^{-1}\}$ is *sparsely unbounded* (which is the same as $\{P_n\}$ is *sparsely vanishing* given the invertibility of each P_n and thanks to Lemma 2.11). □

Remark 6.4. *Theorem 6.3 has a "philosophical" meaning. If we think to the matrices P_n as preconditioners, then Theorem 6.3 states that a good preconditioning sequence $\{P_n\}$ inherits from the original sequence $\{X_n\}$ the distribution, if any. Moreover if the sequence $\{X_n\}$ is* sparsely unbounded *(*sparsely vanishing*) then the same is true for the sequence $\{P_n\}$.*

Remark 6.5. *The* sparsely unboundedness *assumption of $\{X_n\}$ is necessary and cannot be removed as far as we are concerned with Theorem 6.3. For instance, take $X_n = (n+1) I_n$ and $P_n = n I_n$. Then the sequence $\{P_n^{-1} X_n - I_n\} = \left\{\frac{I_n}{n}\right\}$ is clustered at 0, but $\{X_n - P_n\} = \{I_n\}$ is not. However $\{X_n\}$ and $\{P_n\}$ have the same distribution function, since they are both distributed as the constant function ∞.*

Theorem 6.6. *Let $\{X_n\}$, $\{Y_n\}$ and $\{P_n\}$ be three sequences of matrices, with $X_n, Y_n, P_n \in M_{d_n}(\mathbb{C})$ ($d_k < d_{k+1}$ for each k), and P_n invertible for any n. Let $\{I_n\}$ be the sequence of identity matrices, $I_n \in M_{d_n}(\mathbb{R})$. Suppose that*

1. *the sequence $\{X_n\}$ is* sparsely vanishing*,*
2. *the sequence $\{X_n - Y_n\}$ is clustered at 0,*
3. *the sequence $\{P_n^{-1} X_n - I_n\}$ is clustered at 0.*

Then the sequence $\{P_n^{-1} Y_n - I_n\}$ is clustered at 0.

Proof. The matrices $P_n^{-1} Y_n - I_n$ can clearly be split as

$$P_n^{-1} Y_n - I_n = \left(P_n^{-1} X_n - I_n\right) + P_n^{-1}(Y_n - X_n), \tag{6.5}$$

where the sequence $\{P_n^{-1} X_n - I_n\}$ is clustered at 0 by assumption *3.* Moreover the sequence $\{P_n\}$ is *sparsely vanishing* since the sequence $\{X_n\}$ is *sparsely vanishing* (see Remark 6.4). Therefore the application of Lemmas 2.11 and 2.14 proves that the sequence $\{P_n^{-1}(Y_n - X_n)\}$ is clustered at 0. As a final statement, in the light of equation (6.5), the sequence $\{P_n^{-1} Y_n - I_n\}$ is expressed as the sum of two matrix-sequences that are clustered at 0, so that the proof is concluded, by invoking Proposition 2.4 with $\theta = 0$. □

6.2 Preconditioning of g-Toeplitz sequences via g-circulant sequences

We start by analyzing the possibility of a standard preconditioning in the light of the distribution results of Chapter 5 and of the analysis of Section 6.1. Then we consider the preconditioning in a regularizing context.

6.2.1 Consequences of the distribution results on precond. of g-Toeplitz sequences

We study the possibility of a standard preconditioning taking into consideration the distribution results of Chapter 5 and of the analysis of Section 6.1.

First of all Theorem 6.3 tells one that $\{P_n\}$ is a good preconditioning sequence for $\{X_n\}$ (that is $\{P_n^{-1}X_n - I_n\} \sim_\sigma 0$) if and only if $\{X_n - P_n\} \sim_\sigma 0$ and $\{P_n\}$ *sparsely vanishing*, with the matrices P_n all invertible. The consequences below are of paramount importance:

- The vector g has to be strictly positive; if not the original problem $T_{n,g}\mathbf{x} = \mathbf{b}$ is substantially ill-posed since $\{T_{n,g}\} \sim_\sigma 0$ (see the end of Section 5.5) and in addition $C_{n,g}$ is singular and indeed $\{C_{n,g}\} \sim_\sigma 0$ which violates the crucial condition of Theorem 6.3 that $\{P_n\}$ is *sparsely vanishing* with $P_n = C_{n,g}$.
- Even in the case that g is strictly positive, relations (5.86), (5.87), and (5.88) imply that $\{X_n\}$ with $X_n = T_{n,g}$ is *sparsely vanishing* if and only if f is *sparsely vanishing* and $g_i = 1$ (or more generally $g_i = \pm 1$), $i = 1, \ldots, d$. In other words, again by Theorem 6.3, a good preconditioning can be achieved only in standard case of multi-level Toeplitz sequences and in fact the latter is a case widely studied in the literature [24, 26, 80] (for $d = 1$ also with strong clustering when f is continuous [80], while for $d > 1$ the clustering is necessarily weak due to the computational barrier proven in [97]).
- In any case the condition required by Theorem 6.3 that the sequences $\{X_n\}$, $\{P_n\}$, with $X_n = T_{n,g}, P_n = C_{n,g}$, share the same distribution symbol is quite tricky. By comparing (5.86), (5.87), (5.88) and (5.82), (5.83), (5.84), the latter is possible only for the case where $g_i = \gcd(n_i, g_i)$, $i = 1, \ldots, d$, and we have to choose $p = \frac{1}{\tilde{g}} f$.

In conclusion, a good preconditioning can be reached only in the standard multi-level Toeplitz setting. However, by looking at the preconditioning in a different sense, something useful can be said.

6.2.2 Regularizing preconditioning

Suppose that $\{X_n\}$ is a sequence of matrices with $X_n \in M_{d_n}(\mathbb{C})$ ($d_k < d_{k+1}$ for each k) and there exists a sequence of subspaces $\{\mathcal{S}_n\}$ of size r_n being the integer part of cd_n, $c \in (0, 1)$ for which $\forall \epsilon > 0$, $\exists n_\epsilon$ and

$$\|X_n \mathbf{v}\|_2 \leq \epsilon \|\mathbf{v}\|_2, \qquad \forall \mathbf{v} \in \mathcal{S}_n, \forall n \geq n_\epsilon.$$

This situation naturally arises when $\{X_n\} \sim_\sigma (\theta, G)$ with θ vanishing on $\hat{G} \subset G$ with $\frac{m\{\hat{G}\}}{m\{G\}} = c$, $m\{\cdot\}$ being the Lebesgue measure and $|\theta| > 0$ almost everywhere in the complement $G \backslash \hat{G}$. Under such conditions we look for a preconditioning $\{J_n\}$ already in inverse form such that

$$\|J_n X_n \mathbf{v}\|_2 \leq \epsilon \|\mathbf{v}\|_2, \qquad \forall \mathbf{v} \in \mathcal{S}_n, \forall n \geq \tilde{n}_\epsilon,$$

$$\|J_n X_n \mathbf{v} - \mathbf{v}\|_2 \leq \epsilon \|\mathbf{v}\|_2, \qquad \forall \mathbf{v} \in \mathcal{S}_n^\perp, \forall n \geq \tilde{n}_\epsilon.$$

In other words $J_n X_n$ when restricted to $\mathcal{S}_n$ is close to the null matrix, while it is close to the identity matrix in the orthogonal complement. These conditions, amounting in writing that $J_n X_n$ is an ϵ-perturbation of

$$\left[\begin{array}{c|c} I_{r_n} & 0 \\ \hline 0 & 0 \end{array}\right],$$

will be verified in the subsequent Subsection 6.2.3.

6.2.3 The analysis of regularizing preconditioners when $p = q = d = 1$ and n chosen such that $(n, g) = 1$

According to the very concise analysis in Subsection 6.2.2, we will prove that a proper choice of the matrix-sequence $\{C_{n,g}\}$ leads to a satisfactory regularizing preconditioning for $\{T_{n,g}\}$, at least when the entries of $T_{n,g}$ comes from the Fourier coefficients of a sparsely vanishing function f.

Theorem 6.7. *Let $\{T_{n,g}\}$ be a sequence of g-Toeplitz matrices generated by a sparsely vanishing function $f \in L^1(Q)$, $Q = (-\pi, \pi)$, then the sequence $\left\{C_{n,g}^{-1}\right\}$, where $\{C_{n,g}\} = \{C_n Z_{n,g}\}$, C_n is the Frobenius distance minimizer of T_n in the standard circulant algebra and $Z_{n,g}$ defined as in (5.9), is a regularizing preconditioning for $\{T_{n,g}\}$.*

Proof. If we denote by T_n the classical Toeplitz matrix

$$T_n = [a_{r-c}]_{r,c=0}^{n-1},$$

where the elements a_j are the Fourier coefficients of the sparsely vanishing function f in $L^1(Q)$, and by $T_{n,g}$ the g-Toeplitz matrix generated by the same function

$$T_{n,g} = [a_{r-gc}]_{r,c=0}^{n-1}, \tag{6.6}$$

where the quantities $r - gc$ are not reduced modulus n. From (5.63) we have that

$$T_{n,g} = T_n \left[\begin{array}{c|c} \widehat{Z}_{n,g} & 0 \end{array}\right] + \left[\begin{array}{c|c} 0 & \widetilde{T}_{n,g} \end{array}\right],$$

where $\widetilde{T}_{n,g} \in M_{n,(n-\mu_g)}(\mathbb{C})$ is the matrix $T_{n,g}$ defined in (6.6) by considering only the $n - \mu_g$ last columns, and $\widehat{Z}_{n,g} \in M_{n,\mu_g}(\mathbb{R})$ is the matrix defined in (5.9) by considering only the μ_g first columns.

The g-circulant matrix $C_{n,g}$ is defined as follows

$$\begin{aligned} C_{n,g} &= \left[a_{(r-gc) \bmod n}\right]_{r,c=0}^{n-1} \\ &= C_n Z_{n,g}, \end{aligned}$$

where C_n is the classical circulant matrix generated from elements of the first column of $C_{n,g}$ and $Z_{n,g}$ is defined as in (5.9) and we suppose that C_n is non-singular and $(n, g) = \gcd(n, g) = 1$, so that $Z_{n,g}$ is a permutation matrix (see Lemma 5.3). If we choose $\{C_n\}$ with C_n the Frobenius distance minimizer of T_n in the standard circulant algebra (the one proposed by Tony Chan in the one-level setting [26]), by the analysis in [91], for $f \in L^1\left(Q^d\right)$, we have $\{C_n\} \sim_\sigma \left(f, Q^d\right)$ so that $\{C_{n,g}\} \sim_\sigma \left(f, Q^d\right)$ whenever $(n_i, g_i) = 1$, $i = 1, \ldots, d$, because $Z_{n,g}$ is a permutation matrix (here we are for the moment interested only in the case where $d = 1$).

Now we consider the product $C_{n,g}^{-1} T_{n,g}$; from (6.7) and since $Z_{n,g}^\top Z_{n,g} = I_n$ we have that

$$C_{n,g}^{-1} T_{n,g} = Z_{n,g}^\top C_n^{-1} T_n \left[\widehat{Z}_{n,g} | 0\right] + C_{n,g}^{-1} \left[0 | \widetilde{T}_{n,g}\right],$$

and, since $\{C_n^{-1}T_n\} \sim_\sigma 1$ or more precisely if $\{C_n^{-1}T_n - I_n\} \sim_\sigma 0$, i.e.,

$$C_n^{-1}T_n = I_n + E_n, \qquad \text{with} \qquad \{E_n\} \sim_\sigma 0,$$

we obtain

$$\begin{aligned} C_{n,g}^{-1}T_{n,g} &= Z_{n,g}^\top C_n^{-1}T_n \left[\widehat{Z}_{n,g}|0\right] + C_{n,g}^{-1}\left[0|\widetilde{T}_{n,g}\right] \\ &= Z_{n,g}^\top \left[I_n + E_n\right]\left[\widehat{Z}_{n,g}|0\right] + C_{n,g}^{-1}\left[0|\widetilde{T}_{n,g}\right] \\ &= Z_{n,g}^\top \left[\widehat{Z}_{n,g}|0\right] + Z_{n,g}^\top E_n \left[\widehat{Z}_{n,g}|0\right] + C_{n,g}^{-1}\left[0|\widetilde{T}_{n,g}\right] \\ &= \left[\begin{array}{c|c} I_{\mu_g} & 0 \\ \hline 0 & 0 \end{array}\right] + Z_{n,g}^\top E_n \left[\widehat{Z}_{n,g}|0\right] + C_{n,g}^{-1}\left[0|\widetilde{T}_{n,g}\right]. \end{aligned}$$

From Lemma 2.6, since $\left\|Z_{n,g}^\top\right\| = 1$ and $\left\|\left[\widehat{Z}_{n,g}|0\right]\right\| = 1$ (indeed the first is a permutation matrix and the second is an "incomplete" permutation matrix), and since $\{E_n\} \sim_\sigma 0$, we infer that $\left\{Z_{n,g}^\top E_n \left[\widehat{Z}_{n,g}|0\right]\right\} \sim_\sigma 0$. Moreover, since $\{C_n\}, \{C_{n,g}\} \sim_\sigma (f, Q)$ with f sparsely vanishing, from Proposition 2.7 and Lemma 2.11 we have that $\left\{C_{n,g}^{-1}\right\}$ is sparsely unbounded. Finally, since in Proposition 5.20 it was shown that

$$\left\{\left[0|\widetilde{T}_{n,g}\right]\right\} \sim_\sigma (0, Q),$$

from Lemma 2.14, we deduce that $\left\{C_{n,g}^{-1}\left[0|\widetilde{T}_{n,g}\right]\right\} \sim_\sigma 0$ and the proof is concluded. □

Remark 6.8. *In Theorem 6.7 any preconditioning sequence $\{C_n\}$ for which $\{C_n^{-1}T_n - I_n\} \sim_\sigma 0$ will lead to a preconditioning sequence $\{C_{n,g}\}$ with regularizing features. In other words the choice of the Frobenius optimal preconditioners is just a possible example.*

6.3 Generalizations

With all the constraints of Subsection 6.2.3 we can allow to have $d > 1$ that is $n = (n_1, \ldots, n_d)$ sequence of integer positive vectors with $(n_i, g_i) = 1$, $i = 1, \ldots, d$, so that $Z_{n,g}$ is still a permutation matrix. The proof reported in Subsection 6.2.3 is identical with the only observation that the cluster of $\left\{C_{n,g}^{-1}T_{n,g} - I_n\right\}$ is weak and not strong, due to the computational barrier proven in [97]. More precisely, under the assumption of positivity and continuity of $|f|$, by using the Korovkin Theory [80] and the Tony Chan preconditioners, we find that the number of outliers of $\left\{C_{n,g}^{-1}T_{n,g} - I_n\right\}$ grows asymptotically as $\hat{n}\left(\sum_{j=1}^{d} n_j\right)$, $\hat{n} = n_1 n_2 \cdots n_d$. Moreover the weak clustering can be achieved by using the mild assumption that f is only Lebesgue integrable and sparsely vanishing (see [91]).

Furthermore, by following the approach in [81], nothing changes if we assume that the multilevel setting is accompanied by the block setting i.e. $p + q \geq 3$ (somehow the only condition is the recourse to the Moore-Penrose inverse when $p \neq q$).

A bit trickier is the case where the assumption $(n_i, g_i) = 1$, $i = 1, \ldots, d$, is dropped. In that case $C_{n,g} = C_n Z_{n,g}$ is inherently singular due to the singularity of $Z_{n,g}$ whose rank is $\hat{n}_g = \frac{n_1}{(n_1,g_1)}\frac{n_2}{(n_2,g_2)}\cdots\frac{n_d}{(n_d,g_d)}$ with $\mu_g \leq n_g < n$, $\mu_g = \left(\left\lceil\frac{n_1}{g_1}\right\rceil, \cdots, \left\lceil\frac{n_d}{g_d}\right\rceil\right)$ (see Lemma 5.3, where all the objects $n, g, \mu_g, n_g, (n, g)$ are d-dimensional vectors of positive integers and the inequalities are componentwise). In this case a good preconditioner already in inverse form is

$$J_n = Z_{n,g}^\top C_n^{-1},$$

with C_n the usual Tony Chan preconditioner (refer to Subsection 6.2.2). Since $\mu_g \leq n_g < n$ (because $1 < (n,g) \leq g$) by Lemma 5.3 we find

$$\widetilde{Z}_{n,g}^{\top} \widehat{Z}_{n,g} = \left[\frac{I_{\mu_g}}{0} \right].$$

As a consequence the proof given in Subsection 6.2.3 is the same and the final result is identical: for the sake of completeness we only observe that the term $C_{n,g}^{-1}$ is always replaced by $Z_{n,g}^{\top} C_n^{-1}$ so that $\left\{ C_n^{-1} \left[0 | \widetilde{T}_{n,g} \right] \right\} \sim_\sigma 0$ because $\{C_n\} \sim_\sigma \left(f, Q^d \right)$ with f sparsely vanishing and $\left\{ \left[0 | \widetilde{T}_{n,g} \right] \right\} \sim_\sigma 0$ (see Proposition 2.7, Lemma 2.11, Lemma 2.14 and (5.73)), and finally $\left\{ Z_{n,g}^{\top} C_n^{-1} \left[0 | \widetilde{T}_{n,g} \right] \right\} \sim_\sigma 0$ because of Proposition 2.4, where $Z_{n,g}^{\top}$ plays the role of Q_n and $C_n^{-1} \left[0 | \widetilde{T}_{n,g} \right]$ plays the role of B_n.

Finally we observe that we have emphasized the role of the Frobenius optimal preconditioner proposed by Tony Chan, for which a very general and rich clustering analysis is available thanks to the Korovkin Theory. However, any other alternative and successful preconditioner for standard Toeplitz structures can be employed thanks to Theorem 6.6, which states a kind of useful transitive property.

6.4 Numerical experiments

Aimed of providing numerical evidences to the theoretical results of the previous section, now we analyze

(i) the distribution of the singular values of g-Toeplitz matrices and related g-circulant preconditioned matrices (Subsection 6.4.1),

(ii) the effectiveness of the g-circulant preconditioning for the solution of the corresponding g-Toeplitz linear system $Ax = b$ (Subsection 6.4.2), and

(iii) a possible real application related to a 2D inverse problem in imaging (Subsection 6.4.3).

In particular, for the first two points (i) and (ii) we consider six well-known test cases, most of them firstly used in pioneer works by G. Strang, T. Chan and E. Tyrtyshnikov for the classical Toeplitz preconditioning (i.e., $g = 1$). For each of any considered test, we report the elements of the first column $A_{1,k}$ for $k = 1, \ldots, n$, and some basic properties of the corresponding basic Toeplitz matrix ($g = 1$).

- Test 1 $\quad A_{k,1} = k^{-1}$
  Strictly positive non-Wiener generating function, Well-conditioned [106, 26].
- Test 2 $\quad A_{k,1} = k^{-2}$
  Strictly positive Wiener generating function, Well-conditioned [106, 26].
- Test 3 $\quad A_{:,1} = (2, -1, 0, \ldots, 0)^{\top}$
  Sparsely vanishing trigonometric polynomial generating function $f(x) = 2 - 2\cos x$, Ill-conditioned, Zero valued (order 2) at the origin [115].
- Test 4 $\quad A_{:,1} = (20, -15, 6, -1, 0, \ldots, 0)^{\top}$
  Sparsely vanishing trigonometric polynomial generating function $f(x) = (2 - 2\cos x)^3$, Ill-conditioned, Zero valued (order 6) at the origin.
- Test 5 $\quad A_{:,1} = \left(\frac{\pi^2}{2}, -2, 0, -\frac{2}{9}, 0, -\frac{2}{25}, 0, \ldots, 0, -\frac{2}{(k-1)^2}, 0, \ldots \right)^{\top}$
  Sparsely vanishing generating function $f(x) = \pi |x|$, Ill-conditioned, Zero valued (order 1) at the origin [38].

- Test 6 $A_{:,1} = \left(2, 0, 2\frac{1}{3}, 0, -2\frac{1}{15}, 0, 2\frac{1}{35}, 0, \ldots, 0, (-1)^{\frac{k+1}{2}} \frac{2}{(k-1)^2-1}, 0, \ldots\right)^\top$
  Sparsely vanishing generating function $f(x) = \pi |\cos x|$, Ill-conditioned, Zero valued (order 1) at π [38].

We notice that the generating function f is strictly positive in the two (well-conditioned) test cases 1 and 2, and sparsely vanishing in the four (ill-conditioned) test cases 3,4,5 and 6.

According to Subsection 6.2.2, for any g-Toeplitz test matrix we consider both (i) the Natural g-circulant preconditioner and (ii) the Optimal g-circulant preconditioner (see [106, 26] for the classical Toeplitz case $g = 1$). The numerical test have been developed with MatLab, and the singular value decomposition has been computed by the built-in MatLab function svd().

6.4.1 The distribution of the singular values

First, we plot the distribution of the singular values of the g-Toeplitz matrix $A \in M_n(\mathbb{C})$, the corresponding g-circulant preconditioner P, and the preconditioned matrix $P^{-1}A$, for $n = 1000$ and $g = 2, 3, 7$ (n and g are coprime for $g = 3$ and $g = 7$, and are not coprime for $g = 2$). In particular, we have:

I) Fig. 6.2 and Fig. 6.4 show the singular values of the g-Toeplitz matrices A, the Natural (top) and Optimal (bottom) g-circulant preconditioners P and the corresponding preconditioned matrices $P^{-1}A$ in the coprime cases, respectively for $g = 3$ and $g = 7$;

II) Fig. 6.6 shows the singular values of the optimal preconditioning in the non-coprime case $g = 2$, for two test examples (Test 1 and Test 5).

Before dealing with the preconditioned matrix $P^{-1}A$, it is quite interesting to notice that the plotted distribution of the singular values of the g-Toeplitz matrix A and its g-circulant preconditioner P "exactly" agrees with the corresponding expected distributions (5.86)-(5.87)-(5.88) and (5.82)-(5.83)-(5.84). Indeed, for $g > 1$ and sparsely vanishing generating functions, we have:

(i) regarding the g-Toeplitz matrix A, the first $\frac{n}{g}$ singular values are positive, and equals to $\sqrt{\widehat{|f|^{(2)}}(x)}$, while the remaining $n - \frac{n}{g}$ are null, as stated by (5.87) (see the blue line in Figg. 6.2, 6.4 and 6.6);

(ii) regarding the g-circulant preconditioner P, by introducing the positive integer value $\gamma = \gcd(n, g)$, if $\gamma = 1$ then the singular values are bounded away from zero or sparsely vanishing as well as the generating function is (see the green line in Figg. 6.2 and 6.4), and, if $\gamma > 1$, the first $\frac{n}{\gamma}$ singular values are always bounded away from zero (regardless the sparsely vanishing generating function is or is not bounded away from zero), and equals to $\sqrt{\widehat{|p|^{(2)}}(x)}$, while the remaining $n - \frac{n}{\gamma}$ are null, as stated by (5.83) (see the green line in Fig. 6.6).

In particular, since $n = 1000$, then $\gamma = 1$ for $g = 3, 7$, and $\gamma = 2$ for $g = 2$: in Fig. 6.2 and Fig. 6.4 the singular values of the for both the natural and optimal g-circulant preconditioners are bounded away from zero in the ill-conditioned test cases 1 and 2, and sparsely vanishing in the well-conditioned test cases 3,4,5 and 6, while one half of the singular values are always null in Fig. 6.6 (green lines).

It is now interesting to analyze the distribution of the preconditioned matrix.

Any coprime case (Figg. 6.2 and 6.4, red line) gives rise to a good clustering at unity, in the first $\frac{n}{g}$ singular values, while the remaining ones are null. This is a results which was expected in the light of Theorem 6.7: the preconditioned matrix $P^{-1}A$ guarantees a good

clustering in a subspace which is the most large possible (remember that the rank of A is $\frac{n}{g}$, so that the rank of $P^{-1}A$ is just no larger than $\frac{n}{g}$). This good clustering at unity of the preconditioned matrix $P^{-1}A$ occurs in both the well-conditioned case (see, in Figg. 6.2 and 6.4. the cases Test1 and Test 2, where the preconditioners have no vanishing singular values) and the ill-conditioned case (see, again in Figg. 6.2 and 6.4, the cases from Test 3 to Test 6, where the preconditioners have always a zero measure vanishing singular subspace). We can also observe that the singular values' distributions of the natural preconditioned matrix and the optimal preconditioned matrix are very similar. This agrees with the classical and widely studied Toeplitz case (i.e., $g = 1$), where both the distributions tend to the generating function, as n grows. However, as expected, the optimal preconditioner, which is the closest-to-A g-circulant matrix with respect to the Frobenius distance, gives a bit better clustering than the natural one: as instance, see in particular the clustering at unity of Test 3 in the optimal preconditioning (bottom) and in the natural preconditioning (top) in Figg. 6.2 and 6.4.

The situation is different in the non-coprime case, as Fig. 6.6 shows. Before going on, according to Subsection 5.6.1, we mention that in this case instead of the inverse P^{-1} we have to consider the Moore-Penrose generalized inverse $P^{\dagger}$, being P a singular matrix. Due to the non-coprime circularity, now the g-circulant preconditioner has a lot of cyclically repeated, hence linearly dependent, columns. Heuristically, the g-circulant preconditioner P "loose" a lot of information which were contained in the related g-Toeplitz matrix A, which means that P become less correlate to A, and a good clustering is no more possible. This is well explained by Fig. 6.6, red line, where just a couple of test examples are reported (all the others behave similarly, so they are not reported). In particular, in the first two columns we can see that the singular values of the preconditioned matrix $P^{\dagger}A$ are not clustered (moreover they tend to diverge, giving rise to high instability in real applications). To avoid such an amplification, instead of using $P^{\dagger}$ for the preconditioned matrix, we can consider a regularized version $P_{\alpha}^{\dagger}$ of $P^{\dagger}$, where the singular values of P smaller than a regularization parameter $\alpha > 0$ are not inverted. As very first attempt, we plot the singular values of the preconditioned matrix $P_{\alpha}^{\dagger}A$, being $\alpha = 10^{-12}\,\|P\|$. As we can notice, a good clustering is found also for the non-coprime case.

6.4.2 The preconditioning effectiveness

In this subsection we give a first evaluation of the behavior of the optimal g-circulant preconditioning for the solution of the g-Toeplitz linear system $Ax = b$, with $g = 3 > 1$. First of all we mention that, since the square g-Toeplitz matrix A has no full rank (recall that here $g > 1$), we necessarily have to consider the least square solution $A^*Ax = A^*b$. Accordingly, we consider the solution of the linear system by means of the (P)CGNR method, that is, the (preconditioned) conjugate gradient method applied on the normal equations.

In order to evaluate the restoration errors, we choose the true data vector x, ad then we compute the known data b simply as $b = Ax$. In particular, we consider a true data vector x whose (i)th component, for $i = 0, \ldots, n$, is given by $\cos\left(\frac{g\pi i}{n}\right)$, so that the first $\frac{n}{g}$ values of the true data are a sampling on a uniform grid of an entire semi-period of the cosine function.

