Università degli Studi dell'Insubria

Approximation and Spectral Analysis For Large Structured Linear Systems

Thesis submitted to the Facoltà di Scienze MM.FF.NN in
partial fulfilment of the requirements for the
Degree of Philosophiæ Doctor (Ph.D)
in Mathematics

by

Eric Ngondiep

Dipartimento di Fisica e Matematica
Facoltà di Scienze MM.FF.NN

March 2011

Università degli Studi dell'Insubria
Facoltà di Scienze MM.FF.NN

This Thesis is entitled:

Approximation and Spectral Analysis For Large Structured Linear Systems

Presented by:

Eric Ngondiep

has been evaluated by the committee established by the following persons:

Reporter
Thomas Huckle, Prof. Doctor, Technische Universität München, Germany

External Examiner
Dimitrios Noutsos, Professor, University of Ioannina, Greece

Examiner
Francesco Romani, Professor, Università di Pisa, Italy

Examiner
Nicola Guglielmi, Professor, Università Dell'Aquila, Italy

Research Advisor
Stefano Serra Capizzano, Professor, Università Dell'Insubria, Italy

Representative of the Dean of Faculty

Dissertation accepted: November 15th, 2010

Abstract

In this work we are interested in standard and less standard structured linear systems coming from applications in various fields of computational mathematics and often modeled by integral and/or differential equations. Starting from classical Toeplitz and Circulant structures, we consider some extensions as g-Toeplitz and g-Circulants matrices appearing in several contexts in numerical analysis and applications. Then we consider special matrices arising from collocation methods for differential equations: also in this case, under suitable assumptions we observe a Toeplitz structure. More in detail we first propose a detailed study of singular values and eigenvalues of g-circulant matrices and then we provide an analysis of distribution of g-Toeplitz sequences. Furthermore, when possible, we consider Krylov space methods with special attention to the minimization of the computational work. When the involved dimensions are large, the Preconditioned Conjugate Gradient (PCG) method is recommended because of the much stronger robustness with respect to the propagation of errors. In that case, crucial issues are the convergence speed of this iterative solver, the use of special techniques (preconditioning, multilevel techniques) for accelerating the convergence, and a careful study of the spectral properties of such matrices. Finally, the use of radial basis functions allow of determining and studying the asymptotic behavior of the spectral radii of collocation matrices approximating elliptic boundary value problems.

Key words: circulant matrices, Toeplitz sequences, spectral properties, approximations, preconditioning, g-circulant matrices, g-Toeplitz sequences, singular values, eigenvalues, distribution, linear systems, Krylov space methods, multigrid methods, regularizing techniques, collocation matrices, elliptic boundary value problems, RBFs, spectral radii, block Toeplitz matrices with unbounded generating functions.

Acknowledgements

First and foremost I would like to express my sincere gratitude to my advisor Prof. Stefano SERRA CAPIZZANO who, despite its multiple busies has accepted to supervise this Thesis. Its guidance, criticism, encouragement, and assistance have been a great help and I am indebted to both for providing the opportunity for me to develop the results presented in this Thesis.

I wish to thank all the academic staff of the Department. They have always been available for suggestions and advice during several steps of the research.

Special thanks to scientific committee that has evaluated this Thesis. Without their appreciation the problems solved in this dissertation would be not presented.

I also thank my friend Debora SESANA and Dr. Marco DONATELLI for their personal contributions. Without them, certain new results would be not obtained.

The experimental work would not have been possible without the contribution of Dr. Claudio Estatico. Estatico spent long hour in the workshop to accurately check and to improve all the computational work presented in this Thesis.

I am grateful to all the PhD student of the $XXIII$ cycle programme of PhD in computational and applied mathematics. In particular to: Luisa De FRANCESCO ALBASINI, Stefano HAJEK and Debora SESANA. I enjoyed their friendship and support. Many instructive discussions were held with them and several unwinding and enthusiastic events were also appreciated. Special thanks are due to all my colleagues selected for the PhD courses in computational and applied mathematics for sharing happy as well as anxious times.

A heartfelt thank to the Italian Government for the PhD fellowship, the scholarship was very useful during the three years of PhD studies.

Finally, my deepest gratitude is kept for the people closest to me. Without their encouragement, love and understanding I would not have been able to complete this dissertation.

Declaration

This Thesis presents a work carried out in the Department of Physics and Mathematics of the Insubria university (Italy). The obtained new results in this dissertation are the own works of my supervisor Stefano SERRA CAPIZZANO, my friend Debora SESANA, and myself. Any quotation from, or description of the works of others is acknowledged herein by references to the sources, whether published or unpublished.

This dissertation is not the same as any that I have submitted for any degree, diploma or other qualification at any university. No part of this Thesis has been or is being concurrently submitted for any such degree, diploma or other qualification.

Eric Ngondiep

November 29th, 2010.

Contents

Lists of Figures, Tables and Abbreviations

LIST OF FIGURES

Figure 9.1. Shows that the eigenvalues of $\mathcal{P}_{d_n}(f;g)$, for $d_n = 999$, plotted with respect to a uniform grid points $x_k = \frac{k\pi}{1000}$, $k = 0, 1, 2, ..., 999$, form a curve which has the expected shape of f/g.

Figure 9.2. Shows perfect argument of the spectrum of the preconditioned matrix with the behavior of the function f/g. More precisely, notice that all the eigenvalues belong to the interval with the only exception of a few outliers, and the ones closest to zero (± 0.4) are not very close to zero.

Figure 9.3. Shows that the eigenvalues of the preconditioned matrix $\mathcal{P}_{d_n,d_n}(f;g)$ for $d_n = 99$, plotted with respect to a uniform grid points $x_{jk} = (\frac{j\pi}{100}, \frac{k\pi}{100})$, $j, k = 0, 1, 2, ..., 99$, are concentrated in the interval $[-1, -\frac{1}{2}]$ except few of them which are outliers. It follows that Figure 9.3 shows perfect argument of the spectrum of the preconditioned matrix $\mathcal{P}_{d_n,d_n}(f;g)$ with the behavior of the function $\frac{f}{g}$.

Figure 9.4. Shows the (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").

Figure 9.5. $g = 3$ (coprime case) - Singular values of g-Toeplitz matrices A, Natural (top) and Optimal (bottom) g-circulant preconditioners P and corresponding preconditioned matrices $P^{-1}A$.

Figure 9.6. $g = 7$ (coprime case) - Singular values of g-Toeplitz matrices A, Natural (top) and Optimal (bottom) g-circulant preconditioners P and corresponding preconditioned matrices $P^{-1}A$.

Figure 9.7. $g = 2$ (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 v(center), and analogous spectral distributions related to the regularized preconditioners (right).

Figure 9.8. 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 9.9. Shift-variant blurred data, projected data (shift-invariant blur), deblurred data.

LIST OF TABLES

Table 9.1. $g = 3$: Best relative residual $\|A^\star A x_k - A^\star b_\eta\|/\|A^\star b_\eta\|$, with corresponding iteration number k and relative restoration error $\|x_k - x^\dagger\|/\|x^\dagger\|$, with respect to different noise levels $\delta = \|b - b_\eta\|/\|b\|$ of the CGNR and PCGNR with optimal g-circulant preconditioner.

ABBREVIATIONS

SPD: Symmetric Positive Definite

BSTMSTB: Block Symmetric Toeplitz Matrices with Symmetric Toeplitz Blocks

CG: Conjugate Gradient

PCG: Preconditioned Conjugate Gradient

CGNR Method: Conjugate Gradient Method Applied on the Normal Equations

PCGNR Method: Preconditioned Conjugate Gradient Method Applied on the Normal Equations

GMRES: Generalized Minimal Residual

PGMRES: Preconditioned Generalized Minimal Residual

QGMRES or QGMRES(l): Incomplete Quasi-minimal Generalized Minimal Residual

GMRES(N): Restart Generalized Minimal Residual

QMR: Quasi-minimal Residual

BCG or Bi-CG: Bi-Conjugate Gradient

Bi-CGSTAB: Bi-Conjugate Gradient Stab

TGM: Two Grid Method

MGM: MultiGrid Method

LDL^T: Incomplete Choleski Factorization

ILUT(k): Incomplete LU Decomposition

EL: Equal Localization

ED: Equal Distribution

SEL: Strong Equal Localization

SED: Strong Equal Distribution

ϵ-ED: ϵ Equal Distribution

ϵ-EL: ϵ Equal Localization

ϵ-SED: ϵ Strong Equal Distribution

ϵ-SEL: ϵ Strong Equal Localization

MQ: Multiquadric

IMQ: Inverse Multiquadric

DFT: Discrete Fourier Transform

FFT: Fast Fourier Transform

LPO: Linear Positive Operator

SVD: Singular Value Decomposition

PDEs: Partial Differential Equations

RBFs: Radial Basis Functions

Eig: Set of Eigenvalues

Sval: Set of Singular Values

(n, g)**:** Greater Common Divisor of n and g

$\widetilde{Z}_{n,g}$**:** Submatrix of $Z_{n,g}$ obtained by considering its first $n/(n, g)$ columns

$\widehat{Z}_{n,g}$**:** Submatrix of $Z_{n,g}$ obtained by considering its first $\lceil n/g \rceil$ columns

$\widetilde{T}_{n,g}$**:** Submatrix of $T_{n,g}$ obtained by considering its first $n/(n, g)$ columns

$\widehat{T}_{n,g}$**:** Submatrix of $T_{n,g}$ obtained by considering its first $\lceil n/g \rceil$ columns

Introduction

This Thesis is devoted to the Approximation and Spectral Analysis of Special Classes of Large Structured Systems.

More specifically we consider structures of Toeplitz type and related generalizations. A Toeplitz matrix T_n is characterized by the fact that its entries are constant along diagonals, that is, $(T_n)_{j,k} = a_{j-k}$ for some coefficients $\{a_s\}_{s=1-n}^{n-1}$ with n being the size of the matrices. An interesting variant consists in considering the shift with a stepsize different from 1. If the stepsize is g we encounter g-Toeplitz matrix. More explicitly $T_{n,g}$ is a g-Toeplitz matrix of size n if $(T_{n,g})_{j,k} = a_{j-kg}$ for suitable coefficients and $j, k = 0, 1, ..., n-1$.

g-Toeplitz matrices arises in wavelet analysis [50] and subdivision algorithm or, equivalently, in the associated refinement equations, see [58] and references therein. Systems of linear equations associated with a Toeplitz structure are encountered in many two-dimensional digital signal processing applications, such as linear prediction and estimation [86], [97], [98], image restoration [66], and the approximation by radial basis functions (RBFs) of constant-coefficients elliptic boundary value problems. Toeplitz matrices are also related to Multigrid methods (see the interesting book by G. Strang [150] which gives a lot of useful insights about this topic), and they appear in certain restriction/prolongation operators [61, 1, 78, 162] in the context of discretization of differential and integral equations, spline functions, problems and method in physics, mathematics, digital signal processing, such as linear prediction and estimation [86, 97, 98], image restoration [66]. It is worth noticing that the use of different boundary conditions is quite natural when dealing with signal/image restoration problems or differential equations, see [129], [126]. To approximate and study the spectral radii of matrices A_{d_n}, where A_{d_n} are $d_n \times d_n$ collocation matrices approximating elliptic boundary value problems, we find a sequence $\{T_{d_n}\}_n$ of $d_n \times d_n$ symmetric block Toeplitz matrices with symmetric Toeplitz blocks which is equally distributed and equally localized as the sequence $\{A_{d_n}\}_n$. In order to solve the block Toeplitz system $T_{d_n}u = b$, where T_{d_n} is an $d_n \times d_n$ matrix with $d_n \times d_n$ blocks, by direct methods, such as Levinson-type algorithms, requires $O(d_n^3 \times d_n^2)$ operations [8], [113], [169]. Recently, there has been active research on the application of iterative methods such as the preconditioned conjugate gradient (PCG) method to solution of Toeplitz systems. To accelerate the convergence rate, various preconditioners have been proposed for symmetric positive definite (SPD) Toeplitz matrices [38], [83], [92]. The proposed preconditioning techniques can be easily generalized to block Toeplitz matrices. Simply speaking, we construct the preconditioners in special matrix algebras related to fast transforms such as the circulant algebra, the Tau algebra [12] and the general trigonometric matrix algebras. The key for choosing the preconditioner P_{d_n} is that its spectral behavior has to be as close as possible to that of T_{d_n}. An important condition is that the sequence $\{P_{d_n}^{-1}T_{d_n}\}_n$ is spectrally clustered at 1: in that case the related preconditioned Krylov methods are very fast. Since $P_{d_n}^{-1}u$ and $T_{d_n}u$, where u denotes an arbitrary vector of length d_n, can be performed with $O(d_n log\, d_n)$ operations via two-dimensional fast Fourier transform, the computational complexity per preconditioned Krylov iteration is $O(d_n log\, d_n)$ only. The preconditioned Krylov method can be much more attractive than direct methods for solving block Toeplitz systems if it converges fast.

In this Thesis, we first study the problem of the spectral analysis of circulant and Toeplitz matrix sequences. Then we consider the approximation and preconditioning problem by using the Korovkin theory. Furthermore, we address the problem of characterizing the singular values and eigenvalues of g-circulant matrices 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 $\{a_k\}_k$, defining the entries of the matrices, can be interpreted as the sequence of the Fourier coefficients of an integrable function f over the domain $(-\pi, \pi)$. Thirdly, we present powerful iterative methods for solving of large systems of linear equations $Ax = b$ in which A is a (real) nonsingular $n \times n$ matrix. More precisely, we restrict our study on the Krylov space methods and present a general idea of Multigrid methods. In general, one obtains such systems by using difference methods or finite element methods for solving boundary value problems in partial differential equations. The Krylov space methods start with an initial vector $x^{(0)}$ and subsequently produce a sequence of vectors

$$x^{(0)} \to x^{(1)} \to x^{(2)} \to \cdots \to x^{(m)} \to \cdots$$

which converges toward the desired solution $x = A^{-1}b$. The general characteristics of these methods is that the methods, in exact arithmetic, terminate with the exact solution x_m after at most n steps, that is, $m \leq n$. However, if the spectrum of the matrix has good localization features or is clustered, the convergence, up to a fixed error, can be obtained with m much smaller than n. If the original matrix A does not possess such good spectral properties, then we can use preconditioning. More precisely we look for a matrix P such that

1) A linear system with matrix P is cheap to solve (at most the cost of a matrix vector product with matrix A).

2) $P^{-1}A$ has a spectrum with good localization and/or clustering features (so that the number m is much less than n).

In analogy with the well studied Toeplitz case, we consider the preconditioning problem of g-Toplitz sequences via g-circulants. In particular, we consider the general case with $g \geq 2$ and the interesting result is that the preconditioned sequence $\{\mathcal{P}_n\}_n$ cannot be clustered at 1 so that the case of $g = 1$ is exceptional and, by the way, widely studied in the literature (the clustering at 1 of the preconditioning sequence is referred as optimal preconditioning; see e.g. [50, 59] for the one-level case, [141] for the multilevel case, and [150] for the multilevel block case). However, 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. Generalizations to the block and multilevel case are also considered. Finally, we approximate the elliptic boundary value problems by the linear systems of types $A_{d_n} v = b$. The shown method for approximating is based on the Radial Basis Functions. These types of approximations can be applied for giving a numerical solution of certain PDEs. Under certain conditions, the convergence is very fast (exponential in the number of grid points) when compared with Finite Differences or Finite Elements. The price that is paid is often an extreme ill-conditioning of the resulting structured matrices. In these methods, a radial function is core for approximation space and this space is made by translating a standard radial function with zero as its center (core), to all of the space particles. Here, we present an interesting method using the nodes that most of them are selected out of real domain and the others, in the domain. One of the advantages of meshless methods based on the RBFs is high decrease of computational volume that arises when changing multi-dimensions to one dimension. Kansa, [88] is the first researcher who applied an approximation by RBFs (Pseudo interpolation) to the PDEs. The use of the globally supported **RBFs** leads to dense, poorly conditioned, large linear systems as will be shown in the following. A RBFs must be selected experimentally suitable for the model problem. The Thesis is organized as follows:

The chapters 1 and 2 deal with the spectral analysis of circulant and Toeplitz matrices. In these chapters, we study spectral properties of circulant and Toeplitz matrices and we provide an analysis of optimal approximation for the Toeplitz sequence by a Korovkin-type

theory for finite Toeplitz operators via matrix algebra.

In chapters 3 and 4, we address the problem of characterizing the singular values and eigenvalues of g-circulant matrices 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 $\{a_k\}_k$, defining the entries of the matrices, can be interpreted as the sequence of Fourier coefficients of an integrable function f over the domain $(-\pi, \pi)$. As a byproduct, we show interesting relations with the analysis of convergence of multigrid methods given, e.g., in [141, 1] and we generalize the analysis to the block, multilevel 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 $p \times q$.

The powerful iterative methods, such as Krylov space methods and Multigrid methods, for solving of large systems of linear equations $Ax = b$ in which A is a (real) nonsingular $n \times n$ matrix are presented in chapters 5 and 6. The Krylov space methods generate iterates x_k that approximate the solution of linear equations $Ax = b$ best among all vector x_j such that $x_j - x_0$ belong to $K_k(r_0, A)$, where $K_k(r_0, A)$ is the Krylov space belonging to the matrix A and the starting vector $r_0 := b - Ax_0$ given by the residual of x_0. $K_k(r_0, A) := span[r_0, Ar_0, ..., A^{k-1}r_0]$, $k = 1, 2, ...$ Because of roundoff errors, these methods do not terminate with the desired solution after finitely many steps. As in true iterative methods, an infinite number of steps needs to be carried out to speed of convergence of the iterates x_k. The amount of work per step $x^{(k)} \to x^{(k+1)}$ roughly equals that of multiplying the matrix A by a vector. For this reason, these methods are especially advantageous for sparse unstructured matrices A as they occur, e.g in network calculations. They can be competitive with direct methods also for dense structured and nonstructured matrices or for band matrices, if the convergence speed is good. In any case, iterative methods have to be preferred from the point of view of the accuracy when the linear systems are of large dimension. Among these methods, one can note:

i. The conjugate gradient (CG) method proposed by Hestness and Stiefel (1952, [80]) for systems with a positive definite matrix.

ii. The generalized minimal residual (GMRES) method of Saad and Schultz (1986, [116]) (more expensive) but is defined for general linear systems with a nonsymmetric nonsingular matrix.

iii. The quasi-minimal residual method (QMR method) of Freud and Nachtigal (1991, [65]), for solving arbitrary sparse linear systems of equations. This method is based on the more efficient (but numerically more sensitive) biorthogonalization algorithm of Lanczos (1950, [95]), provides non-orthogonal bases $v_1, v_2, ..., v_k$ for the Krylov spaces $K_k(r_0, A)$ of dimension k. Using these bases, one can compute iterates $x_k \in x_0 + K_k(r_0, A)$ with an approximately minimal residual.

iv. The biconjugate gradient algorithm (Bi-CG) due to Lanczos (1950, [95]) and thoroughly studied by Fletcher (1976, [63]) is also a method for solving linear systems of equations with an arbitrary matrix A. It is an inexpensive, natural generalization of the cg-algorithm, and also generates iterates $x_k \in x_0 + K_k(r_0, A)$.

With regard to the applicability of Krylov space methods, the same remarks apply as for the classical iterative methods [see Varga (2000), Young and Axelsson (1994, [4]), and Saad (1996, [114])]. However, the very large systems of linear equations are related with the solution of boundary-value problems of partial differential equations by finite element techniques and are mainly solved by Multigrid methods; e.g, Hackbusch (1985, [98]), Braess (1997, [18]), Bramble (1993, [19]), Quarteroni and Valli (1997).

Chapter 7 is reserved to new material which is collected as the manuscript entitled: A note on preconditioning g-Toeplitz sequences via g-circulants. As already remarked, we consider the preconditioned problems, and we focalize our concerns in the general case with $g \geq 2$. In this case, the main result is that the spectrum of the preconditioned sequence

$\{\mathcal{P}_n\}_n$ can not be clustered at 1. This result is different from the widely studied case with $g = 1$ (where the spectrum can be clustered at 1, [36, 38, 123, 124]) which cannot be generalized to $g \geq 2$. However, although the clustering at unity cannot be achieved, the g-circulant preconditioning tool is useful because it has good regularizing properties. Generalization to the block multilevel cases are also considered in this chapter.

In chapter 8, we treat the problem of preconditioning of collocation matrices approximating elliptic boundary value problems. The shown method of approximation is based on the RBFs. Here, we present an interesting method using the nodes that most of them are selected out of real domain and the others, in the domain. We determine in any case (unidimension and multi-dimensions) the preconditioners and study the condition numbers of obtained matrices. However, we find a sequence $\{T_{d_n}\}_n$ of symmetric Toeplitz matrices (respectively, symmetric block Toeplitz matrices with symmetric Toeplitz blocks (SBTMSTB) in the case of two-dimensions) which is equally distributed and equally localized as the sequence of collocation matrices $\{A_{d_n}\}_n$ approximating elliptic boundary value problems. We determine the preconditioners in the Tau class for the Toeplitz (respectively, block Toeplitz) matrices and study the asymptotic growth of their spectral radii. For this purpose and since a RBFs must be selected experimentally suitable for the model problem, some of the most commonly used RBFs are:

- Direct Multiquadric: $\phi(t) = (t^2 + c^2)^{\frac{1}{2}}$,
- Inverse Multiquadric: $\phi(t) = (t^2 + c^2)^{-\frac{1}{2}}$,
- Gaussian: $\phi(t) = e^{-\frac{t^2}{c^2}}$,

where c is a shape parameter which determines the "accuracy" and the "stability".

In chapter 9, we apply the PCG algorithm for solving the systems of linear equations $T_{d_n} v = \tilde{f}$, in the case where T_{d_n} are $d_n \times d_n$ symmetric Toeplitz matrices (with bounded generating function) (respectively, symmetric block Toeplitz matrices with symmetric Toeplitz blocks (with generating functions just integrable)) related to the collocation matrices defined in chapter 8, and we present some numerical results for such systems.

We end the dissertation by drawing the general conclusions in chapter 10.

Chapter One

Spectral Analysis of Toeplitz and Circulant Matrix Sequences

1.1 Introduction

Throughout this chapter, we study the spectral properties of the circulant and Toeplitz matrices. We put special attention to the case where the entries of the Toeplitz matrices come from the Fourier coefficients of a given function f, usually called as symbol, defined on $(-\pi, \pi)$. The circulant matrices will be chosen as optimal approximation, to Frobenius distance, of the Toeplitz counterparts. In chapter 2 we will use the Korovkin theory to derive clustering of the related preconditioned sequence, when the preconditioners are selected in trigonometric matrix algebras such as Circulants, Tau, etc...

Definition 1.1.1. *A Toeplitz matrix is an $n \times n$ matrix $T_n = [t_{k,j}]_{k,j=0}^{n-1}$ where $t_{k,j} = t_{k-j}$, i.e., a matrix of the form*

$$T_n = \begin{bmatrix} t_0 & t_{-1} & \dots & t_{-(n-2)} & t_{-(n-1)} \\ t_1 & t_0 & t_{-1} & \dots & t_{-(n-2)} \\ \vdots & t_1 & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & t_0 & t_{-1} \\ t_{n-1} & t_{n-2} & \dots & t_1 & t_0 \end{bmatrix} \tag{1.1}$$

Such matrices arise in many applications. For examples, in solutions to differential and integral equations, spline functions, and problems and methods in physics, mathematics, statistics, and signal processing.

A common special case of Toeplitz matrices which will result in significant simplification and play a fundamental role in developing more general results when every row of the matrix is a right cyclic shift of the row above it so that $t_k = t_{-(n-k)} = t_{k-n}$ for $k = 0, 1, ..., n-1$. In this case, the picture takes the form

$$\begin{bmatrix} t_0 & t_{-1} & \dots & t_{-(n-2)} & t_{-(n-1)} \\ t_{-(n-1)} & t_0 & t_{-1} & \dots & t_{-(n-2)} \\ t_{-(n-2)} & t_{-(n-1)} & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & t_0 & t_{-1} \\ t_{-1} & t_{-2} & \dots & t_{-(n-1)} & t_0 \end{bmatrix} \tag{1.2}$$

Definition 1.1.2. *A circulant matrix $C_n = [c_{k,j}]_{k,j=0}^{n-1}$ is a Toeplitz matrix defined by the form given in* (1.2)*. The structure can also be characterized by noting that the (k, j) entry of C_n, $c_{k,j}$ is given by $c_{k,j} = c_{(k-j) mod n}$.*

Circulant matrices arise, for example, in applications involving the discrete Fourier transform (DFT) and the study of cyclic codes for error correction.

A great deal is known about the behavior of Toeplitz matrices, the most common and complete references being Grenander and Szegö [77] and Widom [170]. A more recent text devoted to the subject is Böttcher and Silbermann [16].