Let x_k be the (k)th iteration of the (P)CGNR method. We compute the relative residual error $RelRes = \frac{\|A^*Ax_k - A^*b\|_2}{\|b\|_2}$ and the relative error on the restored signal $RelErr = \frac{\|x_k - x^{\dagger}\|_2}{\|x^{\dagger}\|_2}$, where $x^{\dagger}$ is the projection on $N(A)^{\perp}$ of the true data (which is obviously the best possible restoration). Since $\frac{n}{g}$ is the rank of A, to obtain $x^{\dagger}$ we compute $x^{\dagger} = \tilde{V}\tilde{V}^*x$, where $\tilde{V} \in M_{n,\frac{n}{g}}(\mathbb{C})$ is the matrix given by the first $\frac{n}{g}$ columns of V, being V the orthogonal matrix of the singular value decomposition $A = U\Sigma V^*$.

By using the built-in MatLab function pcg() within the first 100 iterations, in Table 6.1 the numerical results related to three different levels of noise on the data b are reported. In particular, by denoting as $b_{\eta} = b + \eta$ the noisy data, where η is a white Gaussian noise, we

have the following test cases: in the left columns the data b is noiseless; in the central columns the relative noise level $\frac{\|b_\eta - b\|_2}{\|b\|_2}$ is $10^{-4}\%$; in the central columns the relative noise level $\frac{\|b_\eta - b\|_2}{\|b\|_2}$ is $10^{-1}\%$.

As we can observe, the optimal g-circulant preconditioned conjugate gradient method does not allow to obtain better results then the classical (i.e., "unpreconditioned") method. This fact has been already observed for the Toeplitz case (i.e., $g = 1$), and we can say that now, for g-Toeplitz linear system with $g > 1$, this phenomenon is amplified.

Indeed, most preconditioners for Toeplitz systems with high clustering of the singular values such as the natural and optimal ones give rise to instability and noise amplification. In Fig. 6.7 we plot the first $\frac{n}{g}$ values (i.e., the significative ones) of the restored signals for both the CGNR and PCGNR algorithms ($g = 3$, Test 4, 1% of data noise): As we can see, here the noise amplification of the preconditioned case is higher, and hence some oscillations occur. However, to improve the results (that is, speed up the convergence, without amplifying the instability due to noise or floating point computation), a wide range of regularization techniques can be added to the preconditioners (see [37], for the classical Toeplitz case), and future works will be devoted to this analysis. In this direction, the g-circulant preconditioner can be considered as a basic tool for introducing regularization features, which could provide both speed-up and stability to the PCG method.

6.4.3 Two dimensional g-Toeplitz matrices for structured shift-variant image deblurring

We conclude the numerical section by introducing a real problem of image deblurring [14] which is related to g-Toeplitz matrices. Basically, a blurring model (i.e., the forward model) involves a Fredholm linear operator of the first kind as follows. A blurred version $g \in L^2(\mathbb{R}^2)$ of a true image $f \in L^2(\mathbb{R}^2)$ is given by

$$g(x) = \int_{\mathbb{R}^2} h(x,u) f(u)\, \mathrm{d}u, \tag{6.7}$$

where the integral kernel $h \in L^2(\mathbb{R}^{2\times 2})$ is the known impulse response of the blurring system, also called point spread function (PSF), being $x = (x_1, x_2)$ and $u = (u_1, u_2)$ the system coordinates of the blurred image g and the true image f. Image deblurring is the (inverse) problem of finding (an approximation of) the true data f (i.e., the cause) by means the knowledge of the blurred data g (i.e., the effect).

The value $h(x,u)$ represents the weight of the true image f at point u in the formation of the blurred image g at point x. This way, $g(x)$ is the average on $\mathbb{R}^2$ of the values of f with respect the weights $h(x,\cdot)$. Among the proposed mathematical models, the simplest and most common blurring operator (6.7) involves the so-called *shift-invariant* integral kernel, in which the weight $h(x,u)$ depends only on the relative position of u with respect to x, that is, there exists a function $h_I \in L^2(\mathbb{R})$, of one variable, such that

$$h(x,u) = h_I(x-u).$$

In a shift-invariant blurring system like that, the impulse response does not change as the object position is shifted, which means that exactly the same blur arises all over the image domain $\mathbb{R}^2$. In this case the blurring operator (6.7) becomes a simple convolution, and its discretization gives rise to (classical) Toeplitz matrices. On the other hand, shift-invariant mathematical models are often only basic approximations of real shift-variant imaging systems. Among all the shift-variant imaging systems, we are interested in the ones which are intrinsically shift-invariant as follows: there exist two "coordinate transformations" $b = b(x)$ and $c = c(u)$ such

$$h(x,u) = \tilde{h}_I(b(x), c(u)),$$

Noise level	0 (No noise)		0.0001%		1%	
Preconditioning	Prec.	No prec.	Prec.	No prec.	Prec.	No prec.
Test 1						
Iter. Number	92	48	94	77	96	92
Relative Residual	3.19e-007	6.05e-010	3.16e-007	1.64e-009	1.24e-007	1.32e-007
Relative Error	4.29e-006	1.20e-008	1.89e-005	1.88e-005	1.83e-002	1.83e-002
Test 2						
Iter. Number	47	13	47	36	78	79
Relative Residual	2.87e-010	2.89e-010	3.48e-010	8.94e-010	9.08e-009	7.43e-009
Relative Error	4.64e-010	4.72e-010	8.75e-006	8.75e-006	8.80e-003	8.80e-003
Test 3						
Iter. Number	100*	2	100*	2	100*	2
Relative Residual	2.14e-002	1.10e-016	2.13e-002	1.12e-016	2.13e-002	1.19e-016
Relative Error	1.81e-002	1.19e-016	1.81e-002	2.98e-006	1.83e-002	3.10e-003
Test 4						
Iter. Number	100*	14	100*	20	100*	20
Relative Residual	1.93e-005	3.78e-010	1.93e-005	8.39e-010	2.34e-005	9.50e-010
Relative Error	7.48e-006	2.40e-010	7.62e-006	2.20e-006	2.10e-003	2.10e-003
Test 5						
Iter. Number	100*	8	100*	37	98	99
Relative Residual	6.08e-005	1.04e-010	5.97e-005	9.93e-010	8.95e-005	1.34e-008
Relative Error	3.22e-005	8.93e-011	3.17e-005	2.76e-006	2.70e-003	2.70e-003
Test 6						
Iter. Number	77	4	75	10	76	72
Relative Residual	1.92e-004	2.64e-013	1.93e-004	6.14e-010	2.09e-004	7.89e-010
Relative Error	1.04e-004	2.65e-013	1.05e-004	7.57e-006	7.60e-003	7.60e-003

Table 6.1: $g = 3$: Best relative residual $\frac{\|A^*Ax_k - A^*b_\eta\|_2}{\|A^*b_\eta\|_2}$, with corresponding iteration number k and relative restoration error $\frac{\left\|x_k - x^\dagger\right\|_2}{\left\|x^\dagger\right\|_2}$, with respect to different noise levels $\delta = \frac{\|b - b_\eta\|_2}{\|b\|_2}$ of the CGNR and PCGNR with optimal g-circulant preconditioner.

where $\tilde{h}_I(b,c) = h_I(b-c)$ is a shift-invariant PSF. Indeed, in some cases the discretization of these models leads to two-levels g-Toeplitz matrices. We have

$$g(x) = \int_U h(x,u) f(u)\,\mathrm{d}u = \int_U h_I(b(x) - c(u)) f(u)\,\mathrm{d}u,$$

that is

$$\tilde{g}(\tilde{x}) = \int_{c(U)} h_I(\tilde{x} - \tilde{u}) \tilde{f}(\tilde{u})\,\mathrm{d}\tilde{u}, \tag{6.8}$$

where $\tilde{x} = b(x)$, $\tilde{u} = c(u)$, $\tilde{g} = g \circ b^{-1}$, $\tilde{f} = \left(c^{-1}\right)' \cdot f \circ c^{-1}$. Here the symbols $\circ$ and $\cdot$ denote respectively the composition and the point-wise function products. In practice, by using such these two coordinate transformations b and c in both the blurred image g and true object f, we obtain that the imaging system becomes explicitly shift-invariant, since it is modeled by the shift-invariant PSF h_I of (6.8). The main example is the rotational blur, generated when a moving object rotates with respect to the imaging apparatus. In this case, although the blur changes with respect the object position (in particular, it is small close to and increases far from the center of the rotation), the blurring is intrinsically shift-invariant. If the coordinate systems are changed from Cartesian $x = (x_1, x_2)$ and $u = (u_1, u_2)$ to Polar system (ρ_x, θ_x) and (ρ_u, θ_u), the PSF becomes explicitly shift-invariant. As instance, concerning a blur of uniform circular motion, we have $h(x,u) = h((\rho_x, \theta_x), (\rho_u, \theta_u)) = h_I(\rho_x - \rho_u, \theta_x - \theta_u)$, with $h_I(\rho, \theta) = \frac{1}{\sigma}$ for $(\rho, \theta) \in \{0 \times [0, \sigma]\}$ and 0 elsewhere, being σ the whole angle of the considered rotation.

In the simplest case where the coordinate transformation are linear functions such as $b(x) = vx$ and $c(u) = gu$, with v and g two integer values. With a fixed discretization step d, we have that

$$(A)_{i,j} = h(id, jd) = h_I(b(id) - c(jd)) = h_I(ivd - jgd)\ .$$

If $b(x) = x$, then the PSF matrix A is a g-Toeplitz matrices. However, in general, we have to consider (g,v)-Toeplitz matrices, that is, matrices which obey the rule $A_n = [a_{vr-gs}]_{r,s=0}^{n-1}$, which are simple generations of g-Toeplitz matrices. By recalling that any 3D geometric projectivity is a linear transformation, we have that such (g,v)-Toeplitz matrices arise in many imaging systems related to large scenes, where the projective geometry becomes important due to perspective. As instance (g,v)-Toeplitz blur matrices arise when some objects are moving with approximately the same speed in a plane which is not parallel to the image plane of the imaging apparatus (this is usually called as "non-perpendicular imaging system geometry", see Fig. 6.8). We remark that this is the classical scenario of high-way traffic flow control systems.

A numerical simulation is shown in Fig. 6.9, where a structured shift-variant blurred image related to a synthetic homography (i.e., a projectivity between two planes) has been used (see the shift-invariant blur which corrupts the image on the left). Since a homography is a liner transformation with respect to the homogeny coordinates, the discretization gives rise to two-level (g,v)-Toeplitz matrices. By using the involved algebraic structure, the deblurring process can be done within $O(n^2 \log n)$ as in the classical convolutive (i.e. Toeplitz) case. In Fig. 6.9, center, we show the projectivity under which the blur becomes shift-invariant, which is modeled by a linear transformation of coordinates (see that the same blur all over the domain of the image on the center). By means of such a shift-invariant blurred projected image, we can obtain the deblurred image (left image), by using $O(n^2 \log n)$ computation.

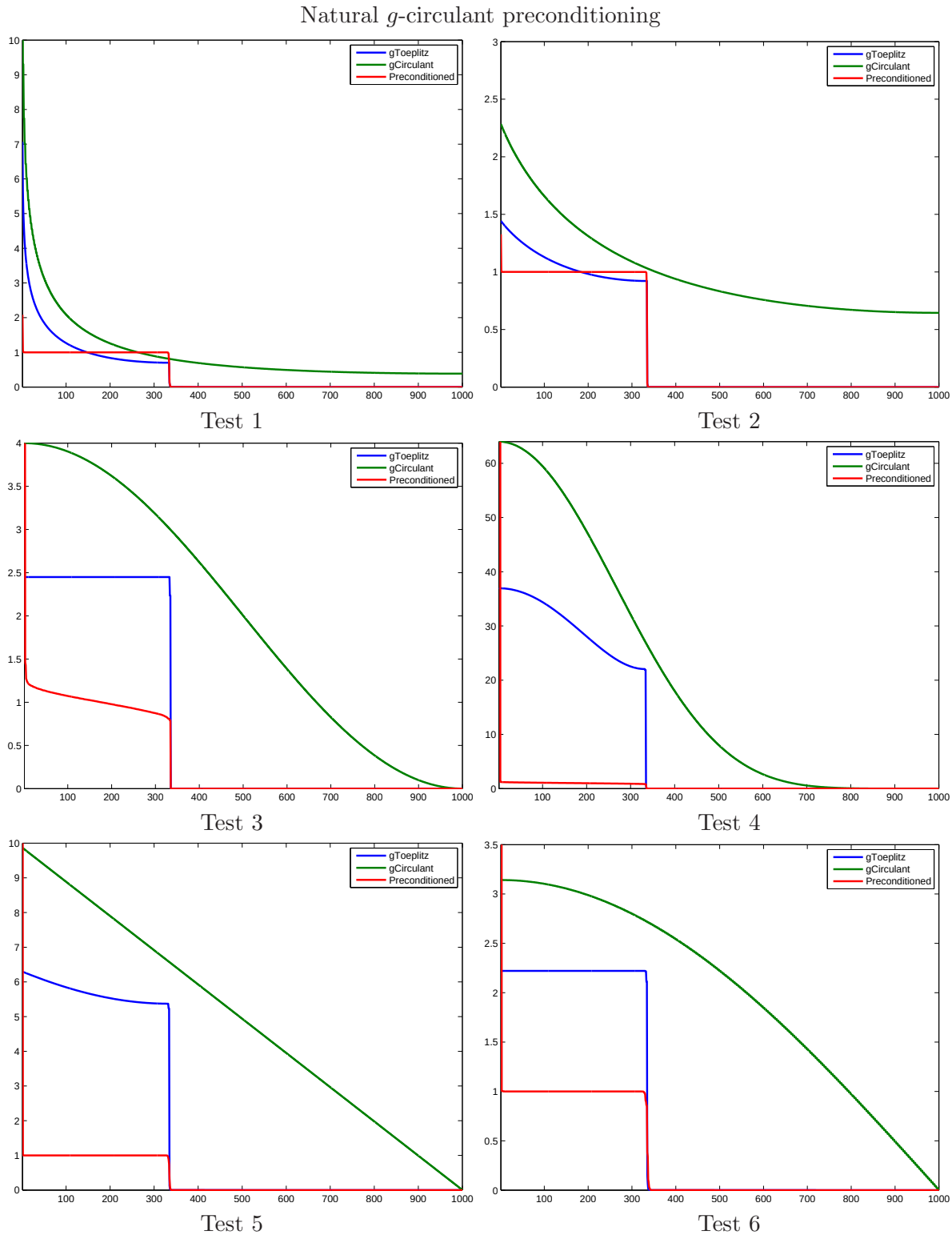

Figure 6.1: $g = 3$ (coprime case) - Singular values of g-Toeplitz matrices A, Natural g-circulant preconditioners P and corresponding preconditioned matrices $P^{-1}A$.

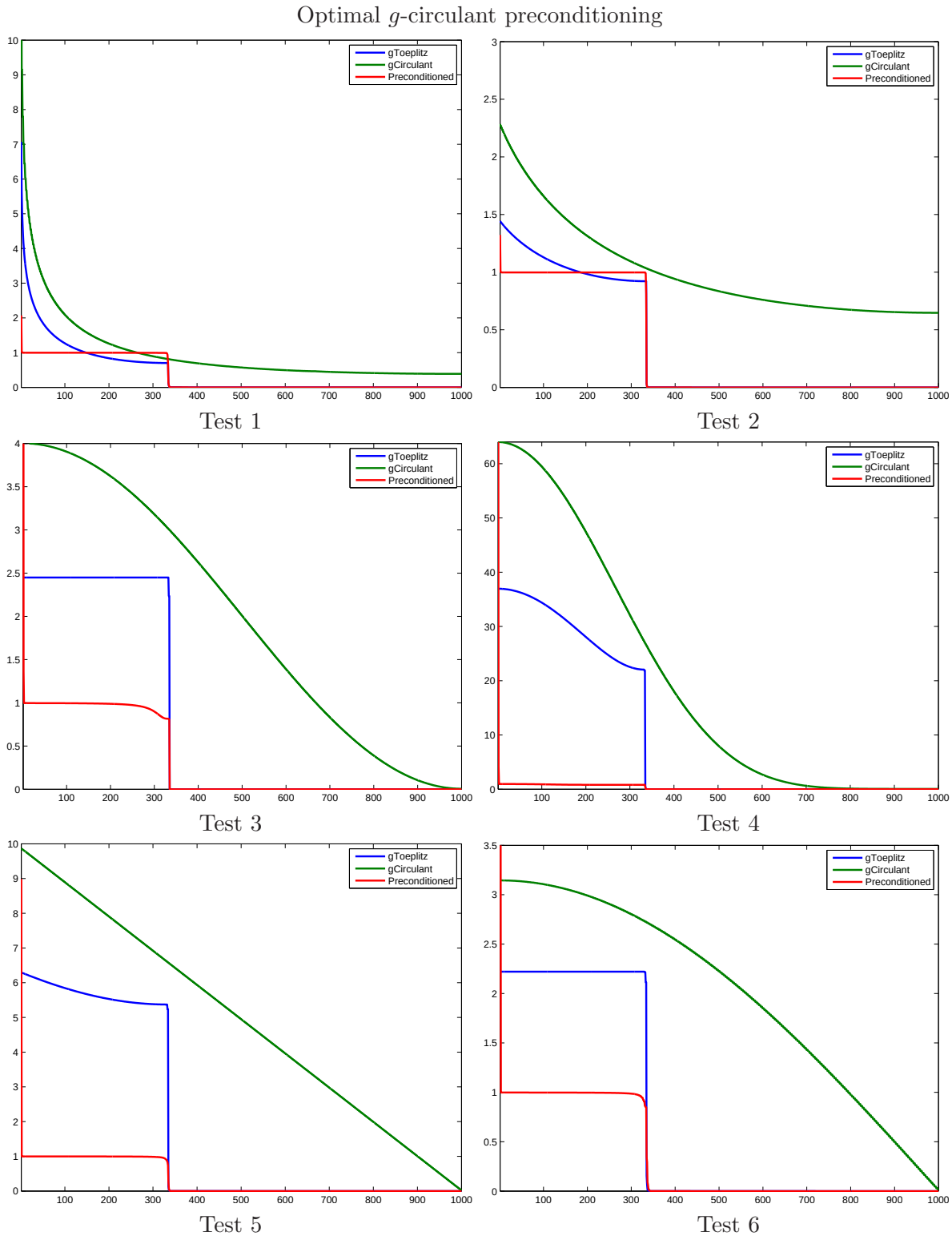

Figure 6.2: $g = 3$ (coprime case) - Singular values of g-Toeplitz matrices A, Optimal g-circulant preconditioners P and corresponding preconditioned matrices $P^{-1}A$.

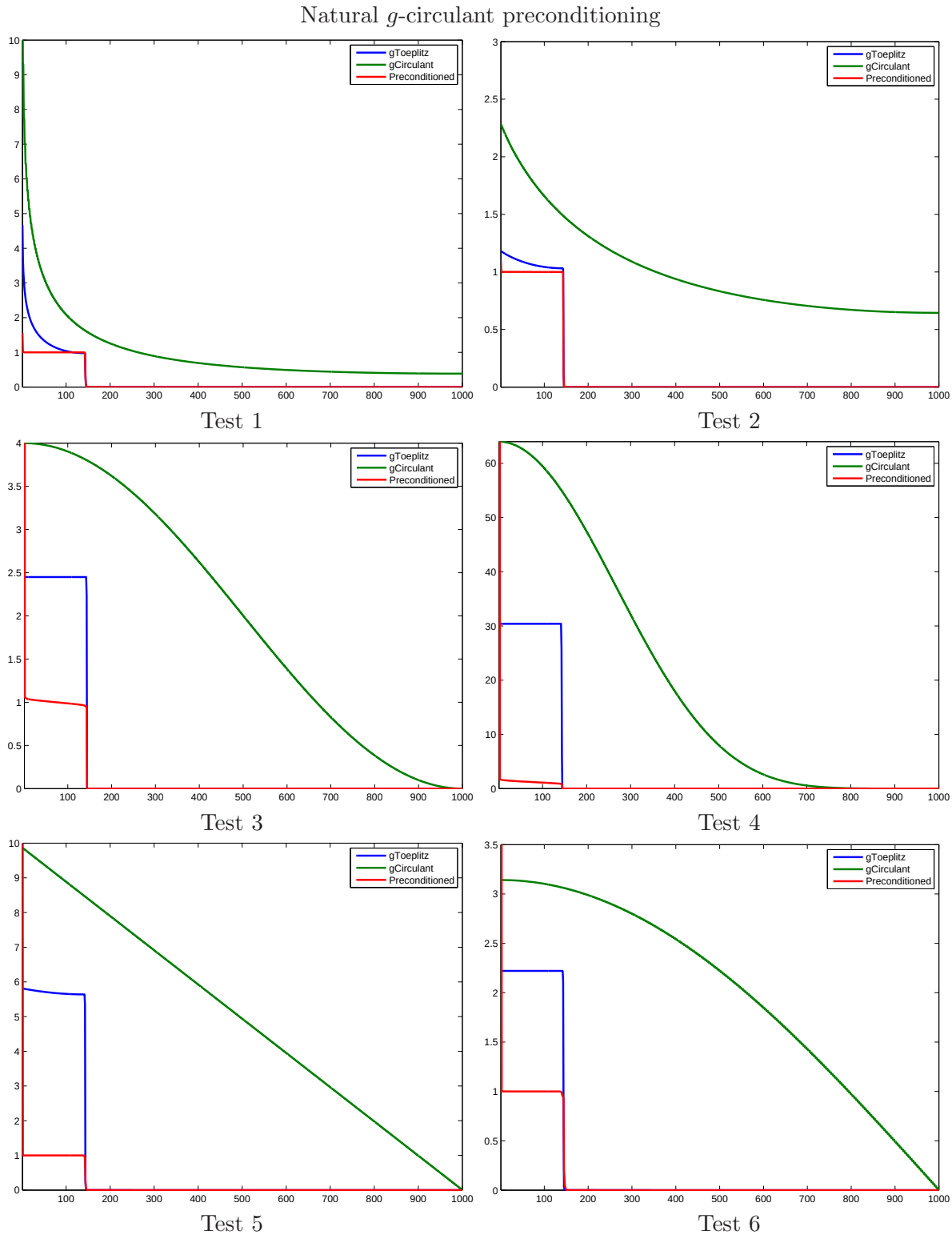

Figure 6.3: $g = 7$ (coprime case) - Singular values of g-Toeplitz matrices A, Natural g-circulant preconditioners P and corresponding preconditioned matrices $P^{-1}A$.

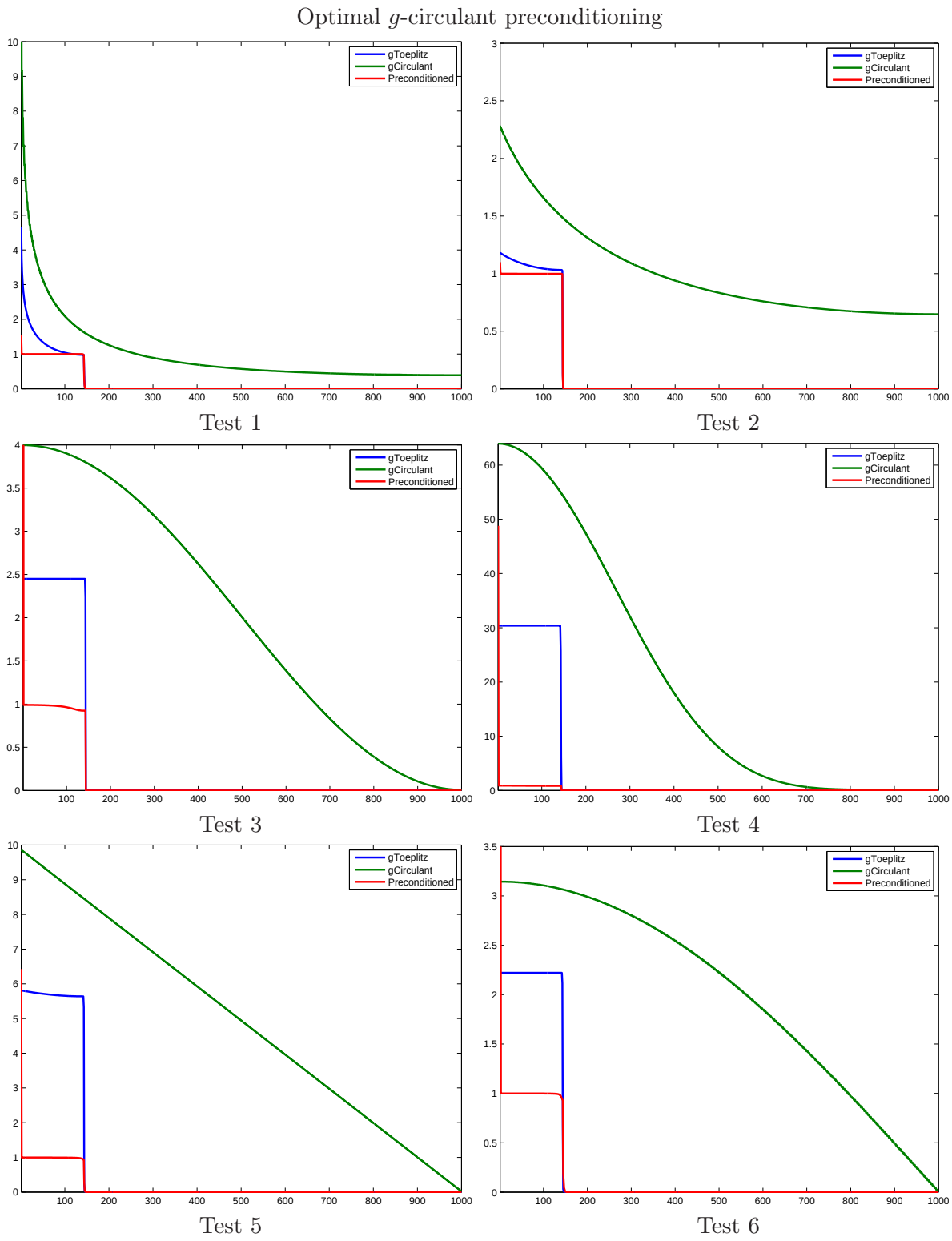

Figure 6.4: $g = 7$ (coprime case) - Singular values of g-Toeplitz matrices A, Optimal g-circulant preconditioners P and corresponding preconditioned matrices $P^{-1}A$.