The most famous and arguably the most important result describing Toeplitz matrices is Szegö theorem for sequences of Toeplitz matrices $\{T_n\}_n$ which deals with the behavior of the eigenvalues as n goes to infinity. Szegö theorem deals with the asymptotic behavior of the eigenvalues of a sequence of Hermitian Toeplitz matrices $T_n = [t_{k-j}]_{k,j=0}^{n-1}$. The theorem requires that several technical conditions be satisfied, including the existence of the Fourier series with coefficients t_k related to each other by

$$f(x) = \sum_{k=-\infty}^{\infty} t_k e^{-\hat{i}kx}, \tag{1.3}$$

$$t_k = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x) e^{-\hat{i}kx} dx. \tag{1.4}$$

Thus the sequence $\{t_k\}_n$ determines the function f and vice-versa, hence the sequence of matrices is often denoted as $T_n = T_n(f)$.

1.2 Spectral properties of circulant matrices.

The following theorems summarize the properties regarding the eigenvalues and eigenvectors of circulant matrices and provides some implications.

Theorem 1.2.1. [73] *Every circulant matrix C_n has eigenvectors*

$$y^{(m)} = \frac{1}{\sqrt{n}} [1, e^{-2\pi\hat{i}m/n}, ..., e^{-2\pi\hat{i}m(n-1)/n}]^T, \quad m = 0, 1, 2, ..., n-1$$

and the corresponding eigenvalues

$$\psi_m = \sum_{k=0}^{n-1} c_k e^{-2\pi\hat{i}km/n}$$

and can be expressed in the form $C_n = U \Psi_n U^\star$, where U has eigenvectors as columns in order and Ψ_n is $diag(\psi_j,\ j = 0, 1, ..., n-1)$. In particular, all circulant matrices share the same eigenvectors, the same unitary matrix U works for all circulant matrices, and any matrix of the form $C = U \Psi U^\star$ is circulant.

Proof. The eigenvalues ψ_m and the eigenvectors $y^{(m)}$ of C_m are the solutions of

$$C_m y = \psi y \tag{1.5}$$

or equivalently, of the n difference equations

$$\sum_{k=0}^{m-1} c_{n-m+k} y_k + \sum_{k=m}^{n-1} c_{k-m} y_k = \psi y_m; \quad m = 0, 1, 2, ..., n-1. \tag{1.6}$$

Changing the summation dummy variable results in

$$\sum_{k=0}^{n-m-1} c_k y_{k+m} + \sum_{k=n-m}^{n-1} c_k y_{k-(n-m)} = \psi y_m; \quad m = 0, 1, 2, ..., n-1. \tag{1.7}$$

One can solve difference equations as one solves differential equations by guessing an intuitive solution and then proving that it works. Since the equation is linear with constant coefficients, a reasonable guess is $y_k = \phi^k$ (analogous to $y(t) = e^{st}$ in linear time invariant differential equations). Substitution into (1.7) and cancelation of ϕ^m yields

$$\sum_{k=0}^{n-m-1} c_k \phi^k + \phi^{-n} \sum_{k=n-m}^{n-1} c_k \phi^k = \psi.$$

Thus, if we choose $\phi^{-n} = 1$, i.e., ϕ is one of the n distinct complex n^{th} roots of unity, then we have an eigenvalue

$$\psi = \sum_{k=0}^{n-1} c_k \phi^k \tag{1.8}$$

with corresponding eigenvector

$$y = n^{-1/2}(1, \phi, \phi^2, ..., \phi^{n-1})^T \tag{1.9}$$

where the normalization is chosen to give the eigenvector unit energy. Choosing $\phi_m = e^{-2\pi\hat{\imath}m/n}$, we have eigenvalue

$$\psi_m = \sum_{k=0}^{n-1} c_k e^{-2\pi\hat{\imath}mk/n} \tag{1.10}$$

and eigenvector

$$y^{(m)} = \frac{1}{\sqrt{n}}[1, e^{-2\pi\hat{\imath}m/n}, e^{-4\pi\hat{\imath}m/n}, ..., e^{-2\pi\hat{\imath}m(n-1)/n}]^T.$$

Thus, from the definition of eigenvalues and eigenvectors

$$C_m y^{(m)} = \psi_m y^{(m)}; \quad m = 0, 1, 2, ..., n-1. \tag{1.11}$$

Equation (1.10) should be familiar to those with standard engineering back grounds as simply the discrete Fourier transform (DFT) of the sequence $\{c_k\}_k$. Thus we can recover the sequence $\{c_k\}_k$ from ψ_k by the Fourier inverse formula. In particular

$$\frac{1}{n}\sum_{m=0}^{n-1} \psi_m e^{2\pi\hat{\imath}ml/n} = \frac{1}{n}\sum_{m=0}^{n-1}\sum_{k=0}^{n-1} \left(c_k e^{-2\pi\hat{\imath}mk/n}\right) e^{2\pi\hat{\imath}ml/n} = \sum_{k=0}^{n-1} c_k \frac{1}{n}\sum_{m=0}^{n-1} e^{2\pi\hat{\imath}m(l-k)/n} = c_l, \tag{1.12}$$

where we have used the orthogonality of the complex exponentials:

$$\sum_{m=0}^{n-1} e^{2\pi\hat{\imath}mk/n} = n\delta_{k(mod\, n)} = \begin{cases} n & \text{if } k(mod\, n) = 0 \\ 0 & \text{otherwise} \end{cases} \tag{1.13}$$

where δ is the Kronecker delta, i.e.,

$$\delta_m = \begin{cases} 1 & \text{if } m = 0 \\ 0 & \text{otherwise} \end{cases}$$

Thus the eigenvalues of a circulant matrix comprise the DFT of the first row of the circulant matrix, and conversely first row of a circulant matrix is the inverse DFT of the eigenvalues.

Equation (1.11) can be written as a single matrix equation

$$C_n U = U\Psi_n \tag{1.14}$$

where

$$U = [y^{(0)}|y^{(1)}|...|y^{(n-1)}] = \frac{1}{\sqrt{n}}[e^{-2\pi\hat{i}mk/n}]_{m,k=0}^{n-1}$$

is the matrix composed of the eigenvectors as columns and $\Psi_n = diag(\psi_k,\ \ k = 0, 1, ..., n-1)$ is the diagonal matrix with diagonal elements $\psi_0, \psi_1, ..., \psi_{n-1}$. Furthermore, (1.13) implies that U is unitary. By way of details, denote that the (k, j)-th element of $UU^\star$ by $a_{k,j}$ and observe that $a_{k,j}$ will be the product of the k-th row of U, which is $\frac{1}{\sqrt{n}}[1, e^{-2\pi\hat{i}k/n}, ..., e^{-2\pi\hat{i}k(n-1)/n}]$, times the j-th column of $U^\star$, which is $\frac{1}{\sqrt{n}}[1, e^{2\pi\hat{i}j/n}, ..., e^{2\pi\hat{i}j(n-1)/n}]^T$, so that:

$$a_{k,j} = \frac{1}{n}\sum_{m=0}^{n-1} e^{2\pi\hat{i}m(j-k)/n} = \delta_{(k-j)mod\, n}$$

and hence $UU^\star = I$. Similarly, $U^\star U = I$. Thus (1.14) implies that

$$C_n = U\Psi_n U^\star, \tag{1.15}$$

$$\Psi_n = U^\star C_n U. \tag{1.16}$$

Since C_n is unitary similar to a diagonal matrix, it is normal. □

Theorem 1.2.2. [75] *Let α and β be two complex numbers, and let $C_n = [c_{k-j}]_{k,j=0}^{n-1}$ and $B_n = [b_{k-j}]_{k,j=0}^{n-1}$ be $n \times n$ circulant matrices with eigenvalues*

$$\psi_m = \sum_{k=0}^{n-1} c_k e^{-2\pi\hat{i}mk/n};\quad \beta_m = \sum_{k=0}^{n-1} b_k e^{-2\pi\hat{i}mk/n},\ \ m = 0, 1, 2, ..., n-1,$$

respectively. Then

(1) *C_n and B_n commute and*

$$C_n B_n = B_n C_n = U\Gamma_n U^\star,$$

where $\Gamma_n = diag(\psi_m\beta_m;\ \ m = 0, 1, 2, ..., n-1)$, and C_nB_n is also a circulant matrix.

(2) *$\alpha C_n + \beta B_n$ is a circulant matrix and*

$$\alpha C_n + \beta B_n = U\Omega_n U^\star \ \ where\ \ \Omega_n = diag(\alpha\psi_m + \beta\beta_m;\ \ m = 0, 1, 2, ..., n-1).$$

(3) *If $\psi_m \neq 0$ for all $m = 0, 1, 2, ..., n-1$, then C_n is nonsingular and*

$$C^{-1} = U\Psi_n^{-1}U^\star.$$

Proof. According to Theorem 1.2.1, we have $C_n = U\Psi_n U^\star$ and $B_n = U\Phi_n U^\star$ where $\Psi_n = diag(\psi_m;\ \ m = 0, 1, 2, ..., n-1)$ and $\Phi_n = diag(\beta_m;\ \ m = 0, 1, 2, ..., n-1)$.

(1) $C_nB_n = U\Psi_nU^\star U\Phi_nU^\star = U\Psi_n\Phi_nU^\star = U\Phi_n\Psi_nU^\star = B_nC_n$. Since $\Psi_n\Phi_n$ is diagonal, the first part of the theorem implies that C_nB_n is circulant.

(2) $\alpha C_n + \beta B_n = \alpha U\Psi_n U^\star + \beta U\Phi_n U^\star = U(\alpha\Psi_n + \beta\Phi_n)U^\star = U\Omega_n U^\star$ where $\Omega_n = \alpha\Psi_n + \beta\Phi_n = diag(\alpha\psi_m + \beta\beta_m;\ \ m = 0, 1, 2, ..., n-1)$. Then $\alpha C_n + \beta B_n$ is a circulant matrix.

(3) If Ψ_n is nonsingular, then

$$CU\Psi_n^{-1}U^\star = U\Psi_nU^\star U\Psi_n^{-1}U^\star = UU^\star = I,\ \ \text{also}$$

$$U\Psi_n^{-1}U^\star C = U\Psi_n^{-1}U^\star U\Psi_nU^\star = UU^\star = I.$$

Then C_n is nonsingular and $C_n^{-1} = U\Psi_n^{-1}U^\star$. □

In the following, we denote by $\mathfrak{A} = \{C \in \mathcal{M}_n(\mathbb{C}) |\ C \text{ is a circulant matrix}\}$ the set of all the $n \times n$ circulant matrices.

Definition 1.2.1. *A set $\mathbb{A} \subset \mathcal{M}_n(\mathbb{C})$ is a matrix algebra if for every complex number α and for $A, B \in \mathbb{A}$:*

(i) $A + B \in \mathbb{A}$,

(ii) $\alpha A \in \mathbb{A}$,

(iii) $AB \in \mathbb{A}$.

Lemma 1.2.1. *The set $\mathfrak{A}$ is a matrix algebra of dimension n.*

Proof. First of all, let us define the circulant matrix

$$\Pi = \begin{bmatrix} 0 & \ddots & \cdots & 0 & 1 \\ 1 & 0 & & & 0 \\ 0 & 1 & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & 1 & 0 \end{bmatrix}$$

and the vector space $\mathbb{P}_{n-1} = \{p(\Pi) | p \text{ is a polynomial of degree at most equal to } n-1\}$. Then $\mathbb{P}_{n-1}$ is an algebra generated by Π. So Π has a minimal polynomial p of degree $n-1$. Now, let $C \in \mathfrak{A}$ and $(c_0, c_1, ..., c_{n-1})^T$ the first column of C. Since for $k = 0, 1, 2, .., n-1$

$$\Pi^k = \begin{pmatrix} 0 & \cdots & 0 & 1 & & 0 \\ \vdots & & \vdots & & \ddots & \\ 0 & \cdots & 0 & 0 & & 1 \\ 1 & & 0 & 0 & \cdots & 0 \\ & \ddots & & \vdots & & \vdots \\ 0 & & 1 & 0 & \cdots & 0 \end{pmatrix}$$

it is obvious that

$$C = \sum_{k=0}^{n-1} c_k \Pi^k \in \mathbb{P}_{n-1}$$

so $\mathfrak{A} \subset \mathbb{P}_{n-1}$.

Conversely, let $r(\Pi) \in \mathbb{P}_{n-1}$, since for all $k = 01, 2, ..., n-1$, Π^k is a circulant matrix, it follows from theorem 1.2.1 that $r(\Pi) \in \mathfrak{A}$, so $\mathbb{P}_{n-1} \subset \mathfrak{A}$. Because $\mathbb{P}_{n-1}$ is a matrix algebra of dimension n, one deduces that $\mathfrak{A}$ is a matrix algebra of dimension n. □

A detailed study of the optimal approximation of Toeplitz sequences $\{T_n(f)\}$ sometimes requires the knowledge of some properties of Toeplitz matrices $T_n(f)$.

1.3 Spectral properties of Toeplitz matrices

This section deals with some properties of the Toeplitz matrices $T_n(f) = [t_{k-j}]_{k,j=0}^{n-1}$ in the case where $\{t_k\}_k$ is the sequence of the Fourier coefficients of an integrable function f over the domain $Q = (-\pi, \pi)$.

Lemma 1.3.1. *Let $f \in L^1(Q)$. The Toeplitz matrix $T_n(f)$ is Hermitian if and only if f is real-valued.*

Proof. First of all, let us suppose that f is real-valued. For $k = 0, 1, 2, ..., n-1$

$$t_{k-j} = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(x)e^{-2\pi\hat{i}(k-j)x}dx,$$

then

$$\begin{aligned} t_{k-j}^{\star} &= \frac{1}{2\pi}\overline{\int_{-\pi}^{\pi} f(x)e^{-2\pi\hat{i}(k-j)x}dx} \\ &= \frac{1}{2\pi}\int_{-\pi}^{\pi} f(x)\overline{e^{-2\pi\hat{i}(k-j)x}}dx \\ &= \frac{1}{2\pi}\int_{-\pi}^{\pi} f(x)e^{2\pi\hat{i}(k-j)x}dx \\ &= t_{-(k-j)}, \end{aligned}$$

then $T_n(f)^{\star} = T_n(f)$.

Conversely, let us suppose that $T_n(f)^{\star} = T_n(f)$, i.e., $t_{k-j}^{\star} = t_{-(k-j)}$, then

$$\begin{aligned} f^{\star}(x) &= \left(\sum_{k=-\infty}^{\infty} t_k e^{\hat{i}kx}\right)^{\star} = \sum_{k=-\infty}^{\infty} t_k^{\star} e^{-\hat{i}kx} \\ &\underset{(a)}{=} \sum_{k=-\infty}^{\infty} t_{-k} e^{-\hat{i}kx} = \sum_{k=-\infty}^{\infty} t_k e^{\hat{i}kx} = f(x). \end{aligned}$$

Whence, f is real-valued. (a) follows from the hypothesis. □

Lemma 1.3.2. *Let $f \in L^1(Q)$. The Toeplitz matrix $T_n(f)$ is symmetric if and only if f is real-valued and even.*

Proof. Let us suppose that f is real-valued and even, i.e., $f(x) = f(-x) \in \mathbb{R}$ for all $x \in (-\pi, \pi)$. Then

$$\begin{aligned} t_{k-j} &= \frac{1}{2\pi}\int_{-\pi}^{\pi} f(x)e^{-2\pi\hat{i}(k-j)x}dx \\ &= -\frac{1}{2\pi}\int_{\pi}^{-\pi} f(-x)e^{2\pi\hat{i}(k-j)x}dx \\ &= \frac{1}{2\pi}\int_{-\pi}^{\pi} f(x)e^{-2\pi\hat{i}(j-k)x}dx \\ &= t_{j-k}, \end{aligned}$$

then $T_n(f)$ is symmetric.

Conversely, assuming that $T_n(f)$ is symmetric, it follows from Lemma 1.3.1 that f is real-valued. Now, let $x \in (-\pi, \pi)$.

$$f(-x) = \sum_{k=-\infty}^{\infty} t_k e^{-\hat{i}kx} = \sum_{k=-\infty}^{\infty} t_{-k} e^{\hat{i}kx} \underset{(b)}{=} \sum_{k=-\infty}^{\infty} t_k e^{\hat{i}kx} = f(x),$$

(b) follows from $t_{-k} = t_k$. Then f is even. □

Lemma 1.3.3. *Let $f \in L^1(Q)$. The Toeplitz matrix $T_n(f)$ is Hermitian and positive definite if f is nonnegative over the domain $(-\pi, \pi)$.*

Proof. Let us suppose that f is nonnegative over the domain $(-\pi, \pi)$. Let $y = (y_0, y_1, ..., y_{n-1})^T \in \mathbb{C}^n - \{0\}$, then

$$\begin{aligned}
y^\star T_n(f) y &= (\bar{y}_0, \bar{y}_1, ..., \bar{y}_{n-1}) \left[\sum_{j=0}^{n-1} t_{k-j} y_j \right]_{k=0}^{n-1} \\
&= \sum_{k=0}^{n-1} \bar{y}_k \sum_{j=0}^{n-1} t_{k-j} y_j \\
&= \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x) \sum_{k=0}^{n-1} \bar{y}_k \sum_{j=0}^{n-1} e^{-2\hat{i}\pi(k-j)x} y_j dx \\
&= \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x) \left(\sum_{k=0}^{n-1} \bar{y}_k e^{-2\hat{i}\pi kx} \sum_{j=0}^{n-1} e^{2\hat{i}\pi jx} y_j \right) dx \\
&= \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x) \left(\sum_{k=0}^{n-1} y_k e^{2\hat{i}\pi kx} \right)^\star \left(\sum_{j=0}^{n-1} e^{2\hat{i}\pi jx} y_j \right) dx \\
&= \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x) \left| \sum_{k=0}^{n-1} y_k e^{2\hat{i}\pi kx} \right|^2 dx > 0
\end{aligned}$$

since $f \geq 0$ and $y \neq 0$. indeed, for $x \neq 0$, $\{e^{2\hat{i}\pi kx};\ k = 0, 1, ..., n-1\}$ is a basis of $\mathbb{C}^n$, so $y = (y_0, y_1, ..., y_{n-1})^T \neq 0$ implies $\sum_{k=0}^{n-1} y_k e^{2\hat{i}\pi kx} \neq 0$. Then, $T_n(f)$ is Hermitian and positive definite. □

In the following, we state without proof some fundamental theorems of linear algebra. These results are very important for the study of the bounds of eigenvalues of Hermitian Toeplitz matrices $T_n(f)$.

Theorem 1.3.1. *(Minimax or Courant-Fischer Theorem, [173]). Let A be an $n \times n$ Hermitian matrix and let $\lambda_1 < \lambda_2 < ... < \lambda_k = \lambda_{\max}$ $(k \leq n)$ be the distinct eigenvalues of A. Setting $\mathfrak{U}_i = \{U \subset \mathbb{C}^n | U$ a vector subspace of dimension $i\}$, then for $i = 1, 2, ..., k$:*

$$\lambda_i = \min_{U \in \mathfrak{U}_i} \max_{\substack{x \in U \\ x \neq 0}} \frac{x^\star A x}{x^\star x}.$$

In particular, $\mathfrak{U}_1 = \{$ lines of $\mathbb{C}^n\}$ and $U \in \mathfrak{U}_1$ implies $U = span < u >$, with $u \in \mathbb{C}^n$ and $u \neq 0$. So,

$$\lambda_{\min}(A) = \min_{\substack{x \in \mathbb{C}^n \\ x \neq 0}} \frac{x^\star A x}{x^\star x}, \quad and \quad \lambda_{\max}(A) = \max_{\substack{x \in \mathbb{C}^n \\ x \neq 0}} \frac{x^\star A x}{x^\star x}.$$

Remark 1.3.1. *Let A be an $n \times n$ Hermitian matrix and let $x \in \mathbb{C}^n$, $x \neq 0$. The quantity*

$$R(x) = \frac{x^\star A x}{x^\star x}$$

is called the RAYLEIGH QUOTIENT of A.

Theorem 1.3.2. *(Cauchy interlace Theorem, [173]). Let A be an $n \times n$ Hermitian matrix and let B be an $(n-1) \times (n-1)$ principale submatrix of A, i.e.,*

$$A = \begin{bmatrix} B & y \\ x^T & z \end{bmatrix}$$

where $x, y \in \mathbb{C}^{n-1}$ and $z \in \mathbb{C}$. Then

$$\lambda_1(A) \leq \lambda_1(B) \leq ... \leq \lambda_{n-1}(A) \leq \lambda_{n-1}(B) \leq \lambda_n(A).$$

Theorem 1.3.3. *(Monotonicity Theorem, [173]). Let B and C be two $n \times n$ Hermitian matrices, with the eigenvalues $\lambda_1(B) \leq \lambda_2(B) \leq ... \leq \lambda_n(B)$ and $\lambda_1(C) \leq \lambda_2(C) \leq ... \leq \lambda_n(C)$ respectively. Setting $A = B + C$, then for $i = 1, 2, ..., n$:*

$$\lambda_i(B) + \lambda_1(C) \leq \lambda_i(A) \leq \lambda_i(B) + \lambda_n(C).$$

Theorem 1.3.4. [170, 117, 154] *Let $f \in L^1(-\pi, \pi)$ be real-valued. Setting*

$$m_f := \inf_{x \in [-\pi,\pi]} f(x) \quad \text{and} \quad M_f := \sup_{x \in [-\pi,\pi]} f(x),$$

then

$$\Lambda(T_n(f)) \subset [m_f, M_f].$$

where $\Lambda(T_n(f))$ is the spectrum of $T_n(f)$.

Proof. Since f is real-valued, it follows from Lemma 1.3.1 that $T_n(f)$ is Hermitian. According to Theorem 1.3.1, one has

$$\lambda_{\min}(T_n(f)) = \min_{x \neq 0} \frac{x^\star T_n(f) x}{x^\star x} \text{ and } \lambda_{\max}(T_n(f)) = \max_{x \neq 0} \frac{x^\star T_n(f) x}{x^\star x}. \tag{1.17}$$

Let $x \in \mathbb{C}^n$, $x \neq 0$, then

$$\begin{aligned}
x^\star T_n(f) x &= (\bar{x}_0, \bar{x}_1, ..., \bar{x}_{n-1}) \left[\sum_{j=0}^{n-1} t_{k-j} x_j \right]_{k=0}^{n-1} \\
&= \sum_{k=0}^{n-1} \bar{x}_k \sum_{j=0}^{n-1} t_{k-j} x_j \\
&= \frac{1}{2\pi} \int_{-\pi}^{\pi} f(t) \sum_{k=0}^{n-1} \bar{x}_k \sum_{j=0}^{n-1} e^{-2\hat{i}\pi(k-j)t} x_j dt \\
&= \frac{1}{2\pi} \int_{-\pi}^{\pi} f(t) \left(\sum_{k=0}^{n-1} \bar{x}_k e^{-2\hat{i}\pi kt} \sum_{j=0}^{n-1} e^{2\hat{i}\pi jt} x_j \right) dt \\
&= \frac{1}{2\pi} \int_{-\pi}^{\pi} f(t) \left(\sum_{k=0}^{n-1} x_k e^{2\hat{i}\pi kt} \right)^\star \left(\sum_{j=0}^{n-1} e^{2\hat{i}\pi jt} x_j \right) dt \\
&= \frac{1}{2\pi} \int_{-\pi}^{\pi} f(t) \left| \sum_{k=0}^{n-1} x_k e^{2\hat{i}\pi kt} \right|^2 dt
\end{aligned}$$

then

$$m_f.\min_{x\neq 0}\left(\frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{\left|\sum_{k=0}^{n-1}x_k e^{2\hat{i}\pi kt}\right|^2}{|x|^2}dt\right)\leq \min_{x\neq 0}\frac{x^\star T_n(f)x}{x^\star x} \tag{1.18}$$

and

$$M_f.\max_{x\neq 0}\left(\frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{\left|\sum_{k=0}^{n-1}x_k e^{2\hat{i}\pi kt}\right|^2}{|x|^2}dt\right)\geq \max_{x\neq 0}\frac{x^\star T_n(f)x}{x^\star x}. \tag{1.19}$$

Or

$$\begin{aligned}\int_{-\pi}^{\pi}\left|\sum_{k=0}^{n-1}x_k e^{2\hat{i}\pi kt}\right|^2 dt &= \int_{-\pi}^{\pi}\sum_{k=0}^{n-1}\bar{x}_k\sum_{j=0}^{n-1}e^{-2\hat{i}\pi(k-j)t}x_j dt\\ &= \sum_{k=0}^{n-1}\bar{x}_k\sum_{j=0}^{n-1}x_j\int_{-\pi}^{\pi}e^{-2\hat{i}\pi(k-j)t}dt\\ &\underset{(c)}{=} 2\pi\sum_{j=0}^{n-1}|x_j|^2 = 2\pi|x|^2\end{aligned}$$

(c) follows from

$$\int_{-\pi}^{\pi}e^{-2\hat{i}\pi(k-j)t}dt=\begin{cases}2\pi & \text{if } j=k\\ 0 & \text{otherwise}\end{cases}$$

Then

$$\max_{x\neq 0}\left(\frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{\left|\sum_{k=0}^{n-1}x_k e^{2\hat{i}\pi kt}\right|^2}{|x|^2}dt\right)=\min_{x\neq 0}\left(\frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{\left|\sum_{k=0}^{n-1}x_k e^{2\hat{i}\pi kt}\right|^2}{|x|^2}dt\right)=1. \tag{1.20}$$

It follows from $(1.17)-(1.18)-(1.19)-(1.20)$ that

$$\lambda_{\min}\geq m_f \text{ and } \lambda_{\max}\leq M_f.$$

□

Theorem 1.3.5. *(Density Results, [170]). Let $f\in L^1(-\pi,\pi)$ be real-valued and $\Lambda(T_n(f))=\{\lambda_k^{(n)};\ k=1,2,...,n\}$ the spectrum of $T_n(f)$. Setting*

$$m_f:=\inf_{x\in[-\pi,\pi]}f(x) \text{ and } M_f:=\sup_{x\in[-\pi,\pi]}f(x),$$

where $\inf f$ and $\sup f$ denote the infimum and the supremum of f, then

(i) $\lim_{n\to\infty}\lambda_k^{(n)}(T_n(f))=m_f$ *and* $\lim_{n\to\infty}\lambda_{n-k}^{(n)}(T_n(f))=M_f$.