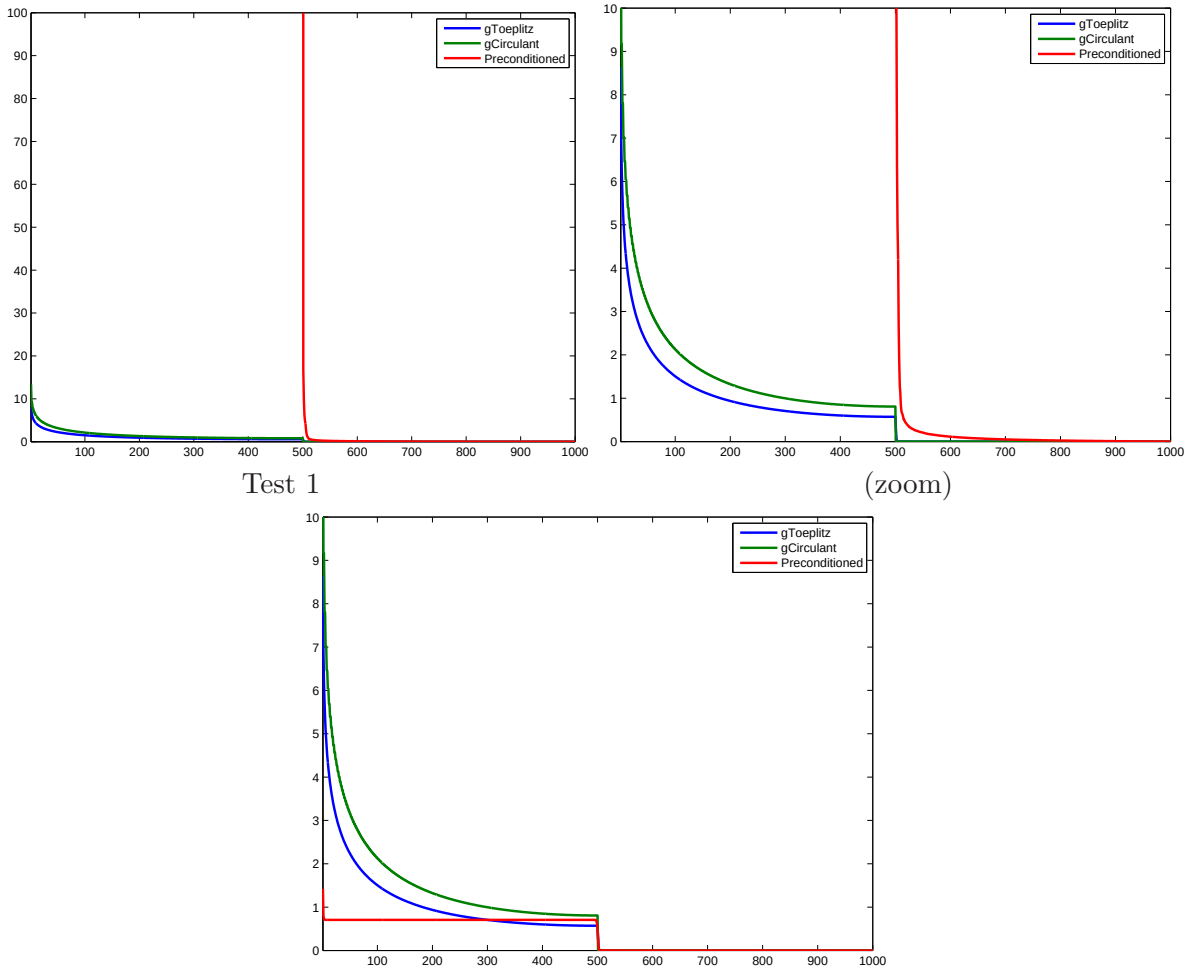

Figure 6.5: $g = 2$ (non-coprime case) - Singular values of g-Toeplitz matrices A, optimal g-circulant preconditioners P and corresponding preconditioned matrices $P^{\dagger}A$ (left), zoom on the small values (center), and analogous spectral distributions related to the regularized preconditioners (right).

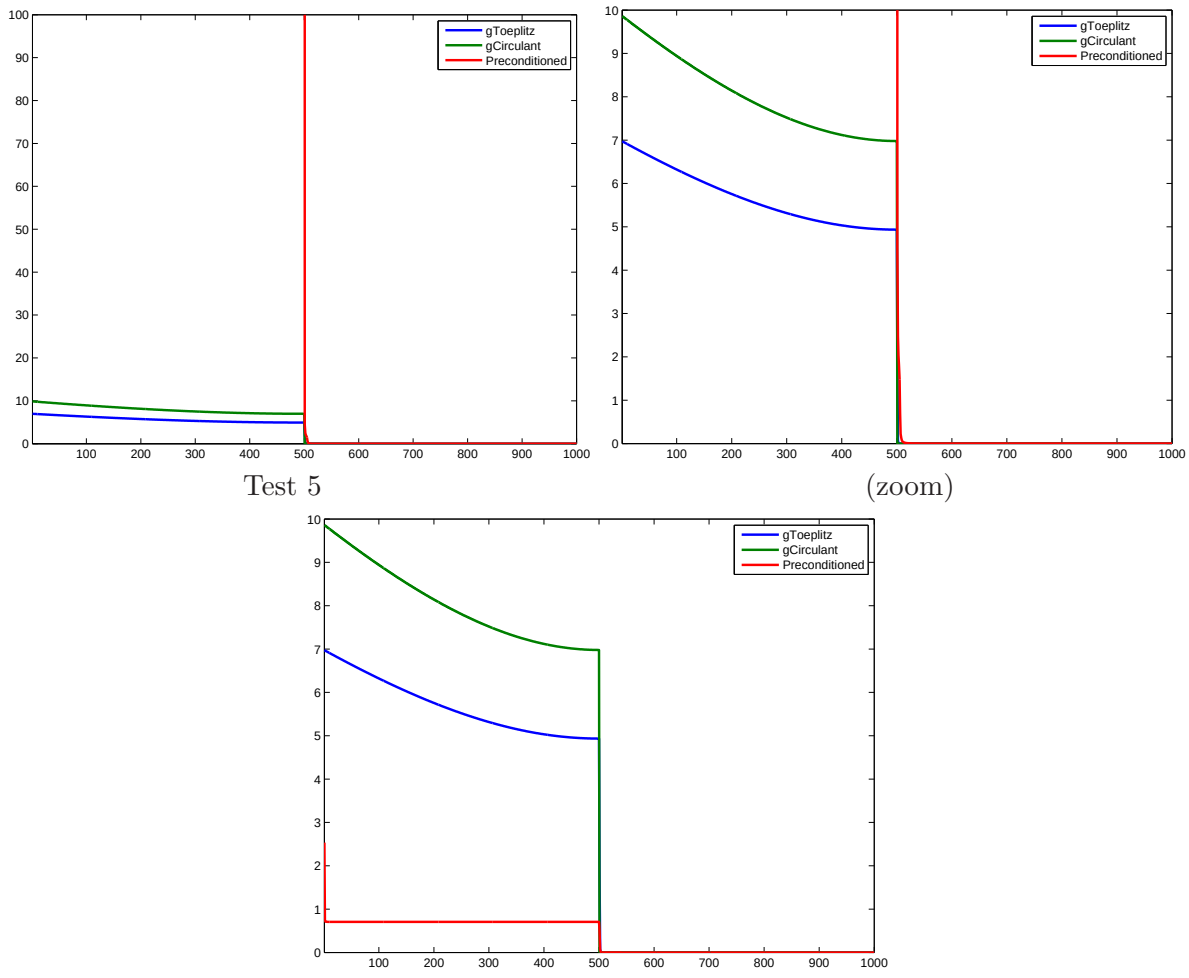

Figure 6.6: $g = 2$ (non-coprime case) - Singular values of g-Toeplitz matrices A, optimal g-circulant preconditioners P and corresponding preconditioned matrices $P^{\dagger}A$ (left), zoom on the small values (center), and analogous spectral distributions related to the regularized preconditioners (right).

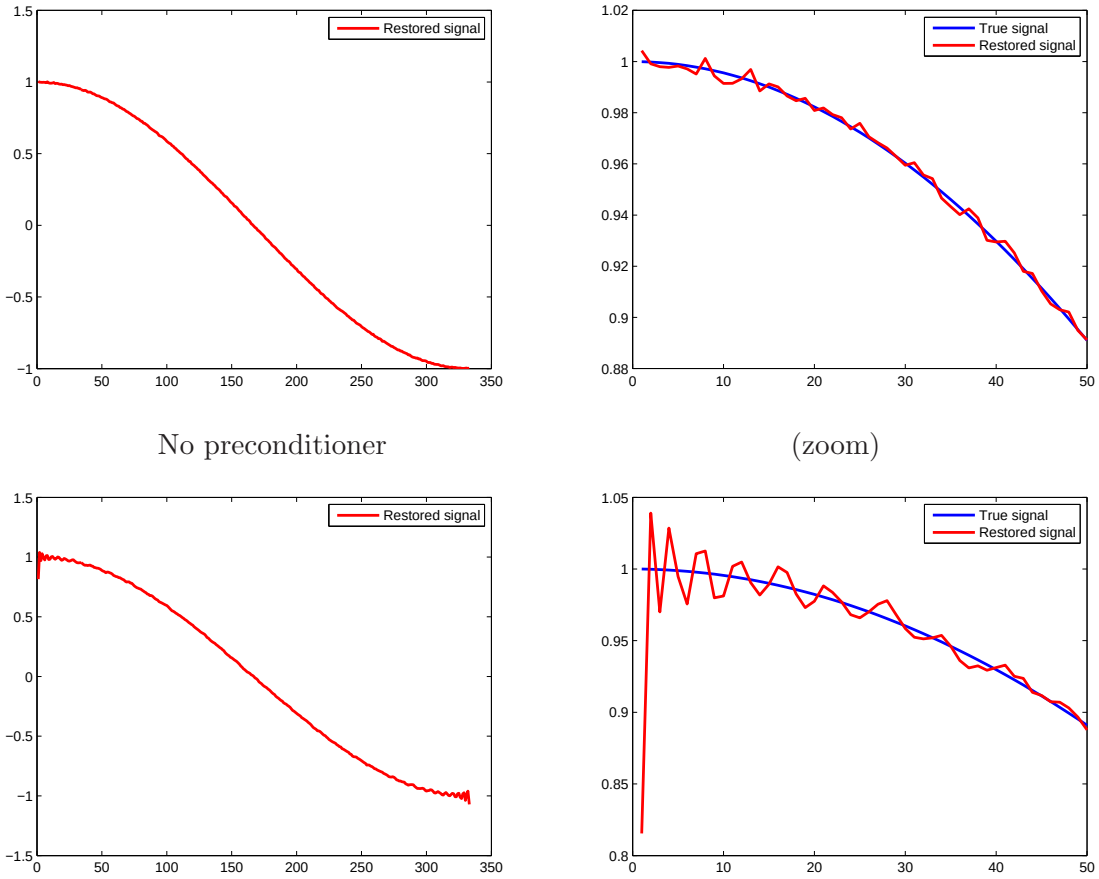

Figure 6.7: Restored signal with (P)CG on the normal equations (1% of data noise, $g = 3$). Left: without preconditioning. Right: with Optimal g-circulant preconditioning.

Figure 6.8: Non-perpendicular imaging system geometry.

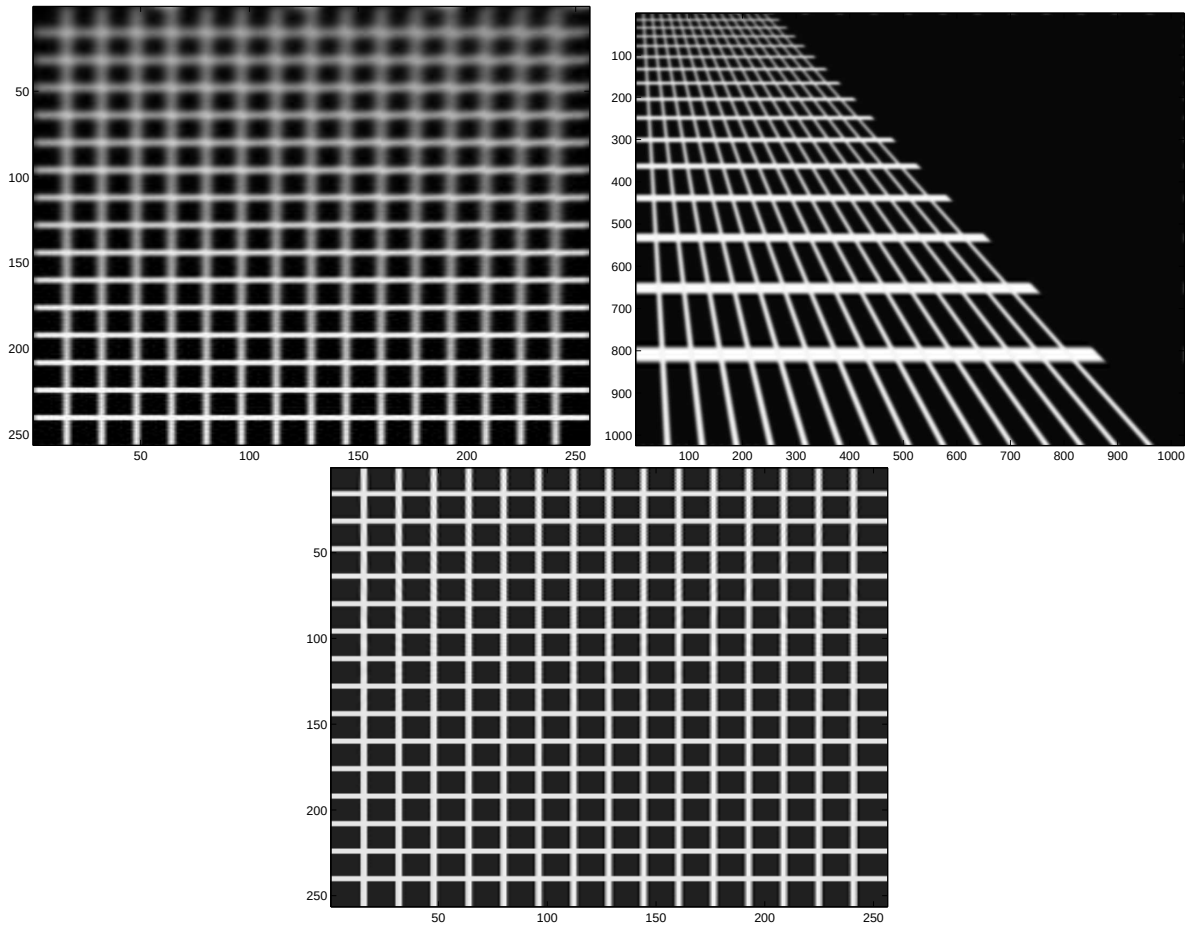

Figure 6.9: Shift-variant blurred data, projected data (shift-invariant blur), deblurred data.

Chapter 7

Multigrid methods for Toeplitz linear systems with different size reduction

In this chapter we analyze the convergence of a multigrid method where the fine problem of size n is projected to a coarser problem of size $\frac{n}{g}$, $g = 2, 3, \ldots$.Nonstandard projection matrices with $g > 2$ arise in a natural way when the aggregation projectors are applied to finite difference or regular finite elements approximations of partial differential equations (see [67] and the references therein). We perform a two-grid analysis using the ideas in [95] for circulant structures and by exploiting the spectral analysis of g-circulant matrices already performed in [68]. As shown in [33], such two-grid analysis is an algebraic generalization of the classical local Fourier analysis, which can be extended from circulant matrices to the more challenging case of Toeplitz matrices. In the multigrid strategy that we propose the coarse problem has size $\frac{n}{g}$ with $g > 2$ and, as an interesting byproduct, we can design multigrid methods with optimal cost and characterized by $g - 1$ recursive calls: in fact, it is enough to perform the analysis of the arithmetic computational cost related to the size reduction $\frac{n}{g}$ between two consecutive levels and to invoke the results in [114]. A further property of the proposed multigrid is that the pathologies induced by the mirror points are bypassed as previously mentioned. Our proposal can be plainly extended to the multilevel case by tensor product arguments, by taking into consideration the larger number of "mirror points" (which is exponential in the number d of levels). Moreover a V-cycle convergence analysis could be performed by following the steps in [3, 2] as a model.

7.1 Two-grid and Multigrid methods

Let $A_n \in M_n(\mathbb{C})$, and x_n, $b_n \in \mathbb{C}^n$. Let $p_n^k \in M_{n,k}(\mathbb{C})$, $k < n$, be a given full-rank matrix and let us consider a class of iterative methods of the form

$$x_n^{(j+1)} = V_n x_n^{(j)} + \tilde{b}_n := \mathcal{V}\left(x_n^{(j)}, \tilde{b}_n\right), \tag{7.1}$$

where $A_n = W_n - N_n$, W_n non-singular matrix, $V_n := I_n - W_n^{-1} A_n \in M_n(\mathbb{C})$, and $\tilde{b}_n := W_n^{-1} b_n \in \mathbb{C}^n$. A Two-Grid Method (TGM) is defined by the following algorithm:

$$\mathrm{TGM}\left(V_{n,\mathrm{pre}}^{\nu_{\mathrm{pre}}}, V_{n,\mathrm{post}}^{\nu_{\mathrm{post}}}, p_n^k\right)\left(x_n^{(j)}\right)$$

0. $\tilde{x}_n = \mathcal{V}_{n,\mathrm{pre}}^{\nu_{\mathrm{pre}}}\left(x_n^{(j)}, \tilde{b}_{n,\mathrm{pre}}\right)$
1. $d_n = A_n\tilde{x}_n - b_n$
2. $d_k = \left(p_n^k\right)^* d_n$
3. $A_k = \left(p_n^k\right)^* A_n p_n^k$
4. Solve $A_k y = d_k$
5. $\hat{x}_n = \tilde{x}_n - p_n^k y$
6. $x_n^{(j+1)} = \mathcal{V}_{n,\mathrm{post}}^{\nu_{\mathrm{post}}}\left(\hat{x}_n, \tilde{b}_{n,\mathrm{post}}\right)$

Steps 1. $\rightarrow$ 5. define the "coarse grid correction" that depends on the projecting operator p_n^k, while Step 0. and Step 6. consist, respectively, in applying ν_{pre} times and ν_{post} times a "pre-smoothing iteration" and a "post-smoothing iteration" of the generic form given in (7.1). The global iteration matrix of the TGM is then given by

$$\mathrm{TGM}\left(V_{n,\mathrm{pre}}^{\nu_{\mathrm{pre}}}, V_{n,\mathrm{post}}^{\nu_{\mathrm{post}}}, p_n^k\right) = V_{n,\mathrm{post}}^{\nu_{\mathrm{post}}}\left[I_n - p_n^k\left(\left(p_n^k\right)^* A_n p_n^k\right)^{-1}\left(p_n^k\right)^* A_n\right] V_{n,\mathrm{pre}}^{\nu_{\mathrm{pre}}}.$$

If k is large, the numerical solution of the linear system at the Step 4. could be computationally expensive. In such case a multigrid procedure is adopted. Consider $0 < m < n$, the sequence $0 < n_m < n_{m-1} < \cdots < n_1 < n_0 = n$ and the full-rank matrices $p_{n_{i-1}}^{n_i} \in M_{n_{i-1},n_i}(\mathbb{C})$, for $i = 1, \ldots, m$. The multigrid method produces the sequence $\left\{x_n^{(k)}\right\}_{k\in\mathbb{N}}$ defined by $x_n^{(j+1)} = \mathrm{MGM}\left(V_{n,\mathrm{pre}}^{\nu_{\mathrm{pre}}}, V_{n,\mathrm{post}}^{\nu_{\mathrm{post}}}, p_n^{n_1}, A_n, b_n, \theta, 0\right)\left(x_n^{(j)}\right)$ with the function MGM defined recursively as follows:

$$x_{n_i}^{(j+1)} = \mathrm{MGM}\left(V_{n_i,\mathrm{pre}}^{\nu_{\mathrm{pre}}}, V_{n_i,\mathrm{post}}^{\nu_{\mathrm{post}}}, p_{n_i}^{n_{i+1}}, A_{n_i}, b_{n_i}, \theta, i\right)\left(x_{n_i}^{(j)}\right)$$

If $i = m$ then Solve $A_{n_i} x_{n_i}^{(j+1)} = b_{n_i}$
Else
0. $\tilde{x}_{n_i} = \mathcal{V}_{n_i,\mathrm{pre}}^{\nu_{\mathrm{pre}}}\left(x_{n_i}^{(j)}, \tilde{b}_{n_i,\mathrm{pre}}\right)$
1. $d_{n_i} = A_{n_i}\tilde{x}_{n_i} - b_{n_i}$
2. $d_{n_{i+1}} = \left(p_{n_i}^{n_{i+1}}\right)^* d_{n_i}$
3. $A_{n_{i+1}} = \left(p_{n_i}^{n_{i+1}}\right)^* A_{n_i} p_{n_i}^{n_{i+1}}$
4. $x_{n_{i+1}}^{(j+1)} = 0$
   for $s = 1$ to θ
   $x_{n_{i+1}}^{(j+1)} = \mathrm{MGM}\left(V_{n_{i+1},\mathrm{pre}}^{\nu_{\mathrm{pre}}}, V_{n_{i+1},\mathrm{post}}^{\nu_{\mathrm{post}}}, p_{n_{i+1}}^{n_{i+2}}, A_{n_{i+1}}, d_{n_{i+1}}, \theta, i+1\right)\left(x_{n_i}^{(j+1)}\right)$
5. $\hat{x}_{n_i} = \tilde{x}_{n_i} - p_{n_i}^{n_{i+1}} x_{n_{i+1}}^{(j+1)}$
6. $x_{n_i}^{(j+1)} = \mathcal{V}_{n_i,\mathrm{post}}^{\nu_{\mathrm{post}}}\left(\hat{x}_{n_i}, \tilde{b}_{n_i,\mathrm{post}}\right)$

The choices $\theta = 1$ and $\theta = 2$ correspond to the well-known V-cycle and W-cycle, respectively.

In this chapter, we are interested in proposing such a kind of techniques in the case where A_n is a Toeplitz matrix. However, for a theoretical analysis, we consider circulant matrices according to the local Fourier analysis for classical multigrid methods (see [33]). Even if we treat in detail the circulant case, in the spirit of the paper [3], the same ideas can be plainly translated to other matrix algebras associated to (fast) trigonometric transforms. First we recall some convergence results from the theory of the algebraic multigrid method given in [77].

Theorem 7.1. [77] *Let $A_n \in M_n(\mathbb{C})$ be a positive definite matrix and let V_n be defined as in the* TGM *algorithm. Suppose that there exists $\alpha_{\mathrm{post}} > 0$ independent of n such that*

$$\|V_{n,\mathrm{post}} x_n\|_{A_n}^2 \leq \|x_n\|_{A_n}^2 - \alpha_{\mathrm{post}} \|x_n\|_{A_n D_n^{-1} A_n}^2, \qquad \forall x_n \in \mathbb{C}^n, \tag{7.2}$$

where D_n is the main diagonal of A_n. Assume that there exists $\gamma > 0$ independent of n such that

$$\min_{y\in\mathbb{C}^k} \left\|x_n - p_n^k y\right\|_{D_n}^2 \leq \gamma \left\|x_n\right\|_{A_n}^2, \qquad \forall x_n \in \mathbb{C}^n. \tag{7.3}$$

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

$$\left\|\text{TGM}\left(I, V_{n,\text{post}}^{\nu_{\text{post}}}, p_n^k\right)\right\|_{A_n} \leq \sqrt{1 - \frac{\alpha_{\text{post}}}{\gamma}}.$$

Conditions (3.7) and (7.3) are usually called as "smoothing property" and "approximation property", respectively.

We note that α_{post} and γ are independent of n and hence, if the assumptions of Theorem 7.1 are satisfied, then the resulting TGM is not only convergent but also optimal. In other words, the number of iterations in order to reach a given accuracy ϵ can be bounded from above by a constant independent of n (possibly depending on the parameter ϵ).

Of course, if the given method is complemented with a convergent pre-smoother, then by the same theorem we get a faster convergence. In fact, it is known that for square matrices A and B the spectra of AB and BA coincide.

Therefore $\text{TGM}\left(V_{n,\text{pre}}^{\nu_{\text{pre}}}, V_{n,\text{post}}^{\nu_{\text{post}}}, p_n^k\right)$ and $\text{TGM}\left(I, V_{n,\text{pre}}^{\nu_{\text{pre}}} V_{n,\text{post}}^{\nu_{\text{post}}}, p_n^k\right)$ have the same eigenvalues so that

$$\begin{aligned}
\left\|\text{TGM}\left(V_{n,\text{pre}}^{\nu_{\text{pre}}}, V_{n,\text{post}}^{\nu_{\text{post}}}, p_n^k\right)\right\|_{A_n} &= \left\|\text{TGM}\left(I, V_{n,\text{pre}}^{\nu_{\text{pre}}} V_{n,\text{post}}^{\nu_{\text{post}}}, p_n^k\right)\right\|_{A_n} \\
&< \sqrt{1 - \frac{\alpha_{\text{post}}^{\text{new}}}{\gamma}} \\
&< \sqrt{1 - \frac{\alpha_{\text{post}}}{\gamma}},
\end{aligned}$$

and hence the presence of a pre-smoother can only improve the convergence.

Concerning multigrid methods, in [77] the V-cycle convergence is considered with a result which could be seen as the analog of Theorem 7.1. For other bounds concerning the convergence rate of the V-cycle see [66] and reference therein. Regarding the convergence of the W-cycle, we point out that a rigorous TGM analysis is sufficient for determining the optimality of the W-cycle (see [114]).

7.2 Projecting operators for circulant matrices

Let $A_n := C_n(f)$ be a circulant matrix generated by a trigonometric polynomial f (see Chapter 5, Section 5.1). In order to provide a general method for obtaining a projecting operator (also called projector) from an arbitrary banded circulant matrix P_n, for some bandwidth independent of n, we introduce the operator $Z_{n,g}^k \in M_{n,k}(\mathbb{R})$, $k \leq n$, where

$$Z_{n,g}^k = \left[\delta_{i-gj}\right]_{i,j}, \qquad \delta_r = \begin{cases} 1 & \text{if } r \equiv 0 \pmod{n}, \\ 0 & \text{otherwise,} \end{cases} \qquad \begin{matrix} i = 0, \ldots, n-1, \\ j = 0, \ldots, k-1, \end{matrix} \tag{7.4}$$

(it is immediate to observe that (7.4) is the matrix defined in (5.9) by considering only the first k columns).

The operator $Z_{n,g}^k$ represents a special link between the space of the frequencies of size n and the corresponding space of frequencies of size k.

The relation between $Z_{n,g}^k$ and the Fourier matrix F_n (see (5.3)) described in Lemma 5.4 (see also Remark 5.7 and [113] for recent findings on these structures) is the key step in defining an algebraic multigrid method, since it allows us to obtain again a circulant matrix at the lower

level. Indeed, denoting by Δ_n the diagonal matrix obtained from the eigenvalues of A_n (see (5.5)), we infer that $\Delta_k := \frac{1}{g} I_{n,g}^\top \Delta_n I_{n,g}$ is again a diagonal matrix. Therefore

$$\begin{aligned}\left(Z_{n,g}^k\right)^\top A_n Z_{n,g}^k &= \left(Z_{n,g}^k\right)^\top F_n \Delta_n F_n^* Z_{n,g}^k \\ &= \frac{1}{g} F_k I_{n,g}^\top \Delta_n I_{n,g} F_k^* \\ &= F_k \Delta_k F_k^* \\ &= A_k,\end{aligned}$$

where A_k is a new circulant matrix. Consequently, starting from the matrix $Z_{n,g}^k$, it is possible to define a generic projector

$$p_{n,g}^k = P_n Z_{n,g}^k, \tag{7.5}$$

where P_n is a circulant matrix. Indeed $P_n^* A_n P_n$ is a circulant matrix and then

$$A_k = \left(p_{n,g}^k\right)^* A_n p_{n,g}^k,$$

is a circulant matrix of size k. We note that, since $k = \frac{n}{g} \in \mathbb{N}$, n must be a multiple of g. We are left to determine the conditions to be satisfied by $P_n = C_n(p)$ (or better by its generating function p), in order to get a projector which is effective in terms of convergence.