(ii) $\bigcup_{n\in\mathbb{N}} \left\{\lambda_k^{(n)}(T_n(f))\right\}_{k=1}^{n}$ *is dense in* $[m_f, M_f]$.

Dealing with iterative methods for linear systems, an important parameter is the convergence speed of the iterations towards the solution of the system. For instance, it is well known that the convergence speed of the CG method depends on the condition number of the system matrix. However, the condition number does not completely force the convergence speed, which is also controlled by the global distribution of the singular values of the system matrix. Basically, the convergence speed is good if the singular values of the system are well clustered "close" to the unity.
In order to increase the convergence speed, it is often useful to replace the system to solve with an equivalent one, in which the clustering of the eigenvalues is improved. This approach leads to the preconditioning: the system $Af = g$ is replaced by the following equivalent system, called preconditioned system,

$$P^{-1}Af = P^{-1}g \tag{1.21}$$

where the invertible matrix P is the system preconditioner. If the CGLR method is now applied to the preconditioned system, then the convergence speed is then related to the distribution of the singular values of the preconditioned matrix.
For Toeplitz systems, we can introduce the following preconditioner, called natural preconditioner.

Definition 1.3.1. *Let $f \in L^1(-\pi,\pi)$ be a real-valued even function, and let $T_n(f)$ be the symmetric Toeplitz matrix generated by f. The natural preconditioner of $T_n(f)$ is defined as*

$$P_n = T_n(f) - H(T_n(f))$$

where $H(T_n(f))$ is the Hankel matrix defined on the following form:

$$H(T_n(f)) = \begin{bmatrix} t_2 & t_3 & \dots & t_{n-1} & 0 & 0 \\ t_3 & & & & 0 & 0 \\ \vdots & & & & & t_{n-1} \\ & & & & & \vdots \\ t_{n-1} & & & & & t_3 \\ 0 & 0 & & & & t_2 \\ 0 & 0 & t_{n-1} & \dots & t_3 & t_2 \end{bmatrix} \quad \textit{when } T_n(f) = \begin{bmatrix} t_0 & t_1 & \dots & t_{n-1} \\ t_1 & t_0 & & t_{n-2} \\ \vdots & \vdots & \ddots & \vdots \\ t_{n-1} & t_{n-2} & \dots & t_0 \end{bmatrix}$$

Lemma 1.3.4. *Let $f \in L^1(-\pi,\pi)$ be an even function, and let $T_n(f)$ be the symmetric Toeplitz matrix generated by f. Setting $P_n = T_n(f) - H(T_n(f))$ and $\Delta_n = T_n(f) - P_n$, if P_n is nonsingular, $\sup_n \|P_n^{-1}\|_2 \le c < \infty$ (where c is a constant number independent of n) and the eigenvalues $\{\lambda(\Delta_n)\}_n$ are clustered around 0, then the eigenvalues $\{\lambda(P_n^{-1}T_n(f))\}_n$ are clustered around 1.*

Proof. First of all, $P_n^{-1}T_n(f) = P_n^{-1}\Delta_n + I_n$ implies $\lambda_k(P_n^{-1}T_n(f)) = \lambda_k(P_n^{-1}\Delta_n) + 1$. Since

$$P_n^{-1}\Delta_n = P_n^{-1/2}(P_n^{-1/2}\Delta_n P_n^{-1/2})P_n^{1/2}, \tag{1.22}$$

according to Theorem 1.3.1,

$$\lambda_k(P_n^{-1/2}\Delta_n P_n^{-1/2}) = \min_{U\in\mathfrak{U}_k} \max_{\substack{x\in U \\ x\neq 0}} \frac{x^\star P_n^{-1/2}\Delta_n P_n^{-1/2} x}{x^\star x}.$$

Setting $y = P_n^{-1/2}x$ for $x \in U$, then $\tilde{U} := \{y = P_n^{-1/2}x | x \in U\}$ is a subspace of dimension k. Then

$$\lambda_k(P_n^{-1/2}\Delta_n P_n^{-1/2}) = \min_{\tilde{U}\in\mathfrak{U}_k} \max_{\substack{y\in\tilde{U} \\ y\neq 0}} \frac{y^\star\Delta_n y}{y^\star P_n y} \le \min_{\tilde{U}\in\mathfrak{U}_k} \max_{\substack{y\in\tilde{U} \\ y\neq 0}} \frac{y^\star\Delta_n y}{y^\star y}\frac{y^\star y}{y^\star P_n y} \le \min_{\tilde{U}\in\mathfrak{U}_k} \max_{\substack{y\in\tilde{U} \\ y\neq 0}} \frac{y^\star\Delta_n y}{y^\star y}\cdot c,$$

since $\frac{y^\star P_n y}{y^\star y} \geq \lambda_{\min}(P_n) = \frac{1}{\lambda_{\max}(P_n)} \geq \frac{1}{c}$, then

$$\lambda_k(P_n^{-1/2}\Delta_n P_n^{-1/2}) \leq c \cdot \lambda_k(\Delta_n). \tag{1.23}$$

Because the eigenvalues $\{\lambda_k(\Delta_n)\}_n$ are clustered around 0, one deduces from (1.23) that the eigenvalues $\{\lambda_k(P_n^{-1/2}\Delta_n P_n^{-1/2})\}_n$ are clustered around 0. It follows from (1.22) that the eigenvalues $\{\lambda_k(P_n^{-1}\Delta_n)\}_n$ are clustered around 0. Hence the eigenvalues $\{\lambda(P_n^{-1}T_n(f))\}_n$ are clustered around 1. □

Lemma 1.3.5. *Let $f \in L^1(-\pi,\pi)$ be a complex-valued function and $T_n(f)$ be the Toeplitz matrix generated by f. Let U be an $n \times n$ unitary matrix. Then the optimal preconditioner of $T_n(f)$ is defined as follows:*

$$\mathcal{P}_U(T_n(f)) = U \cdot diag(U^\star T_n(f) U) \cdot U^\star$$

Proof. Setting $\mathfrak{A}_U := \{U\Delta U^\star |\ \Delta \text{ is diagonal matrix}\}$ and using the usual Frobenius norm defined by

$$\|M\|_F^2 = \sum_{k,j=1}^{n} |m_{kj}|^2$$

one has

$$\mathcal{P}_U(T_n(f)) = \arg\min_{B \in \mathfrak{A}_U} \|T_n(f) - B\|_F$$

or

$$\begin{aligned}
\|T_n(f) - B\|_F^2 &= \|T_n(f) - U\Delta U^\star\|_F^2 \\
&= \|U^\star(T_n(f) - U\Delta U^\star)U\|_F^2 \\
&= \|U^\star T_n(f) U - \Delta\|_F^2 \\
&= \sum_{k\neq j} |(U^\star T_n(f)U)_{kj}|^2 + \sum_{j=1}^{n} |(U^\star T_n(f)U)_{jj} - \Delta_{jj}|^2,
\end{aligned}$$

the optimum is obtained for $\Delta = diag(U^\star T_n(f)U)$.
Then

$$\mathcal{P}_U(T_n(f)) = U.diag(U^\star T_n(f)U).U^\star$$

□

Lemma 1.3.6. *Let $f \in L^1(-\pi,\pi)$ be a real-valued function and $T_n(f)$ be the Toeplitz matrix generated by f. Let U be a unitary matrix. If $f > 0$, then the optimal preconditioner $P_U(T_n(f))$ of $T_n(f)$ is nonsingular and $\sup_n \|P_U(T_n(f))^{-1}\|_2 \leq c < \infty$, where c is a constant number independent of n.*

Proof. According to Lemma 1.3.5, $\lambda_j(\mathcal{P}_U(T_n(f))) = (U^\star T_n(f)U)_{jj} = U_j^\star T_n(f) U_j$, (where U_j is the j-th column of U, $\|U_j\|_2 = 1$). One deduces from Theorem 1.3.1 that $\lambda_{\min}(T_n(f)) \leq \lambda_j(\mathcal{P}_U(T_n(f))) \leq \lambda_{\max}(T_n(f))$. Using Theorem 1.3.5, it follows that $m_f \leq \lambda_k(T_n(f)) \leq M_f$, for all $k \leq n; k, n \in \mathbb{N}$. So,

$$m_f \leq \lambda_{\min}(T_n(f)) \leq \lambda_j(\mathcal{P}_U(T_n(f))) \leq \lambda_{\max}(T_n(f)) \leq M_f$$

and because $f > 0$ over the compact set $[-\pi,\pi]$, it follows that $m_f > 0$. Hence, $P_U(T_n(f))$ is nonsingular and $\sup_n \|P_U(T_n(f))^{-1}\|_2 \leq c < \infty$. □

Conclusion

In this chapter, we have furnished a detailed study of the spectral properties of circulants and Toeplitz matrices. We will prove in Chapter 2 that suitably chosen sequences of circulant matrices asymptotically approximate sequences of Toeplitz matrices. Further, we will use the Korovkin Theory to derive clustering of the related preconditioned sequence when the preconditioners are selected in trigonometric algebras such as Circulants, Tau, etc... This new theory will mainly be based on the approximation of the Toeplitz sequences.

Chapter Two

Preconditioning Toeplitz Sequences via Circulants

2.1 Introduction

The purpose of this chapter is to present a detailed study of a Korovkin-type theory for finite Toeplitz operators via matrix algebra. We consider the approximation of finite self-adjoint Toeplitz operators $T_n(\cdot)$ by means of matrix algebra operators. Here the Hermitian Toeplitz matrix $T_n(f)$ is generated by a Lebesgue-integrable real-valued function f defined in $[-\pi, \pi]$ in the sense that the entries of $T_n(f)$ along the k-th diagonal are given by the k-th Fourier coefficients a_k of f. Denoting by $\mathcal{A} = \{\mathcal{M}(U_n) = \{A = U_n \Delta U_n^\star : \Delta \text{ diagonal}\}\}_n$ the sequence of the matrix algebra associated with the unitary transformations $\{U_n\}_n$ and by $U_n^\star$ the complex conjugate of U_n, we introduce the sequence of operators $\mathcal{P} = \{\mathcal{P}_{U_n} : \mathbb{C}^n \to \mathbb{C}^n\}_n$ so that each operator $\mathcal{P}_{U_n}$ associates with any $n \times n$ matrix A the matrix $\hat{X}$ that minimizes the functional $\mathcal{F}_A(X) = \|A - X\|_F$ in the Frobenius norm over the whole algebra $\mathcal{M}(U_n)$: in this way $\mathcal{P}_{U_n}(A)$ is the matrix where the above defined functional $\mathcal{F}_A(X)$, for $A = T_n(f)$, attains its minimum value.

First, we study the asymptotic equivalence of the matrices sequences $\{C_n(f)\}_n$ and $\{T_n(f)\}_n$ in the case where the generating function is chosen with some particular properties. Furthermore, we derive some general properties of the operator $\mathcal{P}_{U_n}$, starting from above considerations. Then we go on to consider the subset $\mathcal{A}_T$ of all sequences $\{\mathcal{M}(U_n)\}_n$ of algebras whose matrix-sequences $\{U_n\}_n$ are related to trigonometric functions. More precisely, we focus our attention on the algebra $\mathcal{M}(U_n)$ for which the adjoint of U_n, that i.e., $U_n^\star$, is an $n \times n$ Vandermonde-like matrix [67] whose functions are linearly independent and belong to the space of the trigonometric polynomials $\mathcal{P}_{n,2\pi}(I)$ evaluated on a quasi-equispaced set of points (the grid points) in a suitable interval I.

For $\{\mathcal{M}(U_n)\}_n \in \mathcal{A}_T$, we will give two Weierstrass-Jackson type theorems which assure the "approximation" of $\{T_n(f)\}_n$ by $\{\mathcal{P}_{U_n}(T_n(f))\}_n$ under simple conditions on the matrices $\{\mathcal{P}_{U_n}(T_n(p))\}_n$, p ranging among the trigonometric polynomials. In order words, this means that the approximation of the Toeplitz matrices generated by polynomials guarantees that $\{\mathcal{P}_{U_n}(T_n(f))\}_n$ is an approximation process for $\{T_n(f)\}_n$ with f just continuous. It worth mentioning that these two theorems cover, as special cases, the Theorems of R. Chan, Yeung [43, 44], Jin [85] and Serra [122], respectively, regarding the circulant [52], Hartley [12] and τ [10] matrix algebras. This research line is very important from a practical viewpoint since preconditioners belonging to a new algebra (not necessary circulant) can be more suitable for a certain class of applications: refer to [34] for the use of a "cosine" algebra in image restoration, to [36] for the use of circulants in the preconditioning of elliptic boundary value problems ($BVPs$) with periodic boundary conditions and to [62] for the application of the τ-class preconditioning when elliptic BVPs with Dirichlet conditions are considered.

Pursuing further, we notice that the eigenvalues of $\mathcal{P}_{U_n}(T_n(f))$ can be viewed as the values of a trigonometric polynomial, which we denote as $L_n[U_n](f)$, taken on the grid points of the interval I. We show that $f \to L_n[U_n](f)$ is a linear positive operator (LPO) [90] so that

its convergence to f is only dependent on the convergence at the functions 1, $\sin x$, $\cos x$ (or other equivalent test functions) as stated in the famous Korovkin Theorem [90]. In view of this, by using the known trigonometric algebras, we derive both known and new linear operators uniformly converging to the identity operator in the space of the continuous functions equipped with the infinity norm. In particular, the circulant class leads to the classical Cesaro sum, the τ class to a mixed process involving a Cesaro sum and a correction in terms of Chebyshev polynomials of second kind and so on (see also [54]).

Keenly aware of the mathematical elegance and power of the Korovkin Theorem, the Frobenius operator $\mathcal{P}_{U_n}(\cdot)$ is itself a linear positive operator (in the matrix sense) acting on the space of the $n \times n$ complex valued matrices. By exploiting this and other properties of $\mathcal{P}_{U_n}(\cdot)$ and $L_n[U_n](\cdot)$, two matrix-versions of the Korovkin Theorem are obtained: The continuous functions f are replaced by the Toeplitz sequences $\{T_n(f)\}_n$ with continuous f, the polynomials p are replaced by $\{T_n(p)\}_n$ and the approximation process is given by $\{\mathcal{P}_{U_n}(T_n(\cdot))\}_n$ that is, by the sequence of Frobenius-optimal representatives of the sequences $\{\mathcal{P}_{U_n}(T_n(p_i))\}_n$ converges "in a strong or weak sense" to $\{T_n(p_i)\}_n$, where the symbols p_i denote the usual three test functions, then a similar convergence holds for $\{\mathcal{P}_{U_n}(T_n(f))\}_n$ to $\{T_n(f)\}_n$ with f merely continuous. More specifically in section 2.6 we prove the following results.

Theorem 2.1.1. *Let f be a continuous periodic function. If $L_n[U_n](p) = p + \epsilon_n(p)$ for each one of the three test functions p and with $\epsilon_n(p)$ going uniformly to zero, then $\{\mathcal{P}_{U_n}(T_n(f))\}_n$ converges to $\{T_n(f)\}_n$ in the weak sense.*

Theorem 2.1.2. *Under the same assumption of the previous theorem, if $\epsilon_n(p) = O(n^{-1})$ for the three test functions p, and if the grid points of the algebra are uniformly distributed, then the convergence of $\{\mathcal{P}_{U_n}(T_n(f))\}_n$ and $\{T_n(f)\}_n$ is strong.*

Besides the mathematical interest in itself, the above results have dramatic practical implications. In fact, they provide a simple and powerful tool for analyzing the eigenvalue distribution and clustering [43, 122] of the preconditioned matrices and to study the superlinear (or sublinear) convergence of the preconditioned conjugate gradient (PCD) method [6] applied to systems of the form $T_n(f)\mathbf{x} = \mathbf{b}$. See also [77, 76, 70, 42] for several specific applications of these linear systems.

2.2 Asymptotic equivalence of the matrix sequences $\{T_n(f)\}_n$ and $\{C_n(f)\}_n$.

In this section, we consider the case where the generating function f is in the Wiener class, i.e., the case where the sequence $\{t_k\}_k$ of the Fourier coefficients of the Toeplitz matrix $T_n(f)$ is absolutely summable. The basic approach is to find a sequence of circulant matrices $\{C_n(f)\}_n$ that is asymptotically equivalent to the sequence of Toeplitz matrices $\{T_n(f)\}_n$. Obviously, the choice of an appropriate sequence of circulant matrices to approximate a sequence of Toeplitz matrices is not unique, so we are free to choose a construction with the most desirable properties. It will, in fact, prove useful to consider two slightly different circulant approximations.

Definition 2.2.1. *Let $\{A_n\}_n$ and $\{B_n\}_n$ be two matrix sequences of order $n \times n$. $\{A_n\}_n$ and $\{B_n\}_n$ are said to be asymptotically equivalent if*

(i) *$\{A_n\}_n$ and $\{B_n\}_n$ are uniformly bounded in strong norm, i.e.,*

$$\sup_n \|A_n\|_2, \ \sup_n \|B_n\|_2 \leq M < \infty,$$

where M is a constant number independent of n.

(ii) $\|A_n - B_n\|_F = o(\sqrt{n})$.

Notation: *$\{A_n\}_n \sim \{B_n\}_n$ means that the sequences $\{A_n\}_n$ and $\{B_n\}_n$ are asymptotically equivalent.*

Proposition 2.2.1. [75] *Let $T_n(f) = [t_{k-j}]_{k,j=0}^{n-1}$ be the Toeplitz matrix generated by a Wiener function $f \in L^1(-\pi, \pi)$, i.e.,*

$$\sum_{k=-\infty}^{\infty} |t_k| < \infty,$$

and where

$$f(x) = \sum_{k=-\infty}^{\infty} t_k e^{\hat{i}kx}, \quad \hat{f}_n(x) = \sum_{k=-(n-1)}^{n-1} t_k e^{\hat{i}kx}.$$

Define the circulant matrices $C_n(f)$ and $C_n(\hat{f}_n)$ as follows:

1. *$C_n(f)$ is the circulant matrix with top row $(c_0^{(n)}, c_1^{(n)}, ..., c_{n-1}^{(n)})$ where*

$$c_k^{(n)} = \frac{1}{n}\sum_{j=0}^{n-1} f(2\pi j/n) e^{2\pi \hat{i} jk/n}$$

2. *$C_n(\hat{f}_n)$ is the circulant matrix with top row $(\hat{c}_0^{(n)}, \hat{c}_1^{(n)}, ..., \hat{c}_{n-1}^{(n)})$ where*

$$\begin{aligned} \hat{c}_k^{(n)} &= \frac{1}{n}\sum_{j=0}^{n-1} \hat{f}_n(2\pi j/n) e^{2\pi \hat{i} jk/n} \\ &= \sum_{l=-(n-1)}^{n-1} t_l \delta_{(k+l) mod\, n} \\ &= \begin{cases} t_0 & if\ k = 0 \\ t_{-k} + t_{n-k} & for\ k = 1, 2, ..., n-1. \end{cases} \end{aligned}$$

Then,

$$\{C_n(f)\}_n \sim \{C_n(\hat{f}_n)\}_n \sim \{T_n(f)\}_n. \tag{2.1}$$

Proof. Since $C_n(f)$ and $C_n(\hat{f}_n)$ are circulant matrices with the same eigenvectors (cf. Theorem 1.2.1), we have from part 2 of Theorem 1.2.1 and from $\|A\|_F = \left(\sum_{k=0}^{n-1} \lambda_k(A^\star A)\right)^{1/2}$ that

$$\|C_n(f) - C_n(\hat{f}_n)\|_F^2 = \sum_{k=0}^{n-1} |f(2\pi k/n) - \hat{f}_n(2\pi k/n)|^2.$$

Let $\epsilon > 0$, since $f(x) = \sum_{k=-\infty}^{\infty} t_k e^{\hat{i}kx}$, then there exists $N_\epsilon \in \mathbb{N}$ such that for $n \geq N_\epsilon$,

$$\left| f(x) - \sum_{k=-n}^{n} t_k e^{\hat{i}kx} \right| < \epsilon, \text{ for all } x \in [-\pi, \pi]. \tag{2.2}$$

It follows from (2.2) that $\{\hat{f}_n\}_n$ uniformly converges to f. So, for given $\epsilon > 0$, there is an integer $N_\epsilon \in \mathbb{N}$ such that for $n \geq N_\epsilon$, one has

$$|f(2\pi k/n) - \hat{f}_n(2\pi k/n)|^2 < \epsilon, \text{ for } k \leq n-1$$

and hence, for $n \geq N_\epsilon$

$$\|C_n(f) - C_n(\hat{f}_n)\|_F^2 < \sum_{k=0}^{n-1} \epsilon = n\epsilon.$$

Since $\epsilon > 0$ is arbitrary, one has

$$\|C_n(f) - C_n(\hat{f}_n)\|_F = o(\sqrt{n})$$

proving that

$$\{C_n(f)\}_n \sim \{C_n(\hat{f}_n)\}_n. \tag{2.3}$$

Because $C_n(\hat{f}_n)$ is also a Toeplitz matrix, define $C_n(\hat{f}_n) = T_n^{'} = [t_{k-j}^{'}]_{k,j=0}^{n-1}$ with

$$t_k^{'} = \begin{cases} \hat{c}_{-k}^{(n)} = t_k + t_{n+k} & \text{if } k = -(n-1), -(n-2), ..., -1 \\ \hat{c}_0^{(n)} = t_0 & \text{if } k = 0 \\ \hat{c}_{n-k}^{(n)} = t_{-(n-k)} + t_k & \text{if } k = 1, 2, ..., n-1 \end{cases} \tag{2.4}$$

then

$$\|T_n^{'}\|_F = \sum_{k=0}^{n-1}\sum_{j=0}^{n-1} |t_{k-j}^{'}|^2 = \sum_{k=-(n-1)}^{n-1} (n - |k|)|t_k|^2. \tag{2.5}$$

According to (2.4) and (2.5), one obtains

$$\begin{aligned}
\|T_n(f) - C_n(\hat{f}_n)\|_F^2 &= \sum_{k=-(n-1)}^{n-1} (n-|k|)|t_k - t_k^{'}|^2 \\
&= \sum_{k=-(n-1)}^{-1} (n+k)|t_{n+k}|^2 + \sum_{k=1}^{n-1}(n-k)|t_{-(n-k)}|^2 \\
&= n\left(\sum_{k=-(n-1)}^{-1} |t_{n+k}|^2 + \sum_{k=1}^{n-1}|t_{-(n-k)}|^2\right) + \sum_{k=-(n-1)}^{-1} k|t_{n+k}|^2 - \sum_{k=1}^{n-1} k|t_{-(n-k)}|^2 \\
&= n\left(\sum_{k=1}^{n-1}|t_k|^2 + \sum_{k=1}^{n-1}|t_{-(n-k)}|^2\right) + \sum_{k=1}^{n-1}(k-n)|t_k|^2 - \sum_{k=1}^{n-1} k|t_{-(n-k)}|^2 \\
&= n\sum_{k=1}^{n-1}|t_{-(n-k)}|^2 + \sum_{k=1}^{n-1} k|t_k|^2 - \sum_{k=1}^{n-1} k|t_{-(n-k)}|^2 \\
&= n\sum_{k=1}^{n-1}|t_{-k}|^2 + \sum_{k=1}^{n-1} k|t_k|^2 + \sum_{k=1}^{n-1}(k-n)|t_{-k}|^2 \\
&= \sum_{k=1}^{n-1} k|t_k|^2 + \sum_{k=1}^{n-1} k|t_{-k}|^2 = \sum_{k=1}^{n-1} k(|t_k|^2 + |t_{-k}|^2).
\end{aligned}$$

then

$$\|T_n(f) - C_n(\hat{f}_n)\|_F^2 = \sum_{k=1}^{n-1} k(|t_k|^2 + |t_{-k}|^2). \tag{2.6}$$

On the other side, one has

$$\sum_{k=-\infty}^{\infty} |t_k|^2 \leq M \sum_{k=-\infty}^{\infty} |t_k|$$

where M is a positive constant independent of k. **Indeed:** Since the series $\sum_{k=-\infty}^{\infty} |t_k|$ converges, then the sequence $\{t_k\}_k$ is uniformly bounded, so there exists a positive constant M_0 independent of k such that $\sup_k |t_k| \leq M_0$. Further, $\lim_{k\to\pm\infty} |t_k| = 0$, then there exists a positive integer N such that for $|k| > N$, $|t_k| < 1$, whence, for $|k| > N$, $|t_k|^2 < |t_k|$. Let $k \in \mathbb{Z}$,

$$\begin{aligned}
\sum_{k=-\infty}^{\infty} |t_k|^2 &= \sum_{k=-N}^{N} |t_k|^2 + \sum_{k=-\infty}^{-(N+1)} |t_k|^2 + \sum_{k=N+1}^{\infty} |t_k|^2 \\
&\leq M_0 \sum_{k=-N}^{N} |t_k| + \sum_{k=-\infty}^{-(N+1)} |t_k| + \sum_{k=N+1}^{\infty} |t_k| \\
&\leq (1 + M_0) \sum_{k=-\infty}^{\infty} |t_k|
\end{aligned}$$