Definition 7.2. *Given $x \in [0, 2\pi)$, $g \in \mathbb{N}$, $g \geq 2$, the set of* g-corners *of x is $\Omega_g(x) = \{y = \left(x + \frac{2\pi j}{g}\right)$ mod 2π, $\mid j = 0, \ldots, g-1\}$ and the set of* g-mirror points *is $\mathcal{M}_g(x) = \Omega_g(x) \setminus \{x\}$.*

TGM conditions *Let $A_n := C_n(f)$ with f nonnegative, trigonometric polynomial and let $p_{n,g}^k = C_n(p) Z_{n,g}^k$ with p trigonometric polynomial. Assume that $f(x_0) = 0$ for $x_0 \in [0, 2\pi)$, choose p such that the following relations*

$$\lim_{x \to x_0} \frac{p^2(y)}{f(x)} < \infty, \qquad \forall\, y \in \mathcal{M}_g(x), \tag{7.6}$$

$$\sum_{y \in \Omega_g(x)} p^2(y) > 0, \qquad \forall\, x \in [0, 2\pi), \tag{7.7}$$

are fulfilled.

If f has a unique zero $x_0 \in [0, 2\pi)$, then we set $P_n = C_n(p)$ where p is a trigonometric polynomial defined as

$$p(x) = \prod_{\hat{x} \in \mathcal{M}_g(x_0)} (2 - 2\cos(x - \hat{x}))^{\lceil \frac{\beta}{2} \rceil} \sim \prod_{\hat{x} \in \mathcal{M}_g(x_0)} |x - \hat{x}|^{2\lceil \frac{\beta}{2} \rceil}, \tag{7.8}$$

for $x \in [0, 2\pi)$, with

$$\beta \geq \beta_{\min} = \min \left\{ i \,\middle|\, \lim_{x \to x_0} \frac{|x - x_0|^{2i}}{f(x)} < +\infty \right\},$$

thus conditions (7.6) and (7.7) are satisfied.

Before proving (in Subsection 7.3.1) that conditions (7.6) and (7.7) are sufficient to assure the TGM optimality, we consider a crucial result both from a theoretical and a practical point of view.

Proposition 7.3. *Let f be a non-negative function, $k = \frac{n}{g} \in \mathbb{N}$, $p_{n,g}^k = C_n(p) Z_{n,g}^k \in M_{n,k}(\mathbb{C})$, with p trigonometric polynomial satisfying condition (7.6) for any zero of f and globally the*

condition (7.7). Then the matrix $\left(p_{n,g}^k\right)^* C_n(f)\, p_{n,g}^k \in M_k(\mathbb{C})$ *coincides with* $C_k\left(\hat{f}\right)$ *where* $\hat{f}$ *is non-negative and*

$$\hat{f}(x) = \frac{1}{g} \sum_{y \in \Omega_g\left(\frac{x}{g}\right)} f(y) \left|p\right|^2 (y), \tag{7.9}$$

for $x \in [0, 2\pi)$*, i.e., the projected matrix is obtained picking every* (g)*th entry out of the symbol* $f\left|p\right|^2$*. In particular*

1. *if* f *is a polynomial then* $\hat{f}$ *is a polynomial with a fixed degree* $\left\lfloor \frac{q}{g} \right\rfloor$*, where* q *is the degree of* $f\left|p\right|^2$*;*
2. *if* x_0 *is a zero of* f *then* $\hat{f}$ *has a corresponding zero* y_0 *where* $y_0 = (gx_0) \mod 2\pi$*;*
3. *the order of the zero* y_0 *of* $\hat{f}$ *is exactly the same as the one of the zero* x_0 *of* f*, so that at the lower level the new projector can be easily defined in the same way.*

Proof. From (5.4) and (5.5), we find

$$\begin{aligned} \left(p_{n,g}^k\right)^* C_n(f)\, p_{n,g}^k &= \left(Z_{n,g}^k\right)^\top \left(C_n(p)\right)^* C_n(f)\, C_n(p)\, Z_{n,g}^k \\ &= \left(Z_{n,g}^k\right)^\top C_n\left(f\left|p\right|^2\right) Z_{n,g}^k, \end{aligned}$$

Thus the generating function of the circulant matrix $\left(C_n(p)\right)^* C_n(f)\, C_n(p)$ is $f\left|p\right|^2$. Denoting by a_j the (j)th Fourier coefficient of $f\left|p\right|^2$, we have

$$C_n\left(f\left|p\right|^2\right) = \left[a_{(r-s) \bmod n} + a_{(r-s) \bmod n - n}\right]_{r,s=0}^{n-1},$$

and hence, by (7.4), the entries of the matrix $\left(Z_{n,g}^k\right)^\top C_n\left(f\left|p\right|^2\right) Z_{n,g}^k$ are given by

$$\begin{aligned} \left(C_n\left(f\left|p\right|^2\right) Z_{n,g}^k\right)_{r,s} &= \sum_{\ell=0}^{n-1} \left(C_n\left(f\left|p\right|^2\right)\right)_{r,\ell} \left(Z_{n,g}^k\right)_{\ell,s} \\ &= \sum_{\ell=0}^{n-1} \left(a_{(r-\ell) \bmod n} + a_{(r-\ell) \bmod n - n}\right) \delta_{\ell - gs} \\ &\underset{\text{(a)}}{=} a_{(r-gs) \bmod n} + a_{(r-gs) \bmod n - n}, \\ &\qquad\qquad r = 0, \ldots, n-1,\ s = 0, \ldots, k-1, \end{aligned}$$

$$\begin{aligned} \left(\left(Z_{n,g}^k\right)^\top C_n\left(f\left|p\right|^2\right) Z_{n,g}^k\right)_{r,s} &= \sum_{\ell=0}^{n-1} \left(\left(Z_{n,g}^k\right)^\top\right)_{r,\ell} \left(C_n\left(f\left|p\right|^2\right) Z_{n,g}^k\right)_{\ell,s} \\ &= \sum_{\ell=0}^{n-1} \delta_{\ell - gr} \left(a_{(\ell-gs) \bmod n} + a_{(\ell-gs) \bmod n - n}\right) \\ &\underset{\text{(b)}}{=} a_{(gr-gs) \bmod n} + a_{(gr-gs) \bmod n - n}, \\ &\qquad\qquad r, s = 0, \ldots, k-1, \end{aligned}$$

where (a) is true because there exists a unique $\ell \in 1, 2, \ldots, n-1$ such that $\ell - gs \equiv 0 \pmod n$, that is, $\ell \equiv gs \pmod n$ and, since $0 \le gs \le n-1$, we obtain $\ell = gs$. The same argument is applicable for showing the validity of (b). Now if we denote by b_j the (j)th Fourier coefficient of $\hat{f}$ it only remains to show that $\left(C_n\left(\hat{f}\right)\right)_{r,c} = b_{(r-c) \bmod k} + b_{(r-c) \bmod k - k} =$

$a_{(gr-gc) \bmod n} + a_{(gr-gc) \bmod n-n}$, $r, c = 0, \ldots, k-1$. Since $f\,|p|^2$ is a polynomial, we can always write

$$f\,|p|^2\,(x) = \sum_{\ell=-\infty}^{\infty} a_\ell e^{\mathrm{i}\ell x}, \qquad \hat{f}\,(x) = \sum_{\ell=-\infty}^{\infty} b_\ell e^{\mathrm{i}\ell x}. \tag{7.10}$$

From (4.2), (7.9) and (7.10), we have

$$\begin{aligned} b_{r-c} &= \frac{1}{2\pi}\int_0^{2\pi} \frac{1}{g}\sum_{j=0}^{g-1}\sum_{\ell=-\infty}^{+\infty} a_\ell e^{\mathrm{i}\ell\left(\frac{x+2\pi j}{g}\right)} e^{-\mathrm{i}(r-c)x}\mathrm{d}x \\ &= \frac{1}{2\pi g}\int_0^{2\pi} \sum_{\ell=-\infty}^{+\infty} a_\ell \left(\sum_{j=0}^{g-1} e^{\frac{\mathrm{i}2\pi\ell j}{g}}\right) e^{\frac{\mathrm{i}\ell x}{g}} e^{-\mathrm{i}(r-c)x}\mathrm{d}x. \end{aligned}$$

Taking into account

$$\frac{1}{g}\sum_{j=0}^{g-1} e^{\frac{\mathrm{i}2\pi\ell j}{g}} = \begin{cases} 1 & \text{if } \ell = gt, \\ 0 & \text{otherwise,} \end{cases} \qquad \text{and} \qquad \frac{1}{2\pi}\int_0^{2\pi} e^{\mathrm{i}\ell x}\mathrm{d}x = \begin{cases} 1 & \text{if } \ell = 0, \\ 0 & \text{otherwise,} \end{cases}$$

we obtain

$$\begin{aligned} b_{r-c} &= \frac{1}{2\pi g}\int_0^{2\pi} \sum_{t=-\infty}^{+\infty} a_{gt} g e^{\frac{\mathrm{i}gtx}{g}} e^{-\mathrm{i}(r-c)x}\mathrm{d}x \\ &= \sum_{t=-\infty}^{+\infty} a_{gt} \frac{1}{2\pi}\int_0^{2\pi} e^{\mathrm{i}x(t-(r-c))}\mathrm{d}x \\ &= a_{g(r-c)}. \end{aligned} \tag{7.11}$$

So, from (7.11), since $gk = n$ and $g\,((r-c) \bmod k) = (gr - gc) \bmod n$, we have $b_{(r-c) \bmod k} = a_{(gr-gc) \bmod n}$ and, similarly, $b_{(r-c) \bmod k-k} = a_{(gr-gc) \bmod n-n}$.

From the expression of $\hat{f}$, since $f\,(x_0) = 0$ we deduce $p\,(y) = 0\ \forall y \in \mathcal{M}_g\,(x_0)$, whose validity is necessary in order to satisfy relationships (7.6). Thus $y_0 = (gx_0) \bmod 2\pi$ is a zero of $\hat{f}$ (i.e. item 2. is proved).

Moreover, by (7.7), we deduce that $p^2\,(x_0) > 0$ since $p^2\,(y) = 0$, $\forall y \in \mathcal{M}_g\,(x_0)$, and the order of the zero y_0 of $f\left(\frac{x}{g}\right)|p|^2\left(\frac{x}{g}\right)$ is the same as the order of $f\,(x)$ at x_0. Furthermore, by (7.6) we see that $|p|^2\left(\frac{x+2\pi k}{g}\right)$ has at y_0 a zero of order at least equal to the one of $f\,(x)$ at x_0, for any $k = 1, \ldots, g-1$. Since all the contributions in $\hat{f}$ are non-negative the thesis in item 3. is proved.

Finally we have to prove item 1. Since b_j are the Fourier coefficients of $\hat{f}$ and a_j are the Fourier coefficients of the polynomial $f\,|p|^2$ (see (7.10)), from (7.11) we deduce that

$$\hat{f}\,(x) = \sum_j b_j e^{\mathrm{i}jx} = \sum_j a_{gj} e^{\mathrm{i}jx}.$$

Hence, if the polynomial $f\,|p|^2$ has degree q, then $\hat{f}$ has degree at most $\left\lfloor \frac{q}{g} \right\rfloor$. □

7.3 Proof of convergence

Using the results in Section 7.2, we prove the optimality of the TGM and of the W-cycle (for the W-cycle the condition $g > 2$ is required).

7.3.1 TGM convergence

The smoothing property for $g = 2$ was proved in [95] and the proof does not change for $g > 2$.

Lemma 7.4. [95] *Let $A_n := C_n(f)$ with f being a non-negative trigonometric polynomial (not identically zero) and let $V_n := I_n - \omega A_n$, $0 < \omega < \frac{2}{\|f\|_{L^\infty}}$. If we choose α_{post} so that $\alpha_{\text{post}} \leq a_0 \omega (2 - \omega \|f\|_{L^\infty})$, then relation (3.7) holds true.*

If in the previous Lemma we choose $\omega = \|f\|_{L^\infty}^{-1}$, then $\alpha_{\text{post}} \leq \frac{\|f\|_{L^1}}{\|f\|_{L^\infty}}$ and the best value of α_{post} is $\alpha_{\text{post,best}} = \frac{\|f\|_{L^1}}{\|f\|_{L^\infty}}$. Moreover, the result of Lemma 7.4 can be easily generalized when considering both pre-smoothing and post-smoothing as in [2].

The following result shows that TGM conditions (7.6) and (7.7) are sufficient in order to satisfy the approximation property.

Theorem 7.5. *Let $A_n := C_n(f)$, with f being a non-negative trigonometric polynomial (not identically zero), and let $p_{n,g}^k = C_n(p) Z_{n,g}^k$ be the projecting operator, with $Z_{n,g}^k$ defined in (7.4) and with p trigonometric polynomial, satisfying condition (7.6), for any zero of f, and satisfying globally condition (7.7). Then, there exists a positive value γ independent of n such that inequality (7.3) is satisfied.*

Proof. The proof is similar to that of [85, Lemma 8.2], but we report it here for completeness. First, we recall that the main diagonal of A_n is given by $D_n = a_0 I_n$ with $a_0 = \frac{1}{2\pi} \int_Q f = \|f\|_{L^1} > 0$, so that $\|\cdot\|_{D_n}^2 = a_0 \|\cdot\|_2^2$.

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

$$\min_{y \in \mathbb{C}^k} \left\| x_n - p_{n,g}^k y \right\|_{D_n}^2 = a_0 \min_{y \in \mathbb{C}^k} \left\| x_n - p_{n,g}^k y \right\|_2^2 \leq \gamma \|x_n\|_{A_n}^2 ,$$

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

$$\overline{y} = \left[\left(p_{n,g}^k \right)^* p_{n,g}^k \right]^{-1} \left(p_{n,g}^k \right)^* x_n .$$

Therefore, (7.3) is implied by

$$\left\| x_n - p_{n,g}^k \overline{y} \right\|_2^2 \leq \frac{\gamma}{a_0} \|x_n\|_{A_n}^2 , \qquad \forall x_n \in \mathbb{C}^n ,$$

where the latter is equivalent to the matrix inequality

$$W_n(p)^* W_n(p) \leq \frac{\gamma}{a_0} C_n(f) , \tag{7.12}$$

with $W_n(p) = I - p_{n,g}^k \left[\left(p_{n,g}^k \right)^* p_{n,g}^k \right]^{-1} \left(p_{n,g}^k \right)^*$. Since, by construction, $W_n(p)$ is a Hermitian unitary projector, it holds that $W_n(p)^* W_n(p) = W_n^2(p) = W_n(p)$. As a consequence, inequality (7.12) can be rewritten as

$$W_n(p) \leq \frac{\gamma}{a_0} C_n(f) . \tag{7.13}$$

When $k = \frac{n}{g} \in \mathbb{N}$, following the decomposition in (5.18), $p_{n,g}^k = C_n(p) Z_{n,g}^k$ can be expressed according to

$$\left(p_{n,g}^k \right)^* = \frac{1}{\sqrt{g}} F_k \left[\Delta_p^{(0)} | \Delta_p^{(1)} | \cdots | \Delta_p^{(g-1)} \right] F_n^* ,$$

where

$$\Delta_p^{(r)} = \underset{j=0,\ldots,k-1}{\text{diag}} \left(p\left(x_{rk+j,n}^{(n)} \right) \right), \qquad r = 0, \ldots, g-1,$$

with $x_j^{(n)} = \frac{2\pi j}{n}$.

Let $p[\mu] \in \mathbb{C}^g$ whose entries are given by the evaluations of p over the points of $\Omega\left(x_\mu^{(n)}\right)$, for $\mu = 0, \ldots, k-1$. There exists a suitable permutation by rows and columns of $F_n^* W_n(p) F_n$, such that we can obtain a $g \times g$ block diagonal matrix whose μth diagonal block is given by $\frac{I_g - p[\mu](p[\mu])^\top}{\|p[\mu]\|_2^2}$. Therefore, using the same notation for $f[\mu]$ and denoting by diag $(f[\mu])$ the diagonal matrix having the vector $f[\mu]$ on the main diagonal, the condition (7.13) is equivalent to

$$I_g - \frac{p[\mu](p[\mu])^\top}{\|p[\mu]\|_2^2} \leq \frac{\gamma}{a_0} \text{diag}\,(f[\mu]), \tag{7.14}$$

for $\mu = 0, \ldots, k-1$. By the Sylvester inertia law [46], the relation (7.14) is satisfied if every entry of

$$\text{diag}\,(f[\mu])^{-\frac{1}{2}} \left(I_g - \frac{p[\mu](p[\mu])^\top}{\|p[\mu]\|_2^2} \right) \text{diag}\,(f[\mu])^{-\frac{1}{2}},$$

is bounded in modulus by a constant, which follows from the TGM conditions (7.6) and (7.7).

Furthermore, if we put

$$\begin{aligned} z &= \max_{y \in \Omega_g(x)} \left\| \frac{p^2(y)}{f(x)} \right\|_{L^\infty}, \\ h &= \left\| \frac{1}{\sum_{y \in \Omega_g(x)} p^2(y)} \right\|_{L^\infty}, \end{aligned}$$

then condition (7.3) is satisfied by choosing a value of γ such that $\gamma \geq g(g-1) a_0 h z$. □

Combining Lemma 7.4 and Theorem 7.5 with Theorem 7.1, it follows that the TGM convergence speed does not depend on the size of the linear system.

7.3.2 Multigrid convergence

The optimal TGM convergence rate proved in Theorem 7.5 can be extended to a generic recursion level of the multigrid procedure obtaining the so called "level independency" property. The key tools are Proposition 7.3 and an explicit choice of the projector: for instance a good choice is obtained by considering the symbol p reported in (7.8). Indeed, the "level independency" was already proved in literature for $g = 2$ (see [25, 27, 3]) and the proof can be extended to $g > 2$, as in Theorem 7.5.

The "level independency" implies that the W-cycle has a constant convergence rate independent of the problem size [114]. However, the fact that the convergence speed does not depend on the size of the linear system does not imply the optimality of the method, because the computational work at each iteration is not taken into account.

For estimating the computational work at each iteration of a multigrid method, we have to consider the size of the coarse problem and the number θ of recursive calls. In our case the size of the problem at the level i is $n_i = g n_{i-1}$. According to the analysis in [114], we assume that the multigrid components (smoothing, projection, ...) require a number of arithmetic operations which is bounded by cn_i, with c constant independent of n_i. From [114, equation (2.4.14)], the total computational work $C_m(n)$ of one complete multigrid cycle is

$$C_m(n) \doteq \begin{cases} \frac{g}{g-\theta} cn & \text{for } \theta < g, \\ O(n \log n) & \text{for } \theta = g, \end{cases} \tag{7.15}$$

where the symbol $\doteq$ means equality up to lower order terms. It follows that for $g = 2$ the W-cycle can not be optimal even in the presence of "level independency", because each multigrid iteration requires a computational cost of $O(n \log n)$ while the matrix vector product requires $O(n)$ arithmetic operations. On the other hand, for $g = 3$ the W-cycle shows a computational cost growing as $C_m(n) \doteq 3cn$ and hence it is optimal if the "level independency" is satisfied. More in general, the proposed multigrid will be optimal for a number $\theta \in \mathbb{N}$ of recursive calls such that $1 < \theta < g$.

In this chapter we do not provide a convergence analysis for multilevel Circulant matrices. Nevertheless, the computational cost of a multigrid iteration can be easily computed, up to lower order terms, for multidimensional problems as well. For a d-dimensional problem, for simplicity and without loss of generality, we assume the same reduction g along each direction and so the size of the problem at the level i is $n_i = g^d n_{i-1}$, where n_i is the algebraic size of the problem at the level i and $n_0 = n$. Therefore equation (7.15) generalizes into

$$C_m(n) \doteq \begin{cases} \frac{g^d}{g^d - \theta} cn & \text{for } \theta < g^d, \\ O(n \log n) & \text{for } \theta = g^d. \end{cases}$$

It follows that each multigrid iteration has a linear computational cost in n only if $1 \leq \theta < g^d$. In particular, for two-dimensional problems ($d = 2$), when $g \geq 2$ the computational cost of a W-cycle iteration ($\theta = 2$) is linear in n but, for instance, the first order coefficient is equal to $2c$ if $g = 2$, while it is equal to $\frac{9c}{7}$ if $g = 3$.

7.3.3 Some pathologies eliminated when using $g > 2$

From [95, conditions (3.4) and (3.5)], we know that, for $g = 2$, if x_0 is a zero of f, then $f(x_0 + \pi)$ must be positive: otherwise [95, relationship (3.5)] cannot be satisfied no matter which polynomial p we choose. On the other hand, if we consider $g = 3$ then the presence of two zeros at x_0 and $x_0 + \pi$ is no longer a problem, because conditions (7.6) and (7.7) impose that, if x_0 is a zero of f, then $f\left(x_0 + \frac{2}{3}\pi\right)$ and $f\left(x_0 + \frac{4}{3}\pi\right)$ must be positive, while we do not have constraints on the value f at $x_0 + \pi$.

For $g = 3$, if f has a unique zero $x_0 \in [0, 2\pi)$ of finite order, then we consider $\hat{x} = \left(x_0 + \frac{2}{3}\pi\right)$ mod 2π and $\tilde{x} = \left(x_0 + \frac{4}{3}\pi\right)$ mod 2π and we set $P_n = C_n(p)$, where p is a trigonometric polynomial defined as

$$p(x) = (2 - 2\cos(x - \hat{x}))^{\lceil \frac{\beta}{2} \rceil} (2 - 2\cos(x - \tilde{x}))^{\lceil \frac{\beta}{2} \rceil} \sim |x - \hat{x}|^{2\lceil \frac{\beta}{2} \rceil} |x - \tilde{x}|^{2\lceil \frac{\beta}{2} \rceil}, \qquad (7.16)$$

for $x \in [0, 2\pi)$. In this case the parameter β has to be chosen as

$$\beta \geq \beta_{\min} = \min \left\{ i \,\middle|\, \lim_{x \to x_0} \frac{|x - x_0|^{2i}}{f(x)} < +\infty \right\},$$

so that conditions (7.6) and (7.7) are satisfied. If f shows more than one zero in $[0, 2\pi)$, then we consider a polynomial p which is the product of the basic polynomials of the same type reported in (7.16), satisfying condition (7.6) for any single zero and satisfying globally condition (7.7).

A different strategy was proposed in [25] where symmetric positive definite Toeplitz matrices with entries $a_1 = \cdots = a_l = 0$, $l < n$, are considered (like, for example, Toeplitz matrices generated by $f(x) = 1 - \cos((l+1)x)$). A block symbol analysis and a related generalization was discussed in [55] mainly based on the interpretation of $T_n(f)$ with $f(x) = 1 - \cos((l+1)x)$ as $T_m(F)$ with $m = \frac{n}{l}$ and with $F(x) = (1 - \cos(x)) I_l$, being a matrix-valued function. The projector is defined as a $l \times l$ block matrix. The basic projector (7.4) is replaced by its $l \times l$ block version, where 1 and 0 become the identity and the zero matrix of order l, respectively. The symbol of the projector is $p(x) = 1 + \cos((l+1)x)$ or, equivalently, the block symbol

$P(x) = (1 - \cos(x))\, I_l$. With this approach the size of the coarse problem is one half of the size of the finer problem. A Galerkin strategy for $l > 1$ leads to lose the Toeplitz structure and so in [55] a natural coarse grid operator, previously proposed in [54], was employed.

Comparing the two strategies, usually the coarser matrices have about the same sparsity. Therefore, our proposal with $g = 3$ has a lower computational cost for iteration, since the size reduction factor is 3 instead of 2, but usually requires more iterations to converge than the proposal in [25, 55].

Example 7.6. *The symbol*

$$f(x) = (2 - 2\cos(x))(2 + 2\cos(x)) = 2 - 2\cos(2x), \tag{7.17}$$

vanishes at $x = 0$ and at $x = \pi$ with order two.

For $g = 3$, we have $\mathcal{M}_3(0) = \left\{\frac{2\pi}{3}, \frac{4\pi}{3}\right\}$ and $\mathcal{M}_3(\pi) = \left\{\frac{5\pi}{3}, \frac{7\pi}{3}\right\}$, thus the trigonometric polynomial

$$p(x) = \frac{1}{\sqrt{3}} \prod_{\hat{x} \in \mathcal{M}_3(0) \bigcup \mathcal{M}_3(\pi)} (2 - 2\cos(x - \hat{x})), \tag{7.18}$$

satisfies the TGM conditions (7.6) *and* (7.7) *and defines an optimal TGM. The only nonzero Fourier coefficients of p are $a_0 = \sqrt{3}$, $a_{\pm 2} = \frac{2}{\sqrt{3}}$, and $a_{\pm 4} = \frac{1}{\sqrt{3}}$.*

The coarse function $\hat{f}$ in (5.66) *is equal to f.*

The block projector proposed in [25] *is*

$$\frac{1}{2}\begin{bmatrix} I_2 & 2I_2 & I_2 & 0 & & \\ & 0 & I_2 & 2I_2 & I_2 & 0 \\ & & & & \ddots & \ddots \end{bmatrix}_{n \times \frac{n}{2}}$$

and the coarse matrix has the same symbol f, up to a scaling factor (see [55]*). Therefore both strategies preserve the same symbol f on the coarser levels.*

7.4 Numerical experiments

In this section, we apply the proposed multigrid method to symmetric positive definite circulant and Toeplitz systems $A_n x = b$. We choose as solution the vector x such that $x_i = \frac{i}{n}$, $i = 1, \ldots, n$. The right-hand side vector b is obtained accordingly. As smoother, we use Richardson with $\omega_j = \frac{1}{\|f_j\|_{L^\infty}}$, for $j = 0, \ldots, m-1$ (m is the number of subgrids in the algorithm, $m = 1$ for the TGM), for pre-smoother and the conjugate gradient for post-smoother. In the V-cycle and W-cycle procedure when the coarse grid size is less than or equal to 27, we solve the coarse grid system exactly. The zero vector is used as the initial guess and the stopping criterion is $\frac{\|r_q\|_2}{\|r_0\|_2} \leq 10^{-7}$, where r_q is the residual vector after q iterations and 10^{-7} is the given tolerance.

7.4.1 Cutting operators for Toeplitz matrices

When dealing with circulant matrices, using the projector defined in (7.5), the matrix at the lower level is still a circulant matrix, while for Toeplitz matrices, if we consider $A_n := T_n(f)$ and $p_{n,3}^k = T_n(p)\, Z_{n,3}^k$, where p is defined in accordance with the formula (7.16) and $k = \frac{n}{3} \in \mathbb{N}$, we find that

$$T_n(p)\, T_n(f)\, T_n(p) = T_n\left(fp^2\right) + G_n(f, p).$$

Furthermore, if $2\beta + 1$ is the bandwidth of $T_n(p)$, the matrix $G_n(f, p)$ has rank at most 2β and is formed by a matrix of rank β in the upper left corner and a matrix of the same rank in the bottom right corner. According to the proposal in [3], we take a cutting matrix that

will completely erase the contribution of $G_n(f,p)$, so that, at the lower level, the restriction of the matrix $T_n(p)\,T_n(f)\,T_n(p)$ is still a Toeplitz matrix and thus we can recursively apply the algorithm. The proposed cutting matrix is as follows:

$$\widetilde{Z}^k_{n,3} = \begin{bmatrix} 0^{k'}_{\beta} \\ Z^{k'}_{n-2\beta,3} \\ 0^{k'}_{\beta} \end{bmatrix}_{n\times k'} \qquad (k' \text{defined below}),$$

where $0^{k'}_{\beta} \in M_{\beta,k'}(\mathbb{R})$ is the zero matrix; $\widetilde{Z}^k_{n,3}$ has the first and the last β rows equal to zero and therefore it is able to remove corner corrections of rank less than or equal to 2β. Since $G_n(f,p)$ has rank 2β, we deduce that $A_{k'} = \left(\widetilde{Z}^k_{n,3}\right)^{\top} T_n(p)\,T_n(f)\,T_n(p)\,\widetilde{Z}^k_{n,3}$ is Toeplitz and, in the generic case, we cannot obtain a Toeplitz matrix of size greater than this. As a consequence, for Toeplitz matrices, the projector is then defined as

$$p^k_{n,3} = T_n(p)\,\widetilde{Z}^k_{n,3}.$$

Also the size n of the problem should be chosen in such a way that a recursive application of the algorithm is possible; in our case, if we choose $n = 3^{\alpha} - \xi$ with $\xi = \beta - 1$, $k = \frac{n}{3}$ and $k' = k - \frac{2(\beta-1)}{3}$, then the size of the problem at the lower level becomes $k' = \frac{n-2(\beta-1)}{3} = \frac{3^{\alpha}-(\beta-1)-2(\beta-1)}{3} = 3^{\alpha-1} - (\beta-1) = 3^{\alpha-1} - \xi$.

7.4.2 Zero at the origin and at π.

We present some examples where the generating functions f_0 vanish at the origin ($x = 0$) and at $x = \pi$. First, we consider the Example 7.6 where the symbol

$$f_0(x) = (2 - 2\cos(x))\,(2 + 2\cos(x)),$$

vanishes at $x = 0$ and $x = \pi$ with order 2. According to (7.16), we choose the projector $p^k_{n,3} = C_n(p_0)\,Z^k_{n,3}$ if A_n is a circulant matrix and $p^k_{n,3} = T_n(p_0)\,\widetilde{Z}^k_{n,3}$ if A_n is a Toeplitz matrix, where $p_0 = p$ defined in (7.18). Setting $x^{(1)}_0 = 0$ and $x^{(2)}_0 = \pi$, the position of the new zeros $x^{(j)}_j$, for $j = 1, 2, \ldots, m$ with $i = 1, 2$, moves according to Proposition 7.3 and, in this case, the functions p_j are equal to p for every level $j = 0, 1, \ldots, m-1$. Furthermore, the particular choice of p_j implies that the coarse matrices have the same symbol of the finer matrix, i.e., $f_j = f_0$ for $j = 1, \ldots, m$. Tables 7.1 and 7.2 report the number of iterations required for convergence in the case of circulant and Toeplitz systems, respectively. According to the convergence analysis in Section 7.3, Table 7.1 shows an optimal behavior of the TGM and W-cycle. Moreover, Table 7.1 shows an optimal behavior also of the V-cycle that is not covered from our convergence analysis, but the latter is not surprising since the symbol of all coarse matrices does not change ($f_j = f_0$ for $j = 1, \ldots, m$). Table 7.2 shows that for Toeplitz matrices the optimality is preserved only for the TGM and W-cycle, while for the V-cycle the number of iterations slightly grows with the size n. This is probably due to the properties of the projectors described in Subsection 7.4.1, which are employed in order to preserve the Toeplitz structure at the coarser levels.

In the second example we increase the order of the zero at $x = \pi$ by choosing the function

$$f_0(x) = (2 - 2\cos(x))\,(2 + 2\cos(x))^2,$$

which has a zero at $x = 0$ of order 2 and one at $x = \pi$ of order 4. The polynomial $p_0 = p$ defined in (7.18) still satisfies the TGM conditions (7.6) and (7.7). As in the previous example, the functions p_j and f_j do not change at the lower levels, i.e., $p_j = p$ for $j = 0, \ldots, m-1$ while $f_j = f_0$ for $j = 1, \ldots, m$. In Tables 7.3 and 7.4 we report the number of iterations required for convergence in the case of circulant and Toeplitz systems, respectively. Since f_0 has a zero of order 4 the condition number of A_n is of order n^4 and therefore for $n = 3^7$ we

n	# iterations					
	Two-grid		V-cycle		W-cycle	
	$\nu_{\text{pre}} = \nu_{\text{post}} = 1$	$\nu_{\text{pre}} = \nu_{\text{post}} = 2$	$\nu_{\text{pre}} = \nu_{\text{post}} = 1$	$\nu_{\text{pre}} = \nu_{\text{post}} = 2$	$\nu_{\text{pre}} = \nu_{\text{post}} = 1$	$\nu_{\text{pre}} = \nu_{\text{post}} = 2$
$3^4 = 81$	11	6	11	6	11	6
$3^5 = 243$	11	6	11	7	11	6
$3^6 = 729$	11	6	11	7	11	6
$3^7 = 2187$	11	6	11	7	11	6

Table 7.1: Circulant case: $f_0(x) = (2 - 2\cos(x))(2 + 2\cos(x))$.

n	# iterations					
	Two-grid		V-cycle		W-cycle	
	$\nu_{\text{pre}} = \nu_{\text{post}} = 1$	$\nu_{\text{pre}} = \nu_{\text{post}} = 2$	$\nu_{\text{pre}} = \nu_{\text{post}} = 1$	$\nu_{\text{pre}} = \nu_{\text{post}} = 2$	$\nu_{\text{pre}} = \nu_{\text{post}} = 1$	$\nu_{\text{pre}} = \nu_{\text{post}} = 2$
$3^4 - 3 = 78$	24	14	24	14	24	14
$3^5 - 3 = 240$	24	15	35	20	28	16
$3^6 - 3 = 726$	24	15	43	24	29	16
$3^7 - 3 = 2184$	24	15	49	27	29	16

Table 7.2: Toeplitz case. $f_0(x) = (2 - 2\cos(x))(2 + 2\cos(x))$.

n	# iterations					
	Two-grid		V-cycle		W-cycle	
	$\nu_{\text{pre}} = \nu_{\text{post}} = 1$	$\nu_{\text{pre}} = \nu_{\text{post}} = 2$	$\nu_{\text{pre}} = \nu_{\text{post}} = 1$	$\nu_{\text{pre}} = \nu_{\text{post}} = 2$	$\nu_{\text{pre}} = \nu_{\text{post}} = 1$	$\nu_{\text{pre}} = \nu_{\text{post}} = 2$
$3^4 = 81$	20	9	20	9	20	9
$3^5 = 243$	20	9	18	9	20	9
$3^6 = 729$	20	9	18	9	20	9
$3^7 = 2187$	20	9	18	9	20	9

Table 7.3: Circulant case. $f_0(x) = (2 - 2\cos(x))(2 + 2\cos(x))^2$, tolerance=$10^{-3}$.

n	# iterations					
	Two-grid		V-cycle		W-cycle	
	$\nu_{\text{pre}} = \nu_{\text{post}} = 1$	$\nu_{\text{pre}} = \nu_{\text{post}} = 2$	$\nu_{\text{pre}} = \nu_{\text{post}} = 1$	$\nu_{\text{pre}} = \nu_{\text{post}} = 2$	$\nu_{\text{pre}} = \nu_{\text{post}} = 1$	$\nu_{\text{pre}} = \nu_{\text{post}} = 2$
$3^4 - 3 = 78$	50	31	50	31	50	31
$3^5 - 3 = 240$	48	31	93	35	72	32
$3^6 - 3 = 726$	47	31	74	34	68	31
$3^7 - 3 = 2184$	47	31	76	34	68	31

Table 7.4: Toeplitz case. $f_0(x) = (2 - 2\cos(x))(2 + 2\cos(x))^2$, tolerance=$10^{-3}$.

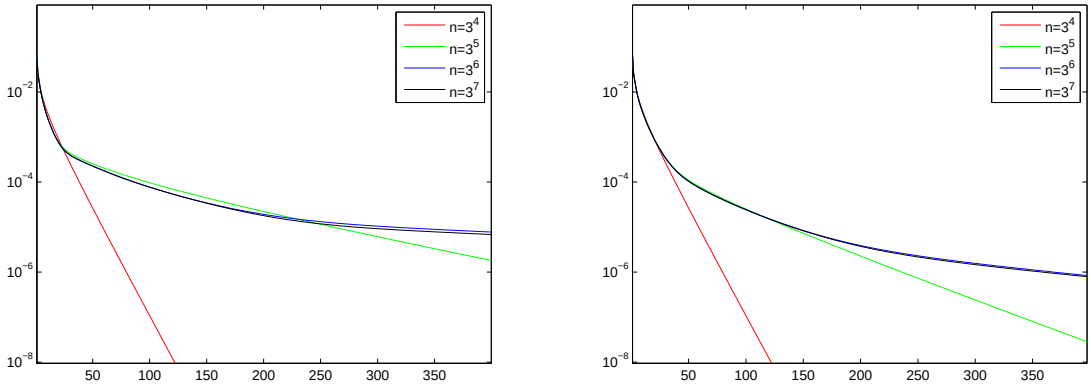

Figure 7.1: Circulant: Graph of the residual in logarithmic scale of the V-cycle (left) and W-cycle (right) with different sizes n, with $\nu_{\text{pre}} = \nu_{\text{post}} = 1$ and a fixed number of iterations $iter = 400$; $f_0(x) = (2 - 2\cos(x))(2 + 2\cos(x))^2$.

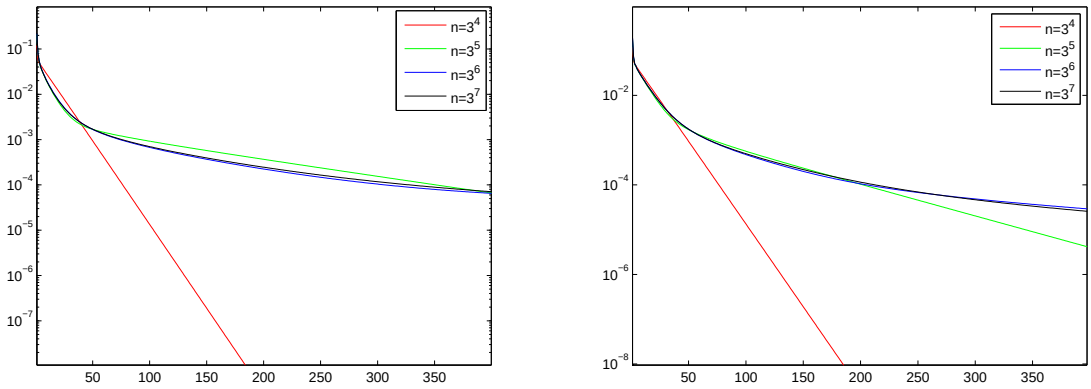

Figure 7.2: Toeplitz: Graph of the residual in logarithmic scale of the V-cycle (left) and W-cycle (right) with different sizes n, with $\nu_{\text{pre}} = \nu_{\text{post}} = 1$ and a fixed number of iterations $iter = 400$; $f_0(x) = (2 - 2\cos(x))(2 + 2\cos(x))^2$.

have a conditioning of magnitude 10^{13}. Therefore, using double precision, for this example we choose a tolerance equal to 10^{-3}. This choice agrees also with the plots in Figures 7.1 and 7.2 where we note an optimal reduction of the residual norm only until about 10^{-3}.

The last example of this subsection is taken from [25]. The generating function

$$f_0(x) = 6 - 4\cos(2x) - 2\cos(4x),$$

vanishes at $x = 0$ and at $x = \pi$ with order 2. The symbol of the projector is again $p_0 = p$ defined in (7.18). The initial guess is a random vector u such that $0 \le u_j \le 1$, the pre-smoother is a step of damped Jacobi with parameter $\omega_j = \frac{(A_j)_{1,1}}{\|f(x)\|_{L^\infty}}$ while the post-smoother is a step of damped Jacobi with parameter $\omega_j = \frac{(A_j)_{1,1}}{\|f(x)\|_{L^\infty}}$ for $j = 0, \ldots, m-1$. The coarser problem is fixed such that it has size lower than 6. Table 7.5 shows that the number of iterations required to achieve the tolerance 10^{-7} remains constant, when increasing the size n of the system like for the multigrid technique proposed in [25], even if the number of iterations is slightly higher. Anyway a direct comparison can not be easily done because of the different cost of each iteration, due to the different size reduction, as discussed in Subsection 7.3.3.

n	# iterations		
	Two-grid	W-cycle	V-cycle
$3^4 - 3 = 78$	15	19	28
$3^5 - 3 = 240$	15	20	39
$3^6 - 3 = 726$	14	20	45
$3^7 - 3 = 2184$	13	20	47

Table 7.5: Toeplitz case. $f_0(x) = 6 - 4\cos(2x) - 2\cos(4x)$, $\nu_{\text{pre}} = \nu_{\text{post}} = 1$, tolerance=$10^{-7}$.

n	# iterations			
	V-cycle		W-cycle	
	$\nu_{\text{pre}} = \nu_{\text{post}} = 1$	$\nu_{\text{pre}} = \nu_{\text{post}} = 2$	$\nu_{\text{pre}} = \nu_{\text{post}} = 1$	$\nu_{\text{pre}} = \nu_{\text{post}} = 2$
$3^4 - 1 = 80$	33	37	33	37
$3^5 - 1 = 242$	30	31	30	31
$3^6 - 1 = 728$	30	31	30	31
$3^7 - 1 = 2186$	30	31	30	31

Table 7.6: Toeplitz case. $f_0(x) = (2 - 2\cos(x - \frac{\pi}{3}))$, tolerance=$10^{-7}$.

7.4.3 Some Toeplitz examples

In this subsection we consider only the more interesting case for practical applications: Toeplitz matrices with a multigrid strategy.

The first example is a function with a zero not at the origin or π:

$$f_0(x) = \left(2 - 2\cos\left(x - \frac{\pi}{3}\right)\right),$$

which vanishes at $x = \frac{\pi}{3}$ with order 2. Moreover, we choose as true solution a random vector instead of a smooth solution. The tolerance is again 10^{-7}. The symbol of the projector at the first level is

$$p_0(x) = (2 - 2\cos(x - \pi))\left(2 - 2\cos\left(x - \frac{5}{3}\pi\right)\right).$$

At the lower levels, the symbols of the projectors will change according to the position of the zero of f_j, which remains unique and which moves as described in Proposition 7.3. Table 7.6 shows an optimal convergence both for the V-cycle and for the W-cycle.

In the second example, we consider the dense Toeplitz matrix generated by the function

n	# iterations			
	V-cycle		W-cycle	
	$\nu_{\text{pre}} = \nu_{\text{post}} = 1$	$\nu_{\text{pre}} = \nu_{\text{post}} = 2$	$\nu_{\text{pre}} = \nu_{\text{post}} = 1$	$\nu_{\text{pre}} = \nu_{\text{post}} = 2$
$3^4 - 1 = 80$	21	11	21	11
$3^5 - 1 = 242$	18	11	21	11
$3^6 - 1 = 728$	18	11	21	11
$3^7 - 1 = 2186$	18	11	21	11

Table 7.7: Toeplitz case. $f(x) = x^2$.

$f(x) = x^2$, which has the Fourier series expansion

$$f(x) = \frac{\pi^2}{3} - 4\left(\frac{\cos(x)}{1^2} - \frac{\cos(2x)}{2^2} + \frac{\cos(3x)}{3^2} - \cdots\right).$$

Such function shows a unique zero at $x = 0$ of order 2 and hence we use the projector with symbol

$$p_0(x) = \left(2 - 2\cos\left(x - \frac{2}{3}\pi\right)\right)\left(2 - 2\cos\left(x - \frac{4}{3}\pi\right)\right).$$

In Table 7.7 we report the number of iterations required for the convergence, with a preassigned accuracy and we note again an optimal behavior.

Chapter 8

Spectral analysis of inexact constraint preconditioning for symmetric saddle point matrices

Large scale non-singular linear systems whose symmetric coefficient matrix has the following saddle point structure

$$\mathcal{A} = \begin{bmatrix} A & B^\top \\ B & -C \end{bmatrix}, \tag{8.1}$$

with $A \in M_n(\mathbb{R})$, $C \in M_m(\mathbb{R})$, with $m \leq n$ and A, C positive definite and semidefinite, respectively, arise in a wide variety of applications in science and engineering; we refer to [10] for a thorough account on the origin of the problem, and on the description of many solution strategies. Conveniently exploiting the matrix structure allows one to devise computational effective acceleration procedures that makes it possible to solve really large two and three-dimensional application problems. In particular, structure-based preconditioners have become a formidable acceleration device, as they can naturally exploit a-priori information on, e.g., the operators leading to the blocks A, B and C. In this chapter we investigate the spectral properties of preconditioned matrices $\mathcal{A}\mathcal{P}^{-1}$, where $\mathcal{P}$ is given by

$$\begin{aligned} \mathcal{P} &= \begin{bmatrix} I_n & B^\top \\ B & -C \end{bmatrix} \\ &= \begin{bmatrix} I_n & 0 \\ B & I_m \end{bmatrix} \begin{bmatrix} I_n & 0 \\ 0 & -H \end{bmatrix} \begin{bmatrix} I_n & B^\top \\ 0 & I_m \end{bmatrix}, \qquad H = BB^\top + C. \end{aligned} \tag{8.2}$$

Here and in the following we assume that A was already preconditioned by means of a preprocessing (e.g. block diagonal preconditioning), so that the (1,1) block in the preconditioner can be taken to be the identity matrix, moreover $I_m, 0_m \in M_m(\mathbb{R})$ denote the identity and zero matrices, respectively; the subscript will be omitted whenever the matrix dimension is clear from the context. The dimensions of zero rectangular matrices will also be deduced from the context.

It can be shown that when solving the transformed problem $\mathcal{A}\mathcal{P}^{-1}$ by means of an iterative method, the iterates satisfy the original constraint (given by the second block-row in the associated system). This major property has made $\mathcal{P}$ very popular in the optimization community, and a lot is now known on the eigenvalue properties of $\mathcal{A}\mathcal{P}^{-1}$, which to a large extent seem to guide the convergence of the iterative system solver; see, e.g., the theoretical developments and algorithmic consequences in [40, 74, 58, 31, 35, 32, 7]. As a key spectral feature, we recall that $\mathcal{A}\mathcal{P}^{-1}$ has all real eigenvalues, and that m Jordan blocks ([105]) of size 2 corresponding to the unit eigenvalue also arise. Moreover, some eigenvector structure was analyzed in [31]. Many

numerical experiments have shown that using $\mathcal{P}$ may be very competitive in various applications, and that the non-symmetry of the resulting matrix, namely $\mathcal{A}\mathcal{P}^{-1}$, does not represent a major problem.

The application of $\mathcal{P}^{-1}$, however, requires solving systems with $H = BB^\top + C$ (see the block form in (8.2)). Explicitly solving with the exact H may be unrealistic when dealing with 3D applications, so that a cheap approximation to H is used, giving rise to an *inexact* constraint preconditioner; see, e.g., [74, 51]. This necessary step seems to jeopardize the whole theory of constraint preconditioning, as in general complex eigenvalues arise and may spread well away from the original values obtained when H is used. This wide spreading is mainly caused by the perturbation of the multiple unit eigenvalue, which is expected to be *non-linear* in the error made in approximating H. Such a phenomenon has created some concern as of the adequacy of this preconditioning procedure in the inexact case, although experimental evidence shows otherwise. Theoretical ground supporting this optimistic numerical experience has remained for long time a largely open issue. First attempts to analyze the spectral modification occurring when using an inexact form of H can be found in [11, 74], but a thorough understanding is still missing. Here we aim to fill this gap. For ease of presentation, we shall use $\mathcal{P}_{\text{ex}} = \mathcal{P}$ to denote the exact preconditioner (namely exact solves with H), and

$$\mathcal{P}_{\text{inex}} = \begin{bmatrix} I_n & 0 \\ B & I_m \end{bmatrix} \begin{bmatrix} I_n & 0 \\ 0 & -H_{\text{inex}} \end{bmatrix} \begin{bmatrix} I_n & B^\top \\ 0 & I_m \end{bmatrix}, \tag{8.3}$$

where H_{inex} is a symmetric and positive definite matrix such that $H_{\text{inex}} \approx H$.

In this chapter we provide a complete spectral characterization of $\mathcal{A}\mathcal{P}_{\text{ex}}^{-1}$ by means of the Weyr canonical form, which highlights the role of the matrix blocks and of the multiple unit eigenvalue. Moreover, we express $\mathcal{A}\mathcal{P}_{\text{inex}}^{-1}$ as a perturbation of $\mathcal{A}\mathcal{P}_{\text{ex}}^{-1}$, and this allows us to easily track the perturbation of its eigenvalues. We can also naturally derive estimates for the condition number of the transformation matrix in the canonical form of $\mathcal{A}\mathcal{P}_{\text{ex}}^{-1}$; these estimates can be used to fully monitor the convergence of an optimal iterative solver such as GMRES [78]; we refer to [75] for early specialized results in this direction, for $C = 0$.

We first discuss the case $C = 0$, and then highlight the difference occurring when C is non-zero. In fact, the original Jordan form can be significantly modified for $C \neq 0$, possibly leading to more favourable properties of the modified preconditioner.

For the sake of simplicity in the exposition we assume that $A - I$ is non-singular.

8.1 Case $C = 0$. Exact constraint preconditioner

The eigenvalues of the preconditioned coefficient matrix $\mathcal{A}\mathcal{P}_{\text{ex}}^{-1}$ may be derived by considering the general eigenvalue problem

$$\begin{bmatrix} A & B^\top \\ B & 0 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \lambda \mathcal{P}_{\text{ex}} \begin{bmatrix} x \\ y \end{bmatrix}, \tag{8.4}$$

which can be written as

$$\begin{bmatrix} I_n & 0 \\ -B & I_m \end{bmatrix} \begin{bmatrix} A & B^\top \\ B & 0 \end{bmatrix} \begin{bmatrix} I_n & -B^\top \\ 0 & I_m \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix} = \lambda \begin{bmatrix} I_n & 0 \\ 0 & -H \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix}, \tag{8.5}$$

where the factorization (8.2) is used, and

$$\begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} I_n & B^\top \\ 0 & I_m \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}.$$

The left-hand side matrix yields

$$\begin{bmatrix} I_n & 0 \\ -B & I_m \end{bmatrix} \begin{bmatrix} A & B^\top \\ B & 0_m \end{bmatrix} \begin{bmatrix} I_n & -B^\top \\ 0 & I_m \end{bmatrix} = \begin{bmatrix} A & (I-A)\,B^\top \\ B\,(I-A) & -B\,(2I-A)\,B^\top \end{bmatrix}. \tag{8.6}$$

After changing sign in the second block-row in (8.5), we can write (8.6) as

$$\begin{aligned}\begin{bmatrix} A & (I-A)B^\top \\ -B(I-A) & B(2I-A)B^\top \end{bmatrix} &= \begin{bmatrix} (A-I) & (I-A)B^\top \\ -B(I-A) & B(I-A)B^\top \end{bmatrix} + \begin{bmatrix} I_n & 0 \\ 0 & H \end{bmatrix} \\ &= \begin{bmatrix} I_n \\ B \end{bmatrix} (A-I) \begin{bmatrix} I_n, & -B^\top \end{bmatrix} + \begin{bmatrix} I_n & 0 \\ 0 & H \end{bmatrix}.\end{aligned}$$

Then the eigenvalue problem (8.4) can be transformed into

$$\left(\begin{bmatrix} I_n \\ B \end{bmatrix} (A-I) \begin{bmatrix} I_n, & -B^\top \end{bmatrix} \right) \begin{bmatrix} u \\ v \end{bmatrix} = (\lambda - 1) \begin{bmatrix} I_n & 0 \\ 0 & H \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix},$$

or equivalently

$$\mathcal{M}_0 w = (\lambda - 1) w \;\Leftrightarrow\; \left(\begin{bmatrix} I_n \\ \widehat{B} \end{bmatrix} (A-I) \left[I_n, -\widehat{B}^\top \right] \right) \begin{bmatrix} u \\ \widehat{v} \end{bmatrix} = (\lambda - 1) \begin{bmatrix} u \\ \widehat{v} \end{bmatrix}, \tag{8.7}$$

where $\widehat{B} = H^{-\frac{1}{2}} B$ and $\widehat{v} = H^{\frac{1}{2}} v$.

For later results it is important to note that the matrix $\widehat{B}^\top = B^\top H^{-\frac{1}{2}}$ has orthonormal columns, so that $\left(I - \widehat{B}^\top \widehat{B} \right)$ is an orthogonal projector.

The following proposition describes a particular form of Jordan decomposition of the matrix $\mathcal{M}_0$, the so-called Weyr canonical form [102], which allows us to derive a clear block structure of the transformation matrix. The complete Weyr canonical form of $\mathcal{A}\mathcal{P}_{\text{ex}}^{-1}$ can be then easily derived, as shown in the subsequent theorem.

Proposition 8.1. *Let $(A-I)\left(I - \widehat{B}^\top \widehat{B} \right) \widehat{X} = \widehat{X}\Theta$ be the partial eigenvalue decomposition of $(A-I)\left(I - \widehat{B}^\top \widehat{B} \right)$ associated with its non-zero eigenvalues. Then the Weyr decomposition of $\mathcal{M}_0$ is given by*

$$\mathcal{M}_0 \mathcal{X}_0 = \mathcal{X}_0 \begin{bmatrix} \Theta & & \\ & 0_m & I_m \\ & & 0_m \end{bmatrix}, \qquad \mathcal{X}_0 = \left[\begin{array}{c|c|c} \widehat{X} & \widehat{B}^\top & (A-I)^{-1}\widehat{B}^\top \\ \widehat{B}\widehat{X} & I_m & 0_m \end{array}\right], \tag{8.8}$$

with $\mathcal{X}_0$ non-singular.