For $M = 1 + M_0$, one has the result. So, the sequence $\{|t_k|^2\}_k$ is summable. Hence, for given $\epsilon > 0$, there exists an integer $N_\epsilon \in \mathbb{N}$ sufficiently large so that

$$\sum_{k=N_\epsilon}^{\infty} (|t_k|^2 + |t_{-k}|^2) \leq \epsilon.$$

Then, for $n > N_\epsilon + 1$

$$\begin{aligned}
\sum_{k=1}^{n-1} k(|t_k|^2 + |t_{-k}|^2) &= \sum_{k=1}^{N_\epsilon - 1} k(|t_k|^2 + |t_{-k}|^2) + \sum_{k=N_\epsilon}^{n-1} k(|t_k|^2 + |t_{-k}|^2) \\
&\leq \sum_{k=1}^{N_\epsilon - 1} k(|t_k|^2 + |t_{-k}|^2) + (n-1) \sum_{k=N_\epsilon}^{n-1} (|t_k|^2 + |t_{-k}|^2) \\
&\leq \sum_{k=1}^{N_\epsilon - 1} k(|t_k|^2 + |t_{-k}|^2) + (n-1)\epsilon.
\end{aligned}$$

It follows from this inequality and (2.6) that

$$\lim_{n\to\infty} \frac{1}{n} \|T_n(f) - C_n(\hat{f}_n)\|_F^2 \leq \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{N_\epsilon - 1} k(|t_k|^2 + |t_{-k}|^2) + \lim_{n\to\infty} \frac{n-1}{n} \epsilon = \epsilon$$

and because $\epsilon > 0$ is arbitrary, one has

$$\|T_n(f) - C_n(\hat{f}_n)\|_F = o(\sqrt{n}).$$

Hence

$$\{T_n(f)\}_n \sim \{C_n(\hat{f}_n)\}_n. \tag{2.7}$$

One deduces from (2.3) and (2.7) that

$$\{C_n(f)\}_n \sim \{T_n(f)\}_n.$$

□

2.3 Matrix algebras and Frobenius-optimal approximation

Let U_n be a unitary complex $n \times n$ matrix, then by $\mathcal{M}(U_n)$ we denote the communicative algebra of all the matrices simultaneously diagonalized by the U_n transform, that is,

$$\mathcal{M}(U_n) = \{A = U_n \Delta U_n^\star : \Delta \text{ diagonal}\}.$$

Here the symbol $\star$ means transpose and conjugate. The operator $\mathcal{P}_{U_n}(\cdot)$ is defined on $\mathbb{C}^{n\times n}$ and takes values in $\mathcal{M}(U_n)$ where both the spaces are equipped with the Frobenius norm $\|X\|_F^2 = \sum_{i,j=0}^{n-1} |x_{i,j}|^2$. In addition, the Frobenius norm is induced by the positive scalar product $(\cdot,\cdot)_F$ on $\mathbb{C}^{n\times n}$ defined as $(A,B)_F = trace(A^\star \cdot B)$. Therefore the existence and the uniqueness of the minimum

$$\mathcal{P}_{U_n}(A) = \arg\min_{X\in\mathcal{M}(U_n)} \|A - X\|_F$$

follows from the fact that the space $(\mathbb{C}^{n\times n}, (\cdot,\cdot)_F)$ is a Hilbert space and $\mathcal{M}(U_n)$ is a closed convex subset since it is a finite dimensional vector space.

By means of simple algebraic arguments, we prove the following Lemma (see also [41] for the circulant case).

Lemma 2.3.1. [55]. *With $A, B \in \mathbb{C}^{n\times n}$ and the previous definition of $\mathcal{P}_{U_n}(\cdot)$, we have*

1. $\mathcal{P}_{U_n}(A) = U_n \sigma(U_n^\star A U_n) U_n^\star$, *with $\sigma(X)$ being the diagonal matrix having $(X)_{ii}$ as diagonal elements,*
2. $\mathcal{P}_{U_n}(\alpha A + \beta B) = \alpha \mathcal{P}_{U_n}(A) + \beta \mathcal{P}_{U_n}(B)$ *with $\alpha, \beta \in \mathbb{C}$,*
3. $\mathcal{P}_{U_n}(A^\star) = (\mathcal{P}_{U_n}(A))^\star$,
4. $trace(\mathcal{P}_{U_n}(A)) = trace(A)$,
5. $\|\mathcal{P}_{U_n}\|_2 = 1$,
6. $\|\mathcal{P}_{U_n}\|_F = 1$,
7. $\|A - \mathcal{P}_{U_n}(A)\|_F^2 = \|A\|_F^2 - \|\mathcal{P}_{U_n}(A)\|_F^2$.

Proof. **1.** The proof of this point is done in Lemma 1.3.5 by replacing A and U_n by $T_n(f)$ and U respectively.

2. Let $A, B \in \mathbb{C}^{n\times n}$ and $\alpha, \beta \in \mathbb{C}$.

$$\begin{aligned}\mathcal{P}_{U_n}(\alpha A+\beta B) &= U_n\sigma\left(U_n^\star(\alpha A+\beta B)U_n\right)U_n^\star\\ &= U_n\sigma\left(\alpha U_n^\star A U_n+\beta U_n^\star B U_n\right)U_n^\star\\ &= \alpha U_n\sigma\left(U_n^\star A U_n\right)U_n^\star+\beta U_n\sigma\left(U_n^\star B U_n\right)U_n^\star\\ &= \alpha\mathcal{P}_{U_n}(A)+\beta\mathcal{P}_{U_n}(B).\end{aligned}$$

3.

$$\begin{aligned}\mathcal{P}_{U_n}(A^\star)=U_n\sigma(U_n^\star A^\star U_n)U_n^\star &= U_n\sigma\left((U_n^\star A U_n)^\star\right)U_n^\star\\ &= U_n\left(\sigma(U_n^\star A U_n)\right)^\star U_n^\star \text{ since } \sigma(X^\star)=(\sigma(X))^\star\\ &= (U_n\sigma(U_n^\star A U_n)U_n^\star)^\star\\ &= \mathcal{P}_{U_n}(A))^\star.\end{aligned}$$

4. $trace(\mathcal{P}_{U_n}(A)) = trace(\sigma(U_n^\star A U_n)) = trace(U_n^\star A U_n) = trace(A)$.

5. First of all, $\max\limits_{1\le i\le n}|(U_n^\star A U_n)_{ii}| \le \|U_n^\star A U_n\|_2$. **Indeed:** Let $x_0=(0,...,0,1,0...,0)^T\in\mathbb{C}^n$.

$$\begin{aligned}x_0^\star(U_n^\star A U_n)^\star(U_n^\star A U_n)x_0 &= (0,...,0,\overline{(U_n^\star A U_n)}_{ii},0,...,0)(0,...,0,(U_n^\star A U_n)_{ii},0,...,0)^T\\ &= |(U_n^\star A U_n)_{ii}|^2,\end{aligned}$$

then $\|(U_n^\star A U_n)x_0\|_2 = |(U_n^\star A U_n)_{ii}|$, so

$$|(U_n^\star A U_n)_{ii}| \le \sup_{\|x\|_2=1}\|(U_n^\star A U_n)x\|_2 = \|U_n^\star A U_n\|_2 = \|A\|_2.$$

Now,

$$\|\mathcal{P}_{U_n}\|_2 = \sup_{A\neq 0}\frac{\|\mathcal{P}_{U_n}(A)\|_2}{\|A\|_2} = \sup_{A\neq 0}\frac{\|U_n\sigma(U_n^\star A U_n)U_n^\star\|_2}{\|A\|_2} = \sup_{A\neq 0}\frac{\|\sigma(U_n^\star A U_n)\|_2}{\|A\|_2}$$

$$= \sup_{A\neq 0}\frac{\max\limits_{1\le i\le n}|(U_n^\star A U_n)_{ii}|}{\|A\|_2} \le \sup_{A\neq 0}\frac{\|U_n^\star A U_n\|_2}{\|A\|_2} = \sup_{A\neq 0}\frac{\|A\|_2}{\|A\|_2} = 1,$$

so $\|\mathcal{P}_{U_n}\|_2 \le 1$.

On the other side,

$$\|\mathcal{P}_{U_n}\|_2 \ge \frac{\|\mathcal{P}_{U_n}(I_n)\|_2}{\|I_n\|_2} = \frac{\|U_n\sigma(U_n^\star I_n U_n)U_n^\star\|_2}{\|I_n\|_2} = \frac{\|\sigma(I_n)\|_2}{\|I_n\|_2} = 1.$$

Whence, $\|\mathcal{P}_{U_n}\|_2 = 1$.

6. The proof of this point is similar to the proof of point (5) by replacing $\|\cdot\|_2$ by $\|\cdot\|_F$.

7.

$$\begin{aligned}\|A-\mathcal{P}_{U_n}(A)\|_F^2 &= \|U_n^\star A U_n - U_n^\star\mathcal{P}_{U_n}(A)U_n\|_F^2\\ &= \|U_n^\star A U_n - \sigma(U_n^\star A U_n)\|_F^2\\ &= \sum_{i\neq j}|(U_n^\star A U_n)_{ij}|^2\\ &= \sum_{i,j=0}^{n-1}|(U_n^\star A U_n)_{ij}|^2 - \sum_{j=0}^{n-1}|(U_n^\star A U_n)_{jj}|^2\\ &= \|U_n^\star A U_n\|_F^2 - \|\sigma(U_n^\star A U_n)\|_F^2\\ &= \|A\|_F^2 - \|U_n\sigma(U_n^\star A U_n)U_n^\star\|_F^2\\ &= \|A\|_F^2 - \|\mathcal{P}_{U_n}(A)\|_F^2.\end{aligned}$$

□

Also of interest is the following result which was proved and used for specific matrix algebras [41, 43, 85, 122, 163], but extended in an abstract way in [55].

Lemma 2.3.2. [55]. *If A is a Hermitian matrix ($A = A^\star$), then the eigenvalues of $\mathcal{P}_{U_n}(A)$ are contained in the closed real interval $[\lambda_1(A), \lambda_n(A)]$ where $\lambda_j(A)$ are the eigenvalues of A ordered in a nondecreasing way. Moreover, when A is positive definite, $\mathcal{P}_{U_n}(A)$ is positive definite as well.*

Proof. According to Lemma 2.3.1, $\lambda_j(\mathcal{P}_{U_n}(A)) = (U_n^\star A U_n)_{jj} = (U_n)_j^\star A (U_n)_j$, (where $(U_n)_j$ is the j-th column of U_n, $\|(U_n)_j\|_2 = 1$). Since A is Hermitian, one deduces from Theorem 1.3.1 that $\lambda_1(A) \leq \lambda_j(\mathcal{P}_U(A)) \leq \lambda_n(A)$ for all $j = 1, 2, ..., n$. □

Trigonometric matrix algebras

Here we define a special subset of sequences of matrix algebras that we call trigonometric matrix algebras and we denote it by $\mathcal{A}_T$. Let $\{v_n\}_{n\in\mathbb{N}}$ with $v_n = \{v_{nj}\}_{j=0}^{n-1}$, be a sequence of trigonometric functions on an interval I. Let $S = \{S_n\}_{n\in\mathbb{N}}$ be a sequence of grids of n points on I, namely, $S_n = \{x_i^{(n)}, i = 0, 1, 2, ..., n-1\}$.

Let us suppose that the generalized Vandermonde matrix

$$V_n = \left(v_{nj}(x_i^{(n)})\right)_{i,j=0}^{n-1}$$

is a unitary matrix. Then, an algebra of the form $\mathcal{M}(U_n)$ is a trigonometric algebra if $U_n = V_n^\star$ with V_n a generalized trigonometric Vandermonde matrix. In addition, given a sequence of unitary generalized trigonometric Vandermonde matrices $\{U_n = V_n^\star\}_n$ and the related sequence of algebras $\{\mathcal{M}(U_n)\}_n$ belonging to $\mathcal{A}_T$, we will call it regular if the grid points form a sequence of quasi-uniformly distributed grid sequences in I. For a formal definition of quasi-uniform distribution, see the following.