Proof. We start by observing that $\Lambda(\mathcal{M}_0) = \{0\} \cup \Lambda\left((A-I)\left(I - \widehat{B}^\top \widehat{B} \right) \right)$. Indeed, the matrix $\mathcal{M}_0 \in M_{(n+m)}(\mathbb{R})$ has rank at most n, therefore it has at least m zero eigenvalues. Moreover, we recall that the non-zero eigenvalues of a low-rank matrix $XY^\top$ are the same as those of the matrix $Y^\top X$ (see Theorem 1.4). Therefore, the remaining n eigenvalues of $\mathcal{M}_0$ are given by those of the matrix

$$(A-I)\left[I_n, -\widehat{B}^\top \right] \begin{bmatrix} I_n \\ \widehat{B} \end{bmatrix} = (A-I)\left(I - \widehat{B}^\top \widehat{B} \right).$$

We then need to establish the dimension of Θ. Since $(A-I)$ is non-singular and $\left(I - \widehat{B}^\top \widehat{B} \right)$ is an orthogonal projector with m zero eigenvalues, the number of non-zero eigenvalues of $(A-I)\left(I - \widehat{B}^\top \widehat{B} \right)$ is $n - m$.

We seek the eigenvectors of the matrix $\mathcal{M}_0$ associated with Θ. It turns out that a distinct analysis for each eigenvalue is not needed. We consider the system

$$\begin{bmatrix} I_n \\ \widehat{B} \end{bmatrix} (A-I) \begin{bmatrix} I_n, & -\widehat{B}^\top \end{bmatrix} \begin{bmatrix} X \\ Y \end{bmatrix} = \begin{bmatrix} X \\ Y \end{bmatrix} \Theta,$$

where $X \in M_{n,(n-m)}(\mathbb{C})$ and $Y \in M_{m,(n-m)}(\mathbb{C})$, or block-wise,

$$(A-I)\left(X - \widehat{B}^\top Y\right) = X\Theta, \qquad \widehat{B}\,(A-I)\left(X - \widehat{B}^\top Y\right) = Y\Theta.$$

Substituting $X = \widehat{X}$, the matrix of eigenvectors of $(A-I)\left(I - \widehat{B}^\top \widehat{B}\right)$, and $Y = \widehat{B}\widehat{X}$ in the first equation we obtain $(A-I)\left(I - \widehat{B}^\top \widehat{B}\right)\widehat{X} = \widehat{X}\Theta$ which is clearly verified by the eigenpairs. Substituting X and Y into the second block equation yields the same equation, multiplied by $\widehat{B}$, and thus that is also verified. This gives us the first block of $n-m$ columns in the matrix $\mathcal{X}_0$.

Now we look for the eigenvectors of the matrix $\mathcal{M}_0$ associated with the zero eigenvalue. We consider the system

$$\begin{bmatrix} I_n \\ \widehat{B} \end{bmatrix} (A-I) \begin{bmatrix} I_n, & -\widehat{B}^\top \end{bmatrix} \begin{bmatrix} X \\ Y \end{bmatrix} = 0,$$

where $X \in M_{n,k}(\mathbb{C})$, $Y \in M_{m,k}(\mathbb{C})$, $k \leq 2m$, or, equivalently,

$$(A-I)\left(X - \widehat{B}^\top Y\right) = 0, \qquad \widehat{B}\,(A-I)\left(X - \widehat{B}^\top Y\right) = 0. \tag{8.9}$$

Since $(A-I)$ is invertible, the first matrix equation in (8.9) is satisfied if and only if $\left(X - \widehat{B}^\top Y\right) = 0$, that is for $X = \widehat{B}^\top Y$. The matrix $Y \in M_{m,k}(\mathbb{C})$, $k \leq 2m$, will consist of at most $k = m$ linearly independent vectors, then we can take $Y = I_m$ and hence $X = \widehat{B}^\top$. We have thus found m eigenvectors of $\mathcal{M}_0$ associated with the zero eigenvalue, matching the second block of columns of $\mathcal{X}_0$. The remaining m vectors must be generalized eigenvectors associated with the zero eigenvalue. We consider the system

$$\begin{bmatrix} I_n \\ \widehat{B} \end{bmatrix} (A-I) \begin{bmatrix} I_n, & -\widehat{B}^\top \end{bmatrix} \begin{bmatrix} X_1 & X_2 \\ Y_1 & Y_2 \end{bmatrix} = \begin{bmatrix} X_1 & X_2 \\ Y_1 & Y_2 \end{bmatrix} \begin{bmatrix} 0_k & I_k \\ & 0_k \end{bmatrix},$$

where $X_1, X_2 \in M_{n,k}(\mathbb{C})$, $Y_1, Y_2 \in M_{m,k}(\mathbb{C})$, $0_k, I_k \in M_k(\mathbb{R})$, $k \leq m$, and we write the two systems of equations associated

$$\begin{cases} (A-I)\left(X_1 - \widehat{B}^\top Y_1\right) = 0 \\ \widehat{B}\,(A-I)\left(X_1 - \widehat{B}^\top Y_1\right) = 0, \end{cases} \qquad \begin{cases} (A-I)\left(X_2 - \widehat{B}^\top Y_2\right) = X_1 \\ \widehat{B}\,(A-I)\left(X_2 - \widehat{B}^\top Y_2\right) = Y_1. \end{cases}$$

The first system is just (8.9) and admits solution $Y_1 = I_m$ and $X_1 = \widehat{B}^\top$, then $k = m$, and it remains to solve the second system that block-wise reads

$$(A-I)\left(X_2 - \widehat{B}^\top Y_2\right) = \widehat{B}^\top, \qquad \widehat{B}\,(A-I)\left(X_2 - \widehat{B}^\top Y_2\right) = I_m.$$

Since $\widehat{B}\widehat{B}^\top = I_m$, the latter system is satisfied by setting $Y_2 = 0_m$ and $X_2 = (A-I)^{-1}\,\widehat{B}^\top$, thus completing the matrix $\mathcal{X}_0$ and the resulting Weyr form in (8.8).

We complete by showing that $\mathcal{X}_0$ is non-singular. Let $\widehat{Z}_0$, $\widehat{Z}$ be the matrices of left eigenvectors of $(A-I)\left(I - \widehat{B}^\top \widehat{B}\right)$ associated with zero and non-zero eigenvalues, respectively. In particular, we recall that $\widehat{B}^\top$ has orthonormal columns and we notice that $\widehat{Z}_0^* \widehat{B}^\top$ must be non-singular, since $\left[\widehat{Z}_0, \widehat{Z}\right]$ is full-rank and the columns of $\widehat{Z}$ span range$\left(I - \widehat{B}^\top \widehat{B}\right)$. Then, using the full-rank assumption of $\widehat{B}\,(A-I)^{-1}\,\widehat{B}^\top$, $\mathcal{X}_0^{-1}$ is given by

$$\mathcal{X}_0^{-1} = \begin{bmatrix} \Theta^{-1}\left(\widehat{Z}^*\widehat{X}\right)^{-1}\widehat{Z}^*\,(A-I) & -\Theta^{-1}\left(\widehat{Z}^*\widehat{X}\right)^{-1}\widehat{Z}^*\,(A-I)\,\widehat{B}^\top \\ \left(\widehat{Z}_0^*\widehat{B}^\top\right)^{-1}\widehat{Z}_0^*\left(I - (A-I)^{-1}\,\widehat{B}^\top G^{-1}\widehat{B}\right) & \left(\widehat{Z}_0^*\widehat{B}^\top\right)^{-1}\widehat{Z}_0^*\,(A-I)^{-1}\,\widehat{B}^\top G^{-1} \\ G^{-1}\widehat{B} & -G^{-1} \end{bmatrix},$$

where $G = \widehat{B}\,(A-I)^{-1}\,\widehat{B}^\top$. Explicit multiplication by $\mathcal{X}_0$ verifies the assertion. □

Theorem 8.2. *With the notation and assumptions of Proposition 8.1, the preconditioned matrix $\mathcal{A}\mathcal{P}_{\text{ex}}^{-1}$ admits the following Weyr decomposition:*

$$\mathcal{A}\mathcal{P}_{\text{ex}}^{-1}\mathcal{X} = \mathcal{X}\begin{bmatrix} I_{n-m}+\Theta & & \\ & I_m & I_m \\ & & I_m \end{bmatrix}, \quad \mathcal{X} = \left[\begin{array}{c|c|c} \widehat{X} & \widehat{B}^\top & (A-I)^{-1}\widehat{B}^\top \\ 0 & 0_m & B(A-I)^{-1}\widehat{B}^\top \end{array}\right],$$

with $\mathcal{X}$ non-singular.

Proof. Let $\mathcal{A}\mathcal{P}_{\text{ex}}^{-1}z = \lambda z$. The derivation leading to (8.7) shows that

$$\mathcal{X} = \begin{bmatrix} I_n & 0 \\ B & -H^{\frac{1}{2}} \end{bmatrix}\mathcal{X}_0.$$

The result readily follows from recalling that the eigenvalues of $\mathcal{A}\mathcal{P}_{\text{ex}}^{-1}$ are given by those of $\mathcal{M}_0$ plus one. □

Direct inspection shows that the inverse of $\mathcal{X}$ can be explicitly written as

$$\mathcal{Y}^* := \mathcal{X}^{-1} = \left[\begin{array}{cc} \widehat{X}^\dagger & -\widehat{X}^\dagger(A-I)^{-1}B^\top G^{-1} \\ \hline H^{\frac{1}{2}}G^{-1}B(A-I)^{-1} & -H^{\frac{1}{2}}G^{-1}B(A-I)^{-2}B^\top G^{-1} \\ \hline 0 & H^{\frac{1}{2}}G^{-1} \end{array}\right]; \tag{8.10}$$

here $G = B(A-I)^{-1}B^\top$, and $\widehat{X}^\dagger$ is the (row) portion of the inverse eigenvector matrix of $(A-I)\left(I - B^\top H^{-1}B\right)$ associated with the non-zero eigenvalues, or the properly scaled conjugate transpose left eigenvector matrix.

The subdivision of $\mathcal{X}$ in three block-columns, and of $\mathcal{Y}^*$ in the corresponding block-rows is used to readily describe the role of each block, which the Weyr decomposition easily emphasizes. The first column block of $\mathcal{X}$ contains all eigenvectors of $\Theta+I_{n-m}$, while the second block collects the m eigenvectors associated with the unit eigenvalue with geometric multiplicity m. The third block is associated with the corresponding m generalized eigenvectors, revealing the occurrence of 2×2 Jordan blocks.

Finally, we notice that with the explicit expression of $\mathcal{X}$ and $\mathcal{X}^{-1}$ at hand, it is possible to give estimates for the condition number of $\mathcal{X}$, which can be used, together with the Weyr decomposition, for estimating the residual norm of optimal Krylov subspace iterative solvers; see [104, Section 6] and the references therein.

8.2 Case $C = 0$. Inexact constraint preconditioner

We start by showing that the inexactly preconditioned problem can be written as a perturbation of the exactly preconditioned one. This formulation will allow us to exploit classical perturbation theory results to derive the desired spectral perturbation bounds.

Theorem 8.3. *With the notation of the previous section, it holds that*

$$\mathcal{A}\mathcal{P}_{\text{inex}}^{-1} = \mathcal{A}\mathcal{P}_{\text{ex}}^{-1} + \mathcal{E}, \qquad \textit{with} \qquad \mathcal{E} = -\mathcal{A}\begin{bmatrix} B^\top \\ -I_m \end{bmatrix} H^{-1}EH_{\text{inex}}^{-1}\left[B,\, -I_m\right].$$

Proof. Let $E = BB^\top - H_{\text{inex}}$. We have

$$\mathcal{P}_{\text{inex}} = \begin{bmatrix} I_n & B^\top \\ B & E \end{bmatrix} = \mathcal{P}_{\text{ex}} + \begin{bmatrix} 0 \\ E \end{bmatrix}[0,\ I_m] = \left(I_{n+m} + \begin{bmatrix} 0 \\ E \end{bmatrix}[0,\ I_m]\,\mathcal{P}_{\text{ex}}^{-1}\right)\mathcal{P}_{\text{ex}}.$$

Therefore,

$$\mathcal{A}\mathcal{P}_{\text{inex}}^{-1} = \mathcal{A}\mathcal{P}_{\text{ex}}^{-1}\left(I_{n+m} + \begin{bmatrix}0\\E\end{bmatrix}[0,\ I_m]\,\mathcal{P}_{\text{ex}}^{-1}\right)^{-1}.$$

Thanks to the Sherman-Morrison formula [46], and using $\mathcal{P}_{\text{ex}}^{-1}\begin{bmatrix}0\\E\end{bmatrix} = \begin{bmatrix}B^\top\\-I_m\end{bmatrix}H^{-1}E$, we obtain

$$\begin{aligned}\mathcal{A}\mathcal{P}_{\text{inex}}^{-1} &= \mathcal{A}\mathcal{P}_{\text{ex}}^{-1}\left(I_{n+m} - \begin{bmatrix}0\\E\end{bmatrix}\left(I_m + [0,\ I_m]\,\mathcal{P}_{\text{ex}}^{-1}\begin{bmatrix}0\\E\end{bmatrix}\right)^{-1}[0,\ I_m]\,\mathcal{P}_{\text{ex}}^{-1}\right)\\ &= \mathcal{A}\mathcal{P}_{\text{ex}}^{-1} - \mathcal{A}\begin{bmatrix}B^\top\\-I_m\end{bmatrix}H^{-1}EH_{\text{inex}}^{-1}[B,\ -I_m],\end{aligned}$$

which is the sought after relation. □

We also notice that

$$\begin{aligned}\|\mathcal{E}\| &\le \|\mathcal{A}\|\left\|\begin{bmatrix}B^\top\\-I_m\end{bmatrix}H^{-\frac{1}{2}}\right\|\left\|H^{-\frac{1}{2}}EH_{\text{inex}}^{-\frac{1}{2}}\right\|\left\|H_{\text{inex}}^{-\frac{1}{2}}[B,\ -I_m]\right\|\\ &= O\left(\|\mathcal{A}\|\left\|H^{-\frac{1}{2}}EH_{\text{inex}}^{-\frac{1}{2}}\right\|\right),\end{aligned}$$

which provides a clear relation between $\|\mathcal{E}\|$ and $\left\|H^{-\frac{1}{2}}EH_{\text{inex}}^{-\frac{1}{2}}\right\| \approx \|H^{-1}E\|$.

To proceed with the spectral analysis, we must distinguish between the unit and non-unit eigenvalues, since the occurrence of multiple eigenvalues with Jordan blocks requires a refined analysis.

We assume that the eigenvalues of the diagonal matrix $I_{n-m} + \Theta$ in $\mathcal{A}\mathcal{P}_{\text{ex}}^{-1}$ are all distinct. Therefore, we can exploit standard perturbation results to evaluate the perturbation that these simple eigenvalues undergo when $\mathcal{A}\mathcal{P}_{\text{ex}}^{-1}$ is perturbed by $\mathcal{E}$ [105]. For each simple eigenvalue $\lambda\left(\mathcal{A}\mathcal{P}_{\text{ex}}^{-1}\right)$ there exists an eigenvalue $\lambda\left(\mathcal{A}\mathcal{P}_{\text{ex}}^{-1} + \mathcal{E}\right)$ such that

$$\lambda\left(\mathcal{A}\mathcal{P}_{\text{ex}}^{-1} + \mathcal{E}\right) = \lambda\left(\mathcal{A}\mathcal{P}_{\text{ex}}^{-1}\right) + \frac{y^*\mathcal{E}x}{y^*x} + O\left(\|\mathcal{E}\|^2\right), \tag{8.11}$$

where x, y are the right and left eigenvectors associated with $\lambda\left(\mathcal{A}\mathcal{P}_{\text{ex}}^{-1}\right)$. Since both the right and left eigenvectors are available, namely they are the columns of the first block of $\mathcal{X}$ (see Theorem 8.2) and of $\mathcal{Y}$ (see (8.10)), the first-order term can be explicitly computed.

To analyze the perturbation of the unit eigenvalues, we first notice that some of them may not be perturbed at all, as the following theorem shows.

Theorem 8.4. *Assume that $E = H - H_{\text{inex}}$ has $k \le m$ zero eigenvalues. Then $\mathcal{A}\mathcal{P}_{\text{inex}}^{-1}$ has $2k$ unit eigenvalues with geometric multiplicity k.*

Proof. Let $x \neq 0$ be such that $Ex = 0$. Setting $y = H^{\frac{1}{2}}x$, from $Hx = H_{\text{inex}}x$ we obtain $y = H^{-\frac{1}{2}}H_{\text{inex}}H^{-\frac{1}{2}}y$. Therefore, using any linear combination of columns from the second block of $\mathcal{X}$, vectors of the form $\begin{bmatrix}\widehat{B}^\top y\\0\end{bmatrix}$ yield

$$\begin{aligned}\mathcal{E}\begin{bmatrix}\widehat{B}^\top y\\0\end{bmatrix} &= -\mathcal{A}\begin{bmatrix}B^\top\\-I_m\end{bmatrix}H^{-1}EH_{\text{inex}}^{-1}H^{\frac{1}{2}}y\\ &= -\mathcal{A}\begin{bmatrix}B^\top\\-I_m\end{bmatrix}H^{-\frac{1}{2}}\left(y - H^{-\frac{1}{2}}H_{\text{inex}}H^{-\frac{1}{2}}y\right) = 0,\end{aligned}$$

so that

$$\mathcal{AP}_{\text{inex}}^{-1} \begin{bmatrix} \widehat{B}^\top y \\ 0 \end{bmatrix} = \mathcal{AP}_{\text{ex}}^{-1} \begin{bmatrix} \widehat{B}^\top y \\ 0 \end{bmatrix} = \begin{bmatrix} \widehat{B}^\top y \\ 0 \end{bmatrix}.$$

Since the dimension of the null space of E is equal to k, the relations above show that $\mathcal{AP}_{\text{inex}}^{-1}$ has k eigenvalues equal to one, with corresponding linearly independent eigenvectors. We also notice that the third block column of $\mathcal{X}$ is in the null space of $\mathcal{E}$, namely

$$\mathcal{E} \begin{bmatrix} (A-I)^{-1} \widehat{B}^\top \\ B (A-I)^{-1} \widehat{B}^\top \end{bmatrix} = 0,$$

so that any k columns of this block represent a set of generalized eigenvectors for the unit eigenvalue of $\mathcal{AP}_{\text{inex}}^{-1}$. □

Theorem 8.4 shows that if H_{inex} is spectrally close to H, in the sense that their eigenstructure partially coincides, then E is singular and it only partially affects the Jordan structure of the unperturbed problem.

The remaining $2(m-k)$ unit eigenvalues, with $k \geq 0$, may be perturbed by a possibly much larger quantity than the simple eigenvalues. In particular, for a multiple eigenvalue with all Jordan blocks of size two it holds that (see, e.g., [65, 61])

$$\lambda\left(\mathcal{AP}_{\text{ex}}^{-1} + \mathcal{E}\right) = \lambda\left(\mathcal{AP}_{\text{ex}}^{-1}\right) + \xi^{\frac{1}{2}} + o\left(\|\mathcal{E}\|^{\frac{1}{2}}\right), \tag{8.12}$$

where ξ are the eigenvalues of $Y^* \mathcal{E} X$ (and both positive and negative square roots are used to compute the first-order term); the columns of X, Y contain all right and left eigenvectors associated with the unit eigenvalue of $\mathcal{AP}_{\text{ex}}^{-1}$. In our case, X (Y) is the second block-column of $\mathcal{X}$ (of $\mathcal{Y}$). Once again, thanks to the explicit expression of $\mathcal{X}$ and $\mathcal{Y}$, all these quantities can be readily computed.

Although we do not dwell here with the subject, we mention that [65, 61] also discuss the non-linear perturbation of eigenvectors, by providing the zero-order perturbation term explicitly. A similar procedure could be applied here; however, the results would be so technical in our setting that they might be difficult to exploit in practice.

We conclude with a result that sheds light into the type of first-order perturbation induced by $Y^* \mathcal{E} X$ (8.12).

Proposition 8.5. *Let X_2, Y_2 be the second block-columns of $\mathcal{X}$ and $\mathcal{Y}$, respectively. Assume that $A - I$ is negative definite. Then the eigenvalues ξ of $Y_2^* \mathcal{E} X_2$ are all real. Moreover, if $E = H - H_{\text{inex}} \geq 0$ ($E = H - H_{\text{inex}} \leq 0$) then $\xi \geq 0$ ($\xi \leq 0$).*

Proof. We show that $Y_2^* \mathcal{E} X_2 = W_1 W_2$ with W_1, W_2 symmetric and $W_1 > 0$. This will ensure that the eigenvalues are real. The definiteness will depend on the definiteness of W_2.

Using the definition of Y_2, $\mathcal{E}$ and X_2, we have

$$\begin{aligned}
Y_2^* \mathcal{E} X_2 &= -H^{\frac{1}{2}} G^{-1} \left(BB^\top - B (A-I)^{-2} B^\top G^{-1} H\right) H^{-1} E H_{\text{inex}}^{-1} H^{\frac{1}{2}} \\
&= -H^{\frac{1}{2}} G^{-1} \left(I - B (A-I)^{-2} B^\top G^{-1}\right) E H_{\text{inex}}^{-1} H^{\frac{1}{2}} \\
&= -H^{\frac{1}{2}} G^{-1} \left(G - B (A-I)^{-2} B^\top\right) G^{-1} E H_{\text{inex}}^{-1} H^{\frac{1}{2}} \\
&= \left(-H^{\frac{1}{2}} G^{-1} B (A-I)^{-1} (A-2I) (A-I)^{-1} B^\top G^{-1}\right) \left(H^{\frac{1}{2}} \left(H_{\text{inex}}^{-1} - H^{-1}\right) H^{\frac{1}{2}}\right) \\
&\equiv W_1 W_2.
\end{aligned}$$

Since $A - 2I < A - I < 0$, it follows that W_1 is positive definite, while the definiteness of W_2 depends on that of $H_{\text{inex}}^{-1} - H^{-1}$. □

$\lambda\left(\mathcal{A}\mathcal{P}_{\text{ex}}^{-1}\right)$	$\lambda\left(\mathcal{A}\mathcal{P}_{\text{inex}}^{-1}\right)$	$\left\lvert\lambda\left(\mathcal{A}\mathcal{P}_{\text{ex}}^{-1}\right)-\lambda\left(\mathcal{A}\mathcal{P}_{\text{inex}}^{-1}\right)\right\rvert$
1	0.99495 - 0.14204i	0.14213
1	0.99495 + 0.14204i	0.14213
1	1.0051 - 0.071435i	0.07162
1	1.0051 + 0.071435i	0.07162
2	1.9798	0.02020

Table 8.1: Example 8.6. Eigenvalues of the exactly and inexactly preconditioned problem, and their difference.

The sign of the eigenvalues ξ influences the type of first-order perturbation of multiple eigenvalues. In particular, if $H \leq H_{\text{inex}}$, then all $\xi \leq 0$, so that $\xi^{\frac{1}{2}}$ are purely imaginary. As a consequence, perturbed unit eigenvalues will all have non-zero imaginary part, and no real eigenvalues occur as first-order perturbations of the unit eigenvalue with non-trivial Jordan block. This is indeed the case for the perturbed spectrum of Example 8.8 when H_{inex} is obtained by an Algebraic Multigrid operator.

8.3 Case $C = 0$. Numerical evidence

In this section we provide experimental validation of our theory. We start with two simple examples with 5×5 matrices, which can be fully replicated.

Example 8.6. We consider the matrix (8.1) with

$$A = \begin{bmatrix} 2 & 1 & 0 \\ 1 & 3 & 1 \\ 0 & 1 & 4 \end{bmatrix}, \qquad B = \begin{bmatrix} 0 & 2 & 0 \\ 0 & 0 & 3 \end{bmatrix}.$$

We precondition this matrix either with $\mathcal{P}_{\text{ex}}$ or with $\mathcal{P}_{\text{inex}}$. In the latter case, the matrix H_{inex} occurring in (8.3) is defined as

$$H_{\text{inex}} = B \begin{bmatrix} 0.985 & 0 & 0 \\ 0 & 0.99 & 0 \\ 0 & 0 & 0.995 \end{bmatrix} B^{\top},$$

yielding a non-singular $E = H - H_{\text{inex}}$, with $\|E\| \approx 4.5 \cdot 10^{-2}$ and $\frac{\|E\|}{\|H\|} \approx 5 \cdot 10^{-3}$. The eigenvalues of $\mathcal{A}\mathcal{P}_{\text{ex}}^{-1}$ and of $\mathcal{A}\mathcal{P}_{\text{inex}}^{-1}$ are reported in Table 8.1, together with their absolute difference.

For the simple eigenvalue $\lambda = 2$ the first-order perturbation in (8.11) predicts a value $\frac{y^*\mathcal{E}x}{y^*x} \approx -0.020202$, which perfectly matches the actual true perturbation. Here x and y^* are the first column of $\mathcal{X}$ and row in $\mathcal{Y}^*$, respectively, while $\mathcal{E}$ is as defined in Theorem 8.3.

For the multiple unit eigenvalue with 2 Jordan blocks of size 2 each, the computed quantity $\xi^{\frac{1}{2}}$ in (8.12) is given by (to the first significant digits) $\xi^{\frac{1}{2}} = 0.11120$ and $\xi^{\frac{1}{2}} = 0.11097\text{i}$. Note that here $\|\mathcal{E}\|^{\frac{1}{2}} = 0.19277$, so that higher order terms will be significantly smaller. As a consequence, the first-order perturbation term $\xi^{\frac{1}{2}}$ provides a sufficiently good correction to the estimate.

Example 8.7. In this example we consider the same data as in Example 8.6, whereas

$$H_{\text{inex}} = B \begin{bmatrix} 0.985 & 0 & 0 \\ 0 & 0.99 & 0 \\ 0 & 0 & 1 \end{bmatrix} B^{\top}.$$

$\lambda\left(\mathcal{AP}_{\text{ex}}^{-1}\right)$	$\lambda\left(\mathcal{AP}_{\text{inex}}^{-1}\right)$	$\left\lvert\lambda\left(\mathcal{AP}_{\text{ex}}^{-1}\right) - \lambda\left(\mathcal{AP}_{\text{inex}}^{-1}\right)\right\rvert$
1	1.0000	$2.220 \cdot 10^{-16}$
1	1.0000	$2.220 \cdot 10^{-16}$
1	1.0050 - 0.10141i	0.10153
1	1.0050 + 0.10141i	0.10153
2	1.9798	0.020198

Table 8.2: Example 8.7. Eigenvalues of the exactly and inexactly preconditioned problem, and their difference.

tol	$\|E\|$	$\|E\| / \|H\|$	$\|\mathcal{E}\|$	# 2×2 Jordan blocks in $\mathcal{AP}_{\text{ex}}^{-1}$	# 2×2 Jordan blocks in $\mathcal{AP}_{\text{inex}}^{-1}$
$1 \cdot 10^{-4}$	1.20	$1.2 \cdot 10^{-4}$	$9.88 \cdot 10^{-2}$	816	93
$5 \cdot 10^{-4}$	9.02	$9.1 \cdot 10^{-4}$	$5.89 \cdot 10^{-1}$	816	65
$1 \cdot 10^{-3}$	24.65	$2.5 \cdot 10^{-3}$	$1.11 \cdot 10^{\ 0}$	816	56
AMG-MI20	164.59	$1.6 \cdot 10^{-2}$	$2.68 \cdot 10^{\ 0}$	816	145

Table 8.3: Example 8.8. Relevant quantities for the inexactly preconditioned problem.

This choice yields a diagonal and singular matrix E, with numerically computed eigenvalues $0, 0.040$. The a-priori perturbation estimate for the simple eigenvalue is again $\frac{y^*\mathcal{E}x}{y^*x} \approx -0.020202$, while for the multiple unit eigenvalue the theory predicts that $\mathcal{AP}_{\text{inex}}^{-1}$ has a unit eigenvalue with a Jordan block of size 2, and two non-unit eigenvalues, with first-order perturbation term equal to $\xi^{\frac{1}{2}} = 0.10050$ and $\xi^{\frac{1}{2}} = 0$. These expectations are fully met in the numerical experiment, as Table 8.2 shows.

Example 8.8. Magnetostatic problem. We consider a 2088×2088 linear system stemming from the mixed finite element discretization of the 2D magnetostatic problem; we refer to [74] for a detailed description of this test problem. In this example $n = 1272$, $m = 816$, and the resulting matrix $\mathcal{A}$ was properly scaled so that the matrix $(A - I)$ is non-singular and negative definite. We approximated $H = BB^\top$ with $H_{\text{inex}} = R^\top R$, where R is the upper triangular factor of the incomplete Cholesky factorization of $BB^\top$ computed using the MatLab function `cholinc` with different dropping threshold [63]. Table 8.3 shows the most relevant quantities for different values of the tolerance `tol` used in `cholinc`. The table also displays the number of Jordan blocks retained after perturbation (we considered as zero all eigenvalues of E less than 10^{-9} in modulo). This provides a feeling of the spectral quality of the incomplete Cholesky factorization. In the last table row we also report relevant information when using an Algebraic MultiGrid (AMG) preconditioner (in this experiment we used MI20 in [1], with all default parameters). In spite of a larger perturbation (in norm), the spectral properties of H are better reproduced, leading to a less perturbed spectrum of the preconditioned matrix; note that the number of eigenvalues of E less than 10^{-5} in absolute value is even higher, namely 196.

For the problem preconditioned with H_{inex} being the incomplete Cholesky factorization, in Fig. 8.1 we report the true spectrum of $\mathcal{AP}_{\text{inex}}^{-1}$ ($\circ$ symbol), together with its approximation using the first-order expansion in (8.11) and in (8.12) (+ symbol). As expected, we can see that the unit eigenvalues of $\mathcal{AP}_{\text{ex}}^{-1}$ are significantly spread both on the real line and on the complex plane. This behavior is qualitatively fully captured by the first-order perturbation terms (eigenvalues with + symbol) although only higher order perturbation terms would be able to capture the actual direction of the complex eigenvalues. Finally, we observe that the real eigenvalues of $\mathcal{AP}_{\text{inex}}^{-1}$ stemming from simple eigenvalues of $\mathcal{AP}_{\text{ex}}^{-1}$ are fully estimated by the first-order term in (8.12): see the real interval $[0.2, 0.4]$ in the plots.

Fig. 8.2 reveals the special features of the Algebraic Multigrid preconditioner. The unit eigenvalue of $\mathcal{AP}_{\text{ex}}^{-1}$ does not spread on the real axis, as all corresponding eigenvalues of $\mathcal{AP}_{\text{inex}}^{-1}$

(∘ symbol) have non-zero imaginary part. This behavior is fully captured by the first-order estimate, as all values ξ in (8.12) are real and negative, so that $\xi^{\frac{1}{2}}$ is purely imaginary. This phenomenon is in agreement with Proposition 8.5, and relies on well-known spectral equivalence properties of AMG which appear to hold, at least numerically, for this matrix; we refer, e.g., to [77] for a thorough discussion on AMG.

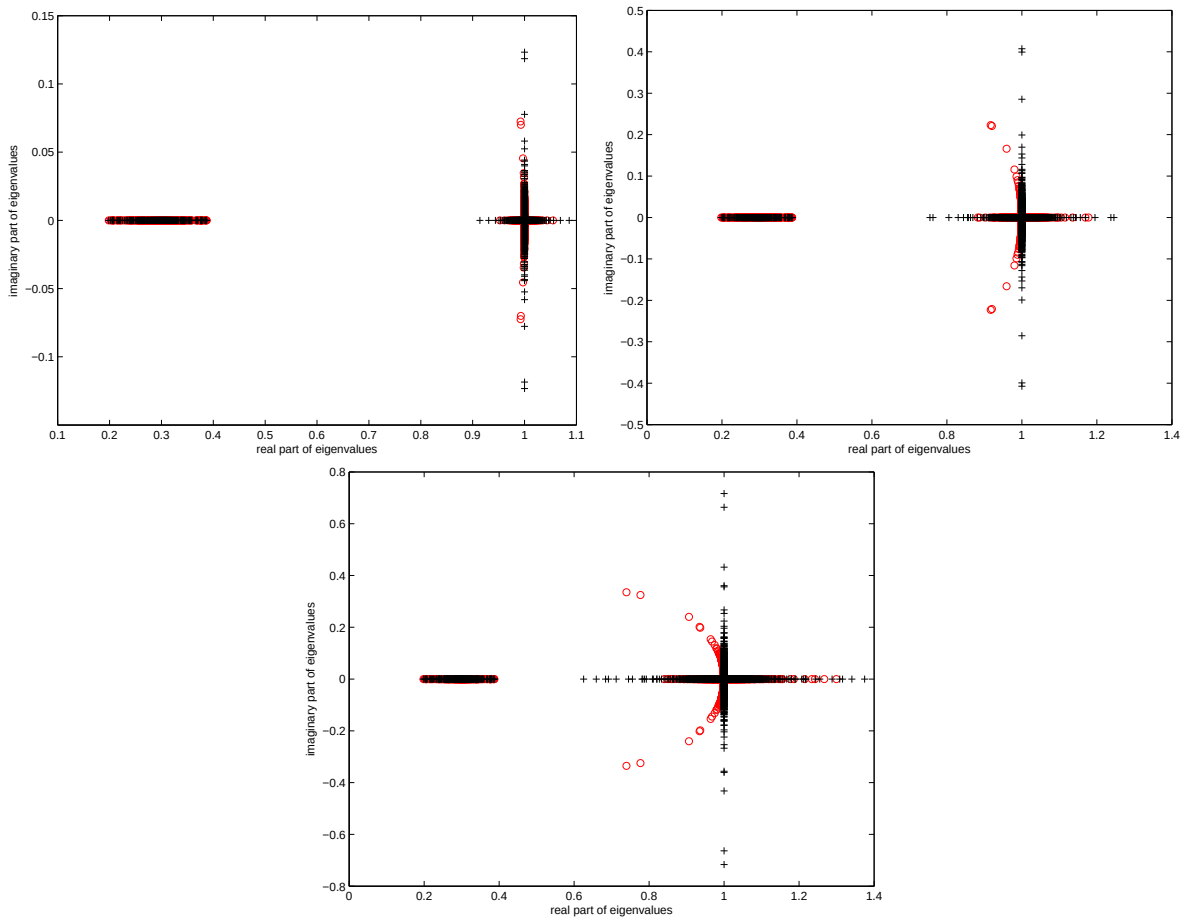

Figure 8.1: Example 8.8. Eigenvalues of $\mathcal{AP}_{\text{inex}}^{-1}$ (∘ symbol) and approximations (+ symbol) as either $\lambda\left(\mathcal{AP}_{\text{ex}}^{-1}\right)+\xi^{\frac{1}{2}}$ or $\lambda\left(\mathcal{AP}_{\text{ex}}^{-1}\right)+\frac{y^*\mathcal{E}x}{y^*x}$, as the `cholinc` tolerance `tol` varies (see (8.11), (8.12)). From the top left corner: `tol` $= 10^{-4}, 5\cdot 10^{-4}, 10^{-3}$.

8.4 The case of non-zero (2,2) block

In this section we generalize our analysis to the case of non-zero (2,2) block in the matrix $\mathcal{A}$. This setting often corresponds to the case when the (1,2) block is column rank deficient, so that the original $\mathcal{A}$ would be singular. We thus assume that $0<\text{rank}\left(B^\top\right)\leq m$; in particular, we exclude the case of $B^\top$ identically zero. Let us thus assume that C is symmetric and positive semidefinite, and such that

$$H := BB^\top + C, \qquad \text{is non-singular.}$$

We thus define

$$\mathcal{P}_{\text{ex}} = \begin{bmatrix} I_n & B^\top \\ B & -C \end{bmatrix}.$$

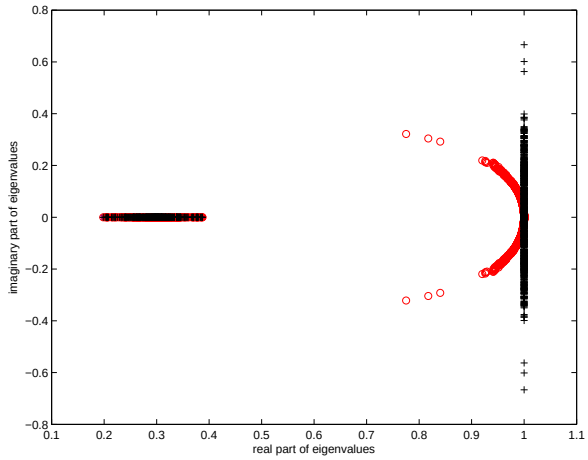

Figure 8.2: Example 8.8. Eigenvalues of $\mathcal{AP}_{\text{inex}}^{-1}$ ($\circ$ symbol) and approximations ($+$ symbol) as either $\lambda\left(\mathcal{AP}_{\text{ex}}^{-1}\right) + \xi^{\frac{1}{2}}$ or $\lambda\left(\mathcal{AP}_{\text{ex}}^{-1}\right) + \frac{y^*\mathcal{E}x}{y^*x}$, for H_{inex} obtained as an Algebraic Multigrid operator.

Using the same steps as in Section 8.1 we can show that the problem

$$\begin{bmatrix} A & B^\top \\ B & -C \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \lambda \mathcal{P}_{\text{ex}} \begin{bmatrix} x \\ y \end{bmatrix},$$

can be transformed into the equivalent eigenproblem

$$\mathcal{M}_0 w = (\lambda - 1)\, w \;\Leftrightarrow\; \left(\begin{bmatrix} I_n \\ \widehat{B} \end{bmatrix} (A - I) \begin{bmatrix} I_n, & -\widehat{B}^\top \end{bmatrix} \right) \begin{bmatrix} u \\ \widehat{v} \end{bmatrix} = (\lambda - 1) \begin{bmatrix} u \\ \widehat{v} \end{bmatrix},$$

where $\widehat{B} = H^{-\frac{1}{2}} B$, and $u, \widehat{v}$ are related to x and y in the same manner. The only difference lies in the definition of H. Clearly, $\widehat{B}$ no longer has all orthonormal rows. The following result generalizes Proposition 8.1 to the new setting. In the following we shall assume that $(A - I)\left(I - \widehat{B}^\top \widehat{B}\right)$ is diagonalizable. This is not a restrictive condition, since A can always be scaled to ensure that $(A - I)$ be definite, from which the diagonalizability follows.

Proposition 8.9. *Let $C \geq 0$ and $H = BB^\top + C$ be non-singular. Let $[Y_*, Y_1, Y_0]$ be the eigenvector basis of $\widehat{B}\widehat{B}^\top$, whose blocks have dimension $\ell_*, \ell_1, \ell_0 \geq 0$, respectively, where Y_1 and Y_0 correspond to the unit and zero eigenvalues, if any. Let $(A - I)\left(I - \widehat{B}^\top \widehat{B}\right) \widehat{X} = \widehat{X}\Theta$ be the partial eigenvalue decomposition of $(A - I)\left(I - \widehat{B}^\top \widehat{B}\right)$ associated with its $n - \ell_1$ non-zero eigenvalues. Then the Weyr decomposition of $\mathcal{M}_0$ is given by*

$$\mathcal{M}_0 \mathcal{X}_0 = \mathcal{X}_0 \begin{bmatrix} \Theta & & & & \\ & 0_{\ell_*} & & & \\ & & 0_{\ell_1} & I_{\ell_1} & \\ & & & 0_{\ell_1} & \\ & & & & 0_{\ell_0} \end{bmatrix},$$

$$\mathcal{X}_0 = \left[\begin{array}{c|c|c|c|c} \widehat{X} & \widehat{B}^\top Y_* & \widehat{B}^\top Y_1 & (A - I)^{-1}\, \widehat{B}^\top Y_1 & 0 \\ \widehat{B}\widehat{X} & Y_* & Y_1 & 0 & Y_0 \end{array}\right],$$

with $\mathcal{X}_0$ non-singular.

Proof. The identity can be easily verified. A constructive proof closely follows the technique used in the proof of Proposition 8.1. The only difference is in the distinction of the Jordan blocks: since $I - \widehat{B}^\top \widehat{B}$ may not be an orthonormal projector, not all eigenvectors of $\widehat{B}^\top \widehat{B}$ contribute into the generation of Jordan blocks. □

We can thus generalize Theorem 8.2 to the case $C \neq 0$. The proof is completely analogous and is thus omitted.

Theorem 8.10. *With the notation and assumptions of Proposition 8.9 and $C \geq 0$, the preconditioned matrix $\mathcal{A}\mathcal{P}_{\text{ex}}^{-1}$ admits the following Weyr decomposition:*

$$\mathcal{A}\mathcal{P}_{\text{ex}}^{-1}\mathcal{X} = \mathcal{X} \begin{bmatrix} I_{n-\ell_1} + \Theta & & & & \\ & I_{\ell_*} & & & \\ & & I_{\ell_1} & I_{\ell_1} & \\ & & & I_{\ell_1} & \\ & & & & I_{\ell_0} \end{bmatrix},$$

with the non-singular matrix $\mathcal{X}$:

$$\begin{aligned} \mathcal{X} &= \begin{bmatrix} I_n & 0 \\ B & -H^{\frac{1}{2}} \end{bmatrix} \mathcal{X}_0 \\ &= \left[\begin{array}{c|c|c|c|c} \widehat{X} & \widehat{B}^\top Y_* & \widehat{B}^\top Y_1 & (A-I)^{-1}\widehat{B}^\top Y_1 & 0 \\ 0 & -CH^{-\frac{1}{2}}Y_* & -CH^{-\frac{1}{2}}Y_1 & B(A-I)^{-1}\widehat{B}^\top Y_1 & -H^{\frac{1}{2}}Y_0 \end{array}\right]. \end{aligned}$$

The decomposition in Theorem 8.10 shows that the number of Jordan blocks may be significantly low, and in particular less than m, if the matrix $\widehat{B}\widehat{B}^\top$ does not have unit eigenvalues, that is, if $BB^\top$ and C do not complement each other (if they do, then $\widehat{B}^\top\widehat{B}$ is an orthogonal projection onto the range of $\widehat{B}^\top$).

To analyze the perturbation induced by an inaccurate computation of $H = BB^\top + C$ by means of some H_{inex}, we can define $E = H - H_{\text{inex}}$ and then write

$$\mathcal{P}_{\text{inex}} = \begin{bmatrix} I_n & B^\top \\ B & -C+E \end{bmatrix} = P_{\text{ex}} + \begin{bmatrix} 0 \\ E \end{bmatrix} [0, I_m].$$

Therefore, precisely the same expression as the one in Theorem 8.3 holds for $C \geq 0$:

$$\mathcal{A}\mathcal{P}_{\text{inex}}^{-1} = \mathcal{A}\mathcal{P}_{\text{ex}}^{-1} + \mathcal{E}, \qquad \text{with } \mathcal{E} = -\mathcal{A}\begin{bmatrix} B^\top \\ -I_m \end{bmatrix} H^{-1}EH_{\text{inex}}^{-1}[B, -I_m].$$

As already said, the Weyr decomposition of Theorem 8.10 reveals that unless $BB^\top$ and C complement each other, we expect fewer than m Jordan blocks to arise in general - these affect the third and forth blocks in $\mathcal{X}$ in Theorem 8.10. In terms of spectral perturbation, fewer eigenvalues will be highly perturbed when using $\mathcal{A}\mathcal{P}_{\text{inex}}^{-1}$ in place of $\mathcal{A}\mathcal{P}_{\text{ex}}^{-1}$. However, there are applications where $BB^\top$ and C do complement each other: in this case $I - B^\top H^{-1}B$ is a projector, and $\ell_1 = m - \ell_0$ Jordan blocks in the exactly preconditioned matrix can be found. For these, the discussion of the previous section on their perturbation applies.

The number of unit eigenvalues that are left unaltered by $\mathcal{A}\mathcal{P}_{\text{inex}}^{-1}$ depends once again on the number of zero eigenvalues of $E = H - H_{\text{inex}}$. A result analogous to Theorem 8.4 is more difficult to obtain, since we would need to distinguish among the occurrence of unit, zero and other eigenvalues of $\widehat{B}\widehat{B}^\top$. As an example, we can easily see that if

$$\text{Null}(E) \cap \text{Range}\left(H_{\text{inex}}^{-1}[B, -I_m]\begin{bmatrix} 0 \\ -H^{\frac{1}{2}}Y_0 \end{bmatrix}\right) \neq \{0\}, \tag{8.13}$$

and k is the dimension of the intersection space, then there will be k unaltered unit eigenvalues in $\mathcal{A}\mathcal{P}_{\text{inex}}^{-1}$ with geometric multiplicity k. Analogously, if

$$\text{Null}(E) \cap \text{Range}\left(H_{\text{inex}}^{-1}[B, -I_m]\begin{bmatrix} \widehat{B}^\top Y_1 \\ -CH^{-\frac{1}{2}}Y_1 \end{bmatrix}\right) \neq \{0\},$$

and k is the dimension of the intersection space, then there will be $2k$ unaltered unit eigenvalues in $\mathcal{AP}_{\text{inex}}^{-1}$ with geometric multiplicity k (note that the forth block of $\mathcal{X}$ is always in the null space of $\mathcal{E}$, hence that block provides the set of generalized eigenvectors). Following the same proving strategy, more extreme cases can be obtained as follows. Recall the definition of ℓ_*, ℓ_1 and ℓ_0 above. Assume $E = H - H_{\text{inex}}$ has k zero eigenvalues. Then

i) If $\ell_* = m$ (i.e. $\ell_0 = \ell_1 = 0$), then k eigenvalues of $\mathcal{AP}_{\text{ex}}^{-1}$ remain unchanged, with geometric multiplicity k;

ii) If $\ell_1 = m$ (i.e. $\ell_* = \ell_0 = 0$), then $2k$ eigenvalues of $\mathcal{AP}_{\text{ex}}^{-1}$ remain unchanged, with geometric multiplicity k.

We conclude with an example that validates the theory described in this section.

Example 8.11. We consider a variant of Example 8.6:

$$A = \begin{bmatrix} 2 & 1 & 0 & 0 \\ 1 & 3 & 1 & 0 \\ 0 & 1 & 4 & 1 \\ 0 & 0 & 1 & 5 \end{bmatrix}, B = \begin{bmatrix} 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 3 \\ 0 & 0 & 0 & 0 \end{bmatrix}, C = \begin{bmatrix} \gamma_1 & 0 & 0 \\ 0 & \gamma_2 & 0 \\ 0 & 0 & 3 \end{bmatrix}.$$

Here B is clearly rank deficient, and depending on the values (zero vs. non-zero) of γ_1, γ_2 we can obtain $\ell_1 = 0, 1, 2$; in particular, notice that for $\gamma_1 = \gamma_2 = 0$ the two matrices $BB^\top$ and C complement each other. At first, we consider

$$H_{\text{inex}} = B \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0.985 & 0 & 0 \\ 0 & 0 & 0.99 & 0 \\ 0 & 0 & 0 & 0.995 \end{bmatrix} B^\top + C.$$

For this choice of H_{inex}, the error matrix E is singular (one zero eigenvalue), irrespective of the choice of C. Table 8.5 summarizes information on the exact and perturbed eigenvalues. For $\gamma_1 = 1$ and $\gamma_2 = 2$ no Jordan blocks are expected even in the exactly preconditioned problem, since $I - \widehat{B}^\top \widehat{B}$ is not a projector, and it is non-singular. Other relevant quantities are collected in Table 8.4 for these and other values in C. First order terms yield the following linear perturbation of the eigenvalues of $\mathcal{AP}_{\text{ex}}^{-1}$, which matches pretty well the actual perturbation taking place:

$\lambda\left(\mathcal{AP}_{\text{ex}}^{-1}\right)$	$y^*\mathcal{E}x$
3.6888	-0.003994
1.8751	-0.040204
1.5754	-0.023661
1.1880	-0.015681

It can also be seen that (8.13) holds with $k = 1$, therefore one unaltered unit eigenvalue can also be observed.

We next chose $\gamma_1 = 0$ and $\gamma_2 = 2$, so that C is singular. With these choices, $\ell_1 = 1$, so that a Jordan block in the canonical form of $\mathcal{AP}_{\text{ex}}^{-1}$ occurs. The theory predicts that the corresponding eigenvalue is perturbed by at least $\xi^{\frac{1}{2}} = 0.11208$ (plus the $o\left(\|\mathcal{E}\|^{\frac{1}{2}}\right)$ terms), and this can be observed in practice (see the second block of rows in Table 8.5). Since (8.13) still holds with $k = 1$, one unaltered unit eigenvalue can also be observed. All remaining eigenvalues show an at most linear perturbation in $\|\mathcal{E}\|$.

Finally, we consider the case $\gamma_1 = \gamma_2 = 0$, so that $BB^\top$ and C complement each other. In this case, $\ell_1 = 2$, so that $B^\top H^{-1} B$ is a projector onto $\text{Range}\left(B^\top\right)$. The theory ensures that the eigenvalues of the 2 Jordan blocks are thus perturbed by at least $\xi^{\frac{1}{2}} = 0.13020\text{i}, 0.10944$, and this can be verified in Table 8.5. Once again, (8.13) holds with $k = 1$, therefore one unaltered unit eigenvalue can also be observed. All other eigenvalues are perturbed at most linearly.

γ_1, γ_2	$\|E\|$	$\|E\| \,/\, \|H\|$	$\|\mathcal{E}\|$
1, 2	0.045	0.00409	0.030879
0, 2	0.045	0.00409	0.044302
0, 0	0.045	0.00500	0.045235

Table 8.4: Example 8.11. Norms of error and perturbation matrices, for various values of the diagonal elements in C.

γ_1, γ_2	$\lambda\left(\mathcal{AP}_{\text{ex}}^{-1}\right)$	$\lambda\left(\mathcal{AP}_{\text{inex}}^{-1}\right)$	$\left\lvert\lambda\left(\mathcal{AP}_{\text{ex}}^{-1}\right) - \lambda\left(\mathcal{AP}_{\text{inex}}^{-1}\right)\right\rvert$
1, 2	1	1.0000	$2.220 \cdot 10^{-16}$
	1.0000	1.0238	0.023763
	1.0000	1.0478	0.047758
	1.1880	1.1657	0.022336
	1.5754	1.5511	0.024283
	1.8751	1.8335	0.041568
	3.6888	3.6848	0.003960
0, 2	1.0000	1.0000	$3.33 \cdot 10^{-16}$
	1.0000	1.0070 - 0.13768i	0.13785
	1.0000	1.0070 + 0.13768i	0.13785
	1.0000	1.0238	0.02377
	1.3820	1.3573	0.02468
	1.7273	1.6885	0.03878
	3.6180	3.6142	0.00381
0, 0	1.0000	1.0000	$2.220 \cdot 10^{-16}$
	1.0000	1.0015 - 0.11657i	0.11657
	1.0000	1.0015 + 0.11657i	0.11657
	1.0000	0.9946 - 0.16696i	0.16705
	1.0000	0.9946 + 0.16696i	0.16705
	1.3820	1.3584	0.02356
	3.6180	3.6142	0.00382

Table 8.5: Example 8.11. Eigenvalues of the exactly and inexactly preconditioned problem, and their difference. The first column shows the choice of the parameters in C.

Conclusions

In the first part of this thesis have provided new tools for working with sequences of matrices: we have extended a recent perturbation result based on a theorem by Mirsky: more in detail our findings concern the eigenvalue distribution and localization of a generic (non-Hermitian) complex perturbation of a bounded Hermitian sequence of matrices; we have studied the stability of the *a.c.s.* notion for sequences formed by Hermitian matrices, under the action of a continuous function defined on the real line; we observed that tools from matrix theory (Mirsky Theorem, see [15]) and approximation theory in the complex field (Mergelyan's Theorem, see [76]), combined with those from asymptotic linear algebra [84, 109, 112], have been crucial in our proof of results concerning the eigenvalue distribution of non Hermitian matrix sequences. In particular, we have exploited these tools to deduce general results that we have applied, as a special case, to the algebra generated by Toeplitz sequences (an interesting side effect, already implicitly contained in the Tilli analysis [112], is a characterization of the range of $L^\infty(\mathbb{T}^d)$ functions obtained as restrictions of functions of several complex variables in the Hardy space H^∞); finally we have studied in detail the singular values and eigenvalues of g-circulant matrices and we have identified the joint asymptotic distribution of g-Toeplitz sequences associated with a given integrable symbol with the generalization to the multilevel block setting.

Some problems remain open.

For example, in Section 3.1 we studied the approximation of PDEs with given boundary conditions. As an important example of application, it would be interesting to extend the CG convergence analysis of Beckermann and Kuijlaars [8] to this quasi-Hermitian setting, in the case where the Hermitian part is positive definite for every n and the global distribution function of the eigenvalues is positive and bounded. The key point would be the definition of proper assumptions on the outliers and on the extreme eigenvalues in order to mimic, if possible, the same analysis performed in [8].

In Section 3.2, with the given assumption of Hermitian character, the results presented generalize those of [83, 94] and have application in determining, if any, the eigenvalues limit distribution of general matrix sequences (refer to [111, Proposition 2.7] and its generalization, i.e., [83, Proposition 2.3]). Given the applications of such a study to the convergence analysis of Krylov-type methods, we believe that the considered field and related tools are worth to be further investigated. Some issues, not in a special order, are the following:

- Extending Theorems 3.2 and 3.3 to the case where A_n is not necessarily Hermitian and $\{A_n - A_n^*\}_n$ is distributed as the zero function. Some precise trace-norm conditions seem to be necessary as emphasized in [45, 52];
- Extending Theorems 3.2 and 3.3 to the case of singular values, when the Hermitianity assumption is dropped;
- Setting and answering to the question whether $\{f(A_n)\}_n$ is a GLT sequence if $\{A_n\}_n$ is;
- Determining specific and interesting examples in which such a theory represents a concrete improvement for the convergence analysis of Krylov-type methods, for stability issues of approximated PDEs etc., in the sense indicated in [90, Section 3];

- Considering the hint given by the knowledge of the spectral symbol in the GLT case for designing new proposals of CG/GMRes preconditioners and prolongation/restriction operators for multigrid procedures.

With regard to the product of sequences of Toeplitz matrices, it would be interesting to extend the results of Chapter 4 to the case where the involved symbols are not necessarily bounded, but just integrable. As already stressed in [90], in that case, the matrix theoretic approach seems more convenient, since the corresponding Toeplitz operators are not well defined if the symbols are not bounded. It should be observed that the conditions described in the Tilli class for the existence of a canonical distribution corresponding to the symbol are sufficient, but not necessary. In fact for $f(t) = e^{-\mathrm{i}t}$ the range of f is the complex unit circle, disconnecting the complex plane, while the eigenvalues are all equal to zero. However, if one takes the symbol $f(t)$ in [20, (3.24), p. 80] ($f(t) = e^{2\mathrm{i}t}$, $t \in [0, \pi)$, $f(t) = e^{-2\mathrm{i}t}$, $t \in [\pi, 2\pi)$), then the range of f is again the complex unit circle, that disconnects the complex plane, but the eigenvalues indeed distribute as the symbol as discussed in [20, Example 5.39, pp. 167-169]. A further step would consist in understanding how to discriminate between these two types of generating functions which do not belong to the Tilli class.

We also would like to study the more involved eigenvalue/eigenvector behavior both for g-circulant and g-Toeplitz structures.

When in a given field of science (mathematics, physics, chemistry, etc.) new discoveries are made, an important next step is to find an application. In the second part of the thesis we dealt with this, especially we have studied in detail the singular values of matrix sequences obtained by preconditioning g-Toeplitz sequences associated with a given integrable symbol via g-circulant sequences: the main point is that the standard preconditioning works only in the classical setting, namely when $g_i = \pm 1$, $i = 1, \ldots, d$. However, when g (or $|g|$) is positive a basic preconditioner for regularizing techniques can be obtained by a clever choice of the g-circulant sequence $\{C_{n,g}\}$. We have presented and discussed several numerical results, also instructive for specific applications in image deblurring and denoising. In particular they have confirmed that the proposed preconditioners can be used as a basic tool for obtaining regularizing features, by means of filtering techniques which will be analyzed and discussed in next works.

We have extended the rigorous two-grid analysis for circulant matrices to the case where the size reduction is performed by a factor g with $g > 2$. An interesting feature of a size reduction by a factor $g > 2$ is that it allows to eliminate some pathologies which occur when $g = 2$. In particular, if the considered matrices come from the approximation of certain integro-differential equations, then we have two sources of ill-conditioning and the zeros of the underlying symbol are located at $x = 0$ and at $x = \pi$: this situation is a special case of mirror point zeros, and when $g = 2$, it is possible to prove that the resulting two-grid iteration cannot be optimal (see [42, 85]). Such difficulty can be overcome using a block projector [25, 55] or choosing a larger g. Moreover, when increasing g, the size of the coarse problems decreases: as a consequence more multigrid recursive calls could be considered, like the W-cycle which is proved to be optimal for $g \geq 3$. We stress that the numerical experiments are encouraging, not only for circulant matrices but also regarding Toeplitz matrices and concerning the use of the V-cycle algorithm. A future line of research must include the multilevel setting, following the approach in [85, 2], and a rigorous proof of convergence for the whole V-cycle procedure in accordance with the proof technique introduced in [3]. The proposed convergence analysis could be then applied to smoothing aggregation projection techniques [67].

We have analyzed in detail the spectral perturbation induced by the use of the inexact constraint preconditioner when solving large scale symmetric algebraic saddle point problems with zero and non-zero (2,2) block. Our results emphasize the role of the spectral properties of the approximation to the core matrix $H = BB^T$ to be able to predict the actual distribution of the spectrum of $\mathcal{A}\mathcal{P}_{\text{inex}}^{-1}$. Moreover, thanks to an explicit description of the transformation

matrix in the canonical form, we were able to fully track the linear and non-linear perturbation of the eigenvalues. Our numerical results show that the analysis can be accurate also on data stemming from real applications.

Bibliography

[1] AEA Technology, *Harwell subroutine library*, Harwell Laboratory, Oxfordshire, England, 1995.

[2] A. Aricò and M. Donatelli, *A V-cycle multigrid for multilevel matrix algebras: proof of optimality*, Numer. Math., 105-4 (2007), pp. 511–547.

[3] A. Aricò, M. Donatelli and S. Serra-Capizzano, *V-cycle optimal convergence for certain (multilevel) structured linear systems*, SIAM J. Matrix Anal. Appl., 26-1 (2004), pp. 186–214.

[4] F. Avram, *On bilinear forms on Gaussian random variables and Toeplitz matrices*, Probab. Theory Related Fields, 79 (1988), pp. 37–45.

[5] O. Axelsson and V. A. Barker, *Finite Element Solution of Boundary Value Problems: Theory and Computation*, Academic Press Inc., New York, 1984.

[6] O. Axelsson and G. Lindskög, *On the rate of convergence of the preconditioned conjugate gradient method*, Numer. Math., 52 (1986), pp. 499–523.

[7] Z.-Z. Bai, M. K. Ng and Z.-Q. Wang, *Constraint preconditioners for symmetric indefinite matrices*, SIAM J. Matrix Anal. Appl., 31 (2009), pp. 410–433.

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

[9] B. Beckermann and S. Serra-Capizzano, *On the asymptotic spectrum of Finite Elements matrices*, SIAM J. Numer. Anal., 45-2 (2007), pp. 746–769.

[10] M. Benzi, G. H. Golub and J. Liesen, *Numerical solution of saddle point problems*, Acta Numer., 14 (2005), pp. 1–137.

[11] M. Benzi and V. Simoncini, *On the eigenvalues of a class of saddle point matrices*, Numer. Math., 103 (2006), pp. 173–196.

[12] B. Bercu and W. Bryc, *Asymptotic results for empirical measures of weighted sums of independent random variables*, Electron. Comm. Probab., 12 (2007), pp. 184–199.

[13] B. Bercu, F. Gamboa and M. Lavielle, *Sharp large deviations for Gaussian quadratic forms with applications*, ESAIM: Probab. Statist., 4 (2000), pp. 1–24.

[14] M. Bertero and P. Boccacci, *Introduction to Inverse Problems in Imaging*, Institute of Physics Publ., Bristol, 1998.

[15] R. Bhatia, *Matrix Analysis*, Springer Verlag, New York, 1997.

[16] R. Bhatia, *Fourier Series*, AMS, Providence, 2005.

[17] A. Böttcher, *Variable-coefficient Toeplitz matrices with symbols beyond the Wiener algebra*, Oper. Theory Adv. Appl., 199 (2010), pp. 192–202.

[18] A. Böttcher and S. Grudsky, *Spectral properties of banded Toeplitz matrices*, SIAM, Philadelphia, PA, 2005.

[19] A. Böttcher and S. Grudsky, *Uniform boundedness of Toeplitz matrices with variable coefficients*, Integr. Equat. Oper. Th., 60-3 (2008), pp. 313–328.

[20] A. Böttcher and B. Silbermann, *Introduction to Large Truncated Toeplitz Matrices*, Springer Verlag, New York, 1999.

[21] A. Böttcher, J. Gutiérrez-Gutiérrez and P. M. Crespo, *Mass concentration in quasicommutators of Toeplitz matrices*, J. Comput. Appl. Math., 205 (2007), pp. 129–148.

[22] A. Böttcher, S. M. Grudsky and E. A. Maksimenko, *The Szegö and Avram-Parter theorems for general test functions*, C. R. Math. Acad. Sci. Paris, 346 (2008), pp. 749–752.

[23] A. Brown and P. R. Halmos, *Algebraic properties of Toeplitz operators*, J. Reine Angew. Math., 213 (1964), pp. 89–102.

[24] R. H. Chan and M. K. Ng, *Conjugate gradient methods for Toeplitz systems*, SIAM Rev., 38-3 (1996), pp. 427–482.

[25] R. H. Chan, Q.-S. Chang and H.-W. Sun, *Multigrid method for ill-conditioned symmetric Toeplitz systems*, SIAM J. Sci. Comput., 19 (1998), pp. 516–529.

[26] T. F. Chan, *An optimal circulant preconditioner for Toeplitz systems*, SIAM J. Sci. Stat. Comp., 9 (1988), pp. 766–771.

[27] Q.-S. Chang, X.-Q. Jin and H.-W. Sun, *Convergence of the multigrid method for ill-conditioned block toeplitz systems*, BIT, 41-1 (2001), pp. 179–190.

[28] G. Codevico, G. Heinig and M. Van Barel, *A superfast solver for real symmetric Toeplitz systems using real trigonometric transformations*, Numer. Linear Algebra Appl., 12 (2005), pp. 699–713.

[29] I. Daubechies, *Ten Lectures on Wavelets*, CBMS-NSF Regional Conf. Ser. Appl. Math., 61, SIAM, Philadelphia, 1992.

[30] P. J. Davis, *Circulant Matrices*, John Wiley and Sons, New York, 1979.

[31] H. S. Dollar, *Constraint-style preconditioners for regularized saddle point problems*, SIAM J. Matrix Anal. Appl., 29 (2007), pp. 672–684.

[32] H. S. Dollar, N. I. M. Gould, W. H. A. Schilders and A. J. Wathen, *Implicit-factorization preconditioning and iterative solvers for regularized saddle-point systems*, SIAM J. Matrix Anal. Appl., 28 (2006), pp. 170–189.

[33] M. Donatelli, *An algebraic generalization of local Fourier analysis for grid transfer operators in multigrid based on Toeplitz matrices*, Numer. Linear Algebra Appl., 17-2/3 (2010), pp. 179–197.

[34] M. Donatelli, S. Serra-Capizzano and D. Sesana, *Multigrid methods for Toeplitz linear systems with different size reduction*, BIT, to appear.

[35] C. Durazzi and V. Ruggiero, *Indefinitely preconditioned conjugate gradient method for large sparse equality and inequality constrained quadratic problems*, Numer. Linear Algebra Appl., 10-8 (2003), pp. 673–688.

[36] N. Dyn and D. Levin, *Subdivision schemes in geometric modelling*, Acta Numer., 11 (2002), pp. 73–144.

[37] C. Estatico, *A classification scheme for regularizing preconditioners, with application to Toeplitz systems*, Linear Algebra Appl., 397 (2005), pp. 107–131.

[38] C. Estatico, *Preconditioners for ill-conditioned Toeplitz matrices with differentiable generating functions*, Numer. Linear Algebra Appl., 16-3 (2009), pp. 237–257.

[39] C. Estatico, E. Ngondiep, S. Serra-Capizzano and D. Sesana, *A note on the (regularizing) preconditioning of g-Toeplitz sequences via g-circulants*, J. Comput. Appl. Math., submitted.

[40] R. E. Ewing, R. D. Lazarov, P. Lu and P. S. Vassilevski, *Preconditioning indefinite systems arising from mixed finite element discretization of second-order elliptic problems*, in Notes in Mathematics, Springer, 1990, pp. 28–43.

[41] D. Fasino and P. Tilli, *Spectral clustering properties of block multilevel Hankel matrices*, Linear Algebra Appl., 306-1/3 (2000), pp. 155–163.

[42] G. Fiorentino and S. Serra-Capizzano, *Multigrid methods for Toeplitz matrices*, Calcolo, 28-3/4 (1991), pp. 283–305.

[43] G. Fiorentino and S. Serra-Capizzano, *Multigrid methods for symmetric positive definite block Toeplitz matrices with nonnegative generating functions*, SIAM J. Sci. Comput., 17-5 (1996), pp. 1068–1081.

[44] H. Gazzah, P. Regalia and J.P. Delmas, *Asymptotic eigenvalue distribution of block Toeplitz matrices and application to blind SIMO channel identification*, IEEE Trans. Inform. Theory, 47-3 (2001), pp. 1243–1251.

[45] L. Golinskii and S. Serra-Capizzano, *The asymptotic properties of the spectrum of nonsymmetrically perturbed Jacobi matrix sequences*, J. Approx. Theory, 144 (2007), pp. 84–102.

[46] G. H. Golub and C. F. Van Loan, *Matrix Computations*, The Johns Hopkins University Press, Baltimore, 1983.

[47] A. Greenbaum, *Iterative Methods for Solving Linear Systems*, SIAM publ., Philadelphia, 1997.

[48] U. Grenander and G. Szegö, *Toeplitz Forms and Their Applications*, Second Edition, Chelsea, New York, 1984.

[49] J. Gutiérrez-Gutiérrez, P. Crespo and A. Böttcher, *Functions of the banded Hermitian block Toeplitz matrices in signal processing*, Linear Algebra Appl., 422-2/3 (2007), pp. 788–807.

[50] W. Hackbush, *Multigrid Methods and Applications*, Springer Verlag, Berlin, 1985.

[51] J. C. Haws and C. D. Meyer, *Preconditioning KKT systems*, Linear Algebra Appl., 1-6 (2001), pp.

[52] S. Holmgren, S. Serra-Capizzano and P. Sundqvist, *Can one hear the composition of a drum?*, Mediterr. J. Math., 3-2 (2006), pp. 227–249.

[53] T. K. Huckle, *Compact Fourier analysis for designing multigrid methods*, SIAM J. Sci. Comput., 31-1 (2008), pp. 644–666.

[54] T. Huckle and J. Staudacher, *Multigrid preconditioning and Toeplitz matrices*, Electr. Trans. Numer. Anal., 13 (2002), pp. 81–105.

[55] T. Huckle and J. Staudacher, *Multigrid methods for block toeplitz matrices with small blocks*, BIT., 46-1 (2006), pp. 61–83.

[56] M. Kac, W. L. Murdoch and G. Szegö, *On the eigenvalues of certain Hermitian forms*, J. Rational Mech. Anal., 2 (1953), pp. 767–800.

[57] N. Kalouptsidis, G. Carayannis and D. Manolakis, *Fast algorithms for block Toeplitz matrices with Toeplitz entries*, Signal Process., 6-1 (1984), pp. 77–81.

[58] C. Keller, N. I. M. Gould and A. J. Wathen, *Constraint preconditioning for indefinite linear systems*, SIAM J. Matrix Anal. Appl., 21 (2000), pp. 1300–1317.

[59] A. B. J. Kuijlaars, *Convergence analysis of Krylov subspace iterations with methods from potential theory*, SIAM Rev., 48-1 (2006), pp. 3–40.

[60] A. B. J. Kuijlaars and S. Serra-Capizzano, *Asymptotic zero distribution of orthogonal polynomials with discontinuously varying recurrence coefficients*, J. Approx. Theory, 113-1 (2001), pp. 142–155.

[61] V. B. Lidskii, *Perturbation theory of non-conjugate operators*, U.S.S.R. Comput. Math. and Math. Phys., 1 (1965), pp. 73–85. Also as Zh. vychisl. Mat. mat. Fiz., 6 (1965), pp. 52–60.

[62] I. Louhichi, E. Strouse and L. Zakariasy, *Products of Toeplitz Operators on the Bergman space*, Integr. Equat. Oper. Th., 54 (2006), pp. 525–539.

[63] The MathWorks, Inc., *MATLAB 7*, September 2004.

[64] E. H. Moore, *General Analysis. Part I*, Amer. Phil. Society, Philadelphia, 1935.

[65] J. Moro, J. V. Burke and M. L. Overton, *On the Lidskii-Vishik-Lyusternik perturbation theory for eigenvalues of matrices with arbitrary Jordan structure*, SIAM J. Matrix Anal. Appl., 18 (1997), pp. 793–817.

[66] A. Napov and Y. Notay, *Comparison of bounds for V-cycle multigrid*, Appl. Numer. Math., 60-3 (2010), pp. 176–192.

[67] A. Napov and Y. Notay, *Algebraic analysis of aggregation-based multigrid*, Numer. Linear Algebra Appl., to appear (2011).

[68] E. Ngondiep, S. Serra-Capizzano and D. Sesana, *Spectral features and asymptotic properties for α-circulants and α-Toeplitz sequences: theoretical results and examples*, preprint available on line from http://arxiv.org/abs/0906.2104 (2009).

[69] E. Ngondiep, S. Serra-Capizzano and D. Sesana, *Spectral features and asymptotic properties for g-circulants and g-Toeplitz sequences*, SIAM J. Matrix Anal. Appl., 31-4 (2010), pp. 1663–1687.

[70] D. Noutsos, S. Serra-Capizzano and P. Vassalos, *Matrix algebra preconditioners for multilevel Toeplitz systems do not insure optimal convergence rate*, Theoret. Computer Sci., 315 (2004), pp. 557–579.

[71] S. V. Parter, *On the eigenvalues of certain generalizations of Toeplitz matrices*, Arch. Rat. Mech. An., 11-1 (1962), pp. 244–257.

[72] S. V. Parter, *On the distribution of the singular values of Toeplitz matrices*, Linear Algebra Appl., 80 (1986), pp. 115–130.

[73] R. Penrose, *A generalized inverse for matrices*, Proc. Cambridge Phil. Soc., 51 (1955), pp. 406–413.

[74] I. Perugia and V. Simoncini, *Block-diagonal and indefinite symmetric preconditioners for mixed finite element formulations*, Numer. Linear Algebra Appl., 7-7/8 (2000), pp. 585–616.

[75] M. Rozložník and V. Simoncini, *Krylov subspace methods for saddle point problem with indefinite preconditioning*, SIAM J. Matrix Anal. Appl., 24-2 (2002), pp. 368–391.

[76] W. Rudin, *Real and Complex Analysis*, McGraw-Hill, New York, 1974.

[77] J. W. Ruge and K. Stüben, *Algebraic multigrid*, in Multigrid Methods, S. McCormick, ed., Frontiers Appl. Math. 3, SIAM, Philadelphia, pp. 73–130 (1987).

[78] Y. Saad, *Iterative methods for sparse linear systems*, SIAM, Society for Industrial and Applied Mathematics, 2nd ed., 2003.

[79] S. Serra-Capizzano, *Multi-iterative methods*, Comput. Math. Appl., 26-4 (1993), pp. 65–87.

[80] S. Serra-Capizzano, *A Korovkin - type Theory for finite Toeplitz operators via matrix algebras*, Numer. Math., 82-1 (1999), pp. 117–142.

[81] S. Serra-Capizzano, *A Korovkin based approximation of multilevel Toeplitz matrices (with rectangular unstructured blocks) via multilevel trigonometric matrix spaces*, SIAM J. Numer. Anal., 36-6 (1999), pp. 1831–1857.

[82] S. Serra-Capizzano, *Spectral and computational analysis of block Toeplitz matrices with nonnegative definite generating functions*, BIT, 39-1 (1999), pp. 152–175.

[83] S. Serra-Capizzano, *Distribution results on the algebra generated by Toeplitz sequences: a finite-dimensional approach*, Linear Algebra Appl., 328-1/3 (2001), pp. 121–130.

[84] S. Serra-Capizzano, *Spectral behavior of matrix sequences and discretized boundary value problems*, Linear Algebra Appl., 337-1/3 (2001), pp. 37–78.

[85] S. Serra-Capizzano, *Convergence analysis of two-grid methods for elliptic Toeplitz and PDEs matrix-sequences*, Numer. Math., 92-3 (2002), pp. 433–465.

[86] S. Serra-Capizzano, *Matrix algebra preconditioners for multilevel Toeplitz matrices are not superlinear*, Linear Algebra Appl., 343-344 (2002), pp. 303–319.

[87] S. Serra-Capizzano, *Test functions, growth conditions and Toeplitz matrices*, Rendiconti Circolo Mat. Palermo, II-68 (2002), pp. 791–795.

[88] S. Serra-Capizzano, *A note on antireflective boundary conditions and fast deblurring models*, SIAM J. Sci. Comput., 25-4 (2003), pp. 1307–1325.

[89] S. Serra-Capizzano, *Generalized Locally Toeplitz sequences: spectral analysis and applications to discretized Partial Differential Equations*, Linear Algebra Appl., 366 (2003), pp. 371–402.

[90] S. Serra-Capizzano, *The GLT class as a Generalized Fourier Analysis and applications*, Linear Algebra Appl., 419-1 (2006), pp. 180–233.

[91] S. Serra-Capizzano, *The spectral approximation of multiplication operators via asymptotic (structured) linear algebra*, Linear Algebra Appl., 424-1 (2007), pp. 154–176.

[92] S. Serra-Capizzano and D. Sesana, *Approximating classes of sequences: the Hermitian case*, Linear Algebra Appl., 434 (2011), pp. 1163–1170.

[93] S. Serra-Capizzano and D. Sesana, *Tools for the eigenvalue distribution in a non-Hermitian setting*, Linear Algebra Appl., 430 (2009), pp. 423–437.

[94] S. Serra-Capizzano and P. Sundqvist, *Stability of the notion of approximating class of sequences and applications*, J. Comput. Appl. Math., 219-2 (2008), pp. 518–536.

[95] S. Serra-Capizzano and C. Tablino-Possio, *Multigrid methods for multilevel circulant matrices*, SIAM J. Sci. Comput., 26-1 (2004), pp. 55–85.

[96] S. Serra-Capizzano and P. Tilli, *On unitarily invariant norms of matrix valued linear positive operators*, J. Inequalities Appl., 7-3 (2002), pp. 309–330.

[97] S. Serra-Capizzano and E. Tyrtyshnikov, *Any circulant-like preconditioner for multilevel matrices is not superlinear*, SIAM J. Matrix Anal. Appl., 21-2 (2000), pp. 431–439.

[98] S. Serra-Capizzano, D. Bertaccini and G. H. Golub, *How to deduce a proper eigenvalue cluster from a proper singular value cluster in the non normal case*, SIAM J. Matrix Anal. Appl., 27-1 (2005), pp. 82–86.

[99] S. Serra-Capizzano, D. Sesana and E. Strouse, *The eigenvalue distribution of products of Toeplitz matrices - Clustering and attraction*, Linear Algebra Appl., 432 (2010), pp. 2658–2678.

[100] D. Sesana, *Spectral theory of matrix sequences: from Szegö to Tyrtyshnikov and applications (Teoria spettrale per successioni di matrici: da Szegö a Tyrtyshnikov e applicazioni)*, in Italian, Master Degree Thesis in Mathematics, Univ. Insubria, 2006.

[101] D. Sesana and V. Simoncini, *Spectral analysis of inexact constraint preconditioning for symmetric saddle point matrices*, Linear Algebra Appl., submitted.

[102] H. Shapiro, *The Weyr characteristic*, The American Mathematical Monthly, 106 (1999), pp. 919–929.

[103] B. Silbermann and O. Zabroda, *Asymptotic behavior of generalized convolutions: an algebraic approach*, J. Integral Equ. Appl., 18-2 (2006), pp. 169–196.

[104] V. Simoncini and D. B. Szyld, *Recent computational developments in Krylov subspace methods for linear systems*, Numer. Linear Algebra Appl., 14-1 (2007), pp. 1–59.

[105] G. W. Stewart and J.-G. Sun, *Matrix Perturbation Theory*, Academic Press, 1990.

[106] G. Strang, *A proposal for Toeplitz matrix calculations*, Stud. Appl. Math., 74-2 (1986), pp. 171–176.

[107] G. Strang, *Wavelets and dilation equations: a brief introduction*, SIAM Rev., 31-4 (1989), pp. 614–627.

[108] H.-W. Sun, R. H. Chan and Q.-S. Chang, *A note on the convergence of the two-grid method for Toeplitz systems*, Comput. Math. Appl., 34 (1997), pp. 11–18.

[109] P. Tilli, *A note on the spectral distribution of Toeplitz matrices*, Linear Multilin. Algebra, 45 (1998), pp. 147–159.

[110] P. Tilli, *Singular values and eigenvalues of non-Hermitian block Toeplitz matrices*, Linear Algebra Appl., 272-1/3 (1998), pp. 59–89.

[111] P. Tilli, *Locally Toeplitz sequences: spectral theory and applications*, Linear Algebra Appl., 278-1/3 (1998), pp. 91–120.

[112] P. Tilli, *Some results on complex Toeplitz eigenvalues*, J. Math. Anal. Appl., 239-2 (1999), pp. 390–401.

[113] W. F. Trench, *Properties of multilevel block $\underline{\alpha}$-circulants*, Linear Algebra Appl., 431 (2009), pp. 1833–1847.

[114] U. Trottenberg, C. Oosterlee and A. Schüller, *Multigrid*, Academic Press, San Diego, 2001.

[115] E. E. Tyrtyshnikov, *Optimal and superoptimal circulant preconditioners*, SIAM J. Matrix Anal. Appl., 13-2 (1992), pp. 459–473.

[116] E. E. Tyrtyshnikov, *A unifying approach to some old and new theorems on distribution and clustering*, Linear Algebra Appl., 232 (1996), pp. 1–43.

[117] E. E. Tyrtyshnikov, *Some applications of a matrix criterion for equidistribution*, Mat. Sb., 192-12 (2001), pp. 1877–1887.

[118] E. E. Tyrtyshnikov and N. L. Zamarashkin, *Spectra of multilevel Toeplitz matrices: advanced theory via simple matrix relationships*, Linear Algebra Appl., 270-1/3 (1998), pp. 15–27.

[119] E. E. Tyrtyshnikov and N. L. Zamarashkin, *Thin structure of eigenvalue clusters for non-Hermitian Toeplitz matrices*, Linear Algebra Appl., 292 (1999), pp. 297–310.

[120] H. Widom, *Eigenvalue distribution of nonselfadjoint Toeplitz matrices and the asymptotics of Toeplitz determinants in the case of nonvanishing index*, Oper. Theory Adv. Appl., 48 (1990), pp. 387–421.

[121] J. H. Wilkinson, *The Algebraic Eigenvalue Problem*, Claredon Press, Oxford, 1965.

[122] H. Woedermann, *Estimates of inverses of multivariable Toeplitz matrices*, Operators and Matrices 2 (2008), pp. 507–515.

[123] A. Zygmund, *Trigonometric Series*, Cambridge University Press, Cambridge, 1959.

Acknowledgments

I would like to thank, in primis, my supervisor Stefano Serra-Capizzano for supporting (and "sopporting") me during these three years of Ph.D..

Thanks to Marco Donatelli, Eric Ngondiep, Valeria Simoncini, Claudio Estatico, Elizabeth Strouse and Matteo Semplice for their friendship and collaboration.

I would also like to thank the referee Dario Bini for his remarks, which have improved the quality and the readability of this Ph.D. thesis.

Now I prefer to continue in italian ...

Un grazie a tutte le persone incontrate durante questi anni che in vario modo mi hanno aiutato a raggiungere questo importante traguardo: tutti i personaggi, più o meno curiosi, che occupano, più o meno abusivamente, l'ufficio matematici al *IV* piano di Via Valleggio 11, i miei nuovi amici/colleghi dell'Università del Piemonte Orientale e i miei compagni di dottorato, in particolare Luca e Luisa ... grazie per la vostra bella amicizia!

Un saluto particolare ad Alessandro e Franca e a tutti i miei studenti del Politecnico di Milano che si sono avvicendati in questi anni.

Colgo l'occasione per ringraziare Don Aldo e tutti i componenti della corale "San Galdino - Don Mario Tocchetti" di Sala al Barro per il supporto, se non tecnico almeno morale, di questi anni: in particolare vorrei augurare a Marco un Buon Cammino in Seminario e a Barbara di riuscire finalmente a discutere la sua tesi di Laurea (è una lunga storia ...).

Infine, un caloroso abbraccio va a tutta la mia famiglia: fratelli e sorelle (my twin sister Diana in primis), zii e zie, cognati e cognate, e parenti vari, ai miei stupendi nipoti (Claudia, Marco, Livio, Nicola e ... (chi sarà?)) e in particolare a mamma e papà: un grazie di cuore a tutti!