Definition 2.3.1. *A sequence of grids $\{S_n = \{x_i^{(n)}, i = 0, 1, 2, ..., n-1\}\}_n$ belonging to an interval I is called quasi-uniform if*

$$\sum_{i=1}^{n-1} \left| \frac{|I|}{n} - (x_i^{(n)} - x_{i-1}^{(n)}) \right| = o(1), \quad \text{for } n \to \infty \tag{2.8}$$

with $|I|$ being the width of I. If the previous relation holds for $o(1) = O(n^{-1})$, then the mesh-sequence $\{S_n\}_n$ is called uniform.

Examples of trigonometric algebras are the circulant, the τ, the Hartley [12] for which the matrix U_n is

$$\begin{aligned} U_n = F_n &= \left(\frac{1}{\sqrt{n}} e^{\hat{i} j x_i^{(n)}}\right), \quad i, j = 0, 1, ..., n-1 \\ S_n &= \left\{x_i^{(n)} = \frac{2i\pi}{n} : i = 0, 1, ..., n-1\right\} \subset I = [0, 2\pi], \end{aligned}$$

$$\begin{aligned} U_n = S_n &= \left(\sqrt{\frac{2}{n+1}} \sin((j+1)x_i^{(n)})\right), \quad i, j = 0, 1, ..., n-1 \\ S_n &= \left\{x_i^{(n)} = \frac{(i+1)\pi}{n+1} : i = 0, 1, ..., n-1\right\} \subset I = [0, \pi], \end{aligned}$$

$$U_n = H_n = \left(\frac{1}{\sqrt{n}} [\sin(jx_i^{(n)}) + \cos(jx_i^{(n)})] \right), \quad i,j = 0,1,...,n-1$$
$$S_n = \left\{ x_i^{(n)} = \frac{2i\pi}{n} : i = 0,1,...,n-1 \right\} \subset I = [0,2\pi].$$

respectively.

Notice that all the associated sequences of algebras are trigonometric and regular with uniform meshes. In addition, these algebras are characterized by a special kind of sequence $\mathbf{v}$ of functions. In particular if $n > m$, then there exists a constant $C(n,m)$ such that

$$(v_{ni})_{[i=0,m-1]} = C(n,m)v_m.$$

This means that, up to a scaling factor, for any n, the functions related to v_n can be viewed as the first n functions of a unique sequence $\{v_j\}_{j\in\mathbb{N}}$.

2.4 A Weierstrass-Jackson matrix theory

In order to properly state the "matrix approximation results", we require the following definitions concerning the concept of "matrix convergence".

Definition 2.4.1. *Given a sequence of algebras $\{\mathcal{M}(U_n)\}_n$ with associated operators $\{\mathcal{P}_{U_n}(\cdot)\}_n$, we say that "$\{\mathcal{P}_{U_n}(T_n(f))\}_n$ (strongly) converges to $\{T_n(f)\}_n$" if, for any $\epsilon > 0$, there exists a nonnegative integer N_ϵ such that, for $n \geq N_\epsilon$, $T_n(f) - \mathcal{P}_{U_n}(T_n(f))$ has eigenvalues in $(-\epsilon, \epsilon)$ except for $N_\epsilon = o(1)$ outliers (proper clustering at zero* [164]*).*

Definition 2.4.2. *Given a sequence of algebras $\{\mathcal{M}(U_n)\}_n$ with associated operators $\{\mathcal{P}_{U_n}(\cdot)\}_n$, we say that "$\{\mathcal{P}_{U_n}(T_n(f))\}_n$ (weakly) converges to $\{T_n(f)\}_n$" if, for any $\epsilon > 0$, there exists a nonnegative integer N_ϵ such that, for $n \geq N_\epsilon$, $T_n(f) - \mathcal{P}_{U_n}(T_n(f))$ has eigenvalues in $(-\epsilon, \epsilon)$ except for $N_\epsilon = o(n)$ outliers (general clustering at zero* [164]*).*

We say that the convergence is also uniform, when the number N_ϵ does not depend on ϵ. In the case where there is strong convergence (strong or proper clustering in the terminology used in [164]) and the function f is strictly positive, we have a superlinear convergence of the related PCG methods having $\{\mathcal{P}_{U_n}(T_n(f))\}_n$ as preconditioner, but we may have a sublinear behavior when the weak convergence (weak or general clustering [164]) case occurs [151, 52].

Moreover, if the convergence is also uniform, that is, N_ϵ does not depend on ϵ, the number of iterations decreases as the dimension n increases and, therefore, the associated PCG method is comparable with the one devised in [119].

The following result due to Tyrtyshnikov provides a criterion to establish if convergence occurs.

Lemma 2.4.1. [164] *Let $\{A_n\}_n$ and $\{B_n\}_n$ be two sequences of $n \times n$ Hermitian matrices. When $\|A_n - B_n\|_F^2 = O(1)$, then we have convergence in the strong sense. If $\|A_n - B_n\|_F^2 = o(n)$, then the convergence is weak.*

Theorem 2.4.1. *Let f be a continuous periodic real-valued function. Then, $\{\mathcal{P}_{U_n}(T_n(f))\}_n$ strongly converges to $\{T_n(f)\}_n$ if $\{\mathcal{P}_{U_n}(T_n(p))\}_n$ strongly converges to $\{T_n(p)\}_n$ for all the trigonometric polynomials p.*

Proof. Let p_k be the polynomial having degree k of best approximation of f in the supremum [84]. For any $\epsilon > 0$, there exists an integer M such that $\|f - p_M\|_\infty < \epsilon/3$. Then, by using the Szegö Theorem (see [77] at page 64) and Lemma 2.3.2 we have $\|T_n(f) - T_n(p_M)\|_2 < \epsilon/3$, $\|\mathcal{P}_{U_n}(T_n(f)) - \mathcal{P}_{U_n}(T_n(p_M))\|_2 < \epsilon/3$. Therefore, from the identity

$$T_n(f) - \mathcal{P}_{U_n}(T_n(f)) = T_n(f) - T_n(p_M) - \mathcal{P}_{U_n}(T_n(f)) + \mathcal{P}_{U_n}(T_n(p_M)) + T_n(p_M) - \mathcal{P}_{U_n}(T_n(p_M))$$

we have that, except for a term of norm bounded by $2\epsilon/3$, the difference $T_n(f) - \mathcal{P}_{U_n}(T_n(f))$ coincides with $T_n(p_M) - \mathcal{P}_{U_n}(T_n(p_M))$. From the hypothesis of convergence, we may split the Hermitian matrix $T_n(p_M) - \mathcal{P}_{U_n}(T_n(p_M))$ into two parts. The first part has a norm bounded by $\epsilon/3$ and the second part has constant rank. Therefore, the claimed result is obtained, by invoking the Cauchy interlace theorem (cf. Theorem 1.3.2). □

Theorem 2.4.2. *Let f be a continuous periodic real-valued function. Then, $\{\mathcal{P}_{U_n}(T_n(f))\}_n$ weakly converges to $\{T_n(f)\}_n$ if $\{\mathcal{P}_{U_n}(T_n(p))\}_n$ weakly converges to $\{T_n(p)\}_n$ for all the trigonometric polynomials p.*

Proof. The proof is the same as the one of Theorem 2.4.1 with the exception of the last part where we split $\{T_n(p_M) - \mathcal{P}_{U_n}(T_n(p_M))\}_n$ into two sequences: the first has a norm bounded by $\epsilon/3$ and the second one has $o(n)$ rank. The use of the Cauchy Interlace theorem completes the proof. □

The following corollaries are particularly useful for deriving and analyzing good preconditioners for the conjugate gradient method.

Corollary 2.4.1. *Under the assumption of Theorem 2.4.1, if f is positive then for any $\epsilon > 0$, for n large enough, the matrix $\mathcal{P}_{U_n}(T_n(f))^{-1}T_n(f)$ has eigenvalues in $(1-\epsilon, 1+\epsilon)$ except $N_\epsilon = O(1)$ outliers, at most.*

Corollary 2.4.2. *With the hypotheses of Theorem 2.4.2, if f is positive then for any $\epsilon > 0$, for n large enough, the matrix $\mathcal{P}_{U_n}(T_n(f))^{-1}T_n(f)$ has eigenvalues in $(1-\epsilon, 1+\epsilon)$ except $N_\epsilon = o(n)$ outliers, at most.*

The proofs of the first corollary and Theorem 2.4.1 for specific algebras can be found in [43], [85], [122]. However, in the cited papers, a bit different definition of strong convergence is used since the notion of convergence is replaced by the fact that $\{T_n(p_M) - \mathcal{P}_{U_n}(T_n(p_M))\}_n$ is viewed as the sum of a matrix with norm bounded by $\epsilon/3$ and another of constant rank. Therefore, we observe that the argument used in [43] proves in effect a more general and abstract formulation.

It is interesting to notice that the main effort made in the aforementioned papers was toward proving of the "matrix convergence" in the polynomial case. Therefore, in section 2.6, in Theorems 2.6.3 and 2.6.4 (cf. [123, 124, 134]) we have given very simple conditions for verifying the "matrix convergence" in the polynomial cases.

Finally, observe that the two corollaries tell us something about the convergence of the associated PCG methods [6], [43] : in particular, if the assumption of corollary 2.4.1 is fulfilled, then we have a superlinear PCG method.

2.5 The LPO sequences related to $\{\mathcal{P}_{U_n}(T_n(\cdot))\}_n$

The behavior of the eigenvalues of $\mathcal{P}_{U_n}(T_n(f))$ is studied in this section. If U_n is completely generic not very much can be said, but under the assumption that $\mathcal{M}(U_n)$ is a trigonometric matrix algebra, a richer analysis can be carried out. Actually, in the light of the concepts introduced in section 2.3, the j-th row of U_n is a vector of trigonometric functions calculated on the grid point $x_j^{(n)}$. Therefore, by exploiting the first part of Lemma 2.3.1, we find that the j-th eigenvalue λ_j of $\mathcal{P}_{U_n}(T_n(f))$ is $\sigma(U_n^\star T_n(f) U_n)_{jj}$. Consequently, λ_j is the value that a trigonometric function takes on $x = x_j^{(n)}$. If f is a real valued function, then λ_i is a Rayleigh quotient related to the Hermitian matrix $T_n(f)$ and is a real-valued polynomial.

Now, let us consider the function $[L_n[U_n](f)](x)$ obtained by replacing $x_j^{(n)}$ with $x \in I$ in the formal expression of $\lambda_j = \sigma(U_n^\star T_n(f) U_n)_{jj}$. To formulate precisely this idea, let us define

$\mathbb{C}_{2\pi}(I,\mathbb{R})$ as the space of the continuous real-valued 2π-periodic functions defined on I and let us define the sequence of operators $\{L_n[U_n](\cdot)\}_n$

$$L_n[U_n](\cdot) : \mathbb{C}_{2\pi}(I,\mathbb{R}) \to \mathbb{C}_{2\pi}(I,\mathbb{R})$$

as

$$L_n[U_n](f) = v(\cdot)T_n(f)v^\star(\cdot) \in \mathbb{C}_{2\pi}(I,\mathbb{R}).$$

Here f is a real valued function and $v(x)$ is the generic row of $U_n^\star$, where the grid points have been replaced by the continuous variable x. In other words, it can be seen that $L_n[U_n](f)$ is nothing other than the continuous expressions of the diagonal elements of $\sigma(U_n^\star T_n(f)U_n)$.

In order to analyze this sequence of operators, let us introduce the following definitions.

Definition 2.5.1. *Let $\mathcal{G}$ be a linear space of functions and Φ be an operator from $\mathcal{G}$ to $\mathcal{G}$. Let us suppose that*

1. $\Phi(\alpha f + \beta g) = \alpha\Phi(f) + \beta\Phi(g)$ *with* $f, g \in \mathcal{G}$ *and* $\alpha, \beta \in \mathbb{C}$;
2. $\Phi(f) \geq 0$ *for any nonnegative function* $f \in \mathcal{G}$.

Under the above mentioned assumptions, the operator Φ is said linear and positive operator (LPO).

Definition 2.5.2. *Let $\mathcal{G}$ be a linear space of matrices and Φ be an operator from $\mathcal{G}$ to $\mathcal{G}$. Let us suppose that*

1. $\Phi(\alpha A + \beta B) = \alpha\Phi(A) + \beta\Phi(B)$ *with* $A, B \in \mathcal{G}$ *and* $\alpha, \beta \in \mathbb{C}$;
2. $\Phi(A)$ *is self-adjoint and nonnegative definite whether the matrix* $A \in \mathcal{G}$ *is self-adjoint and nonnegative definite.*

Under the above mentioned assumptions, the operator Φ is said linear and positive (matrix) operator (LPO).

Under the former notations, the following lemma holds.

Lemma 2.5.1. *$L_n[U_n](\cdot)$ is a linear positive operator. In addition $\mathcal{P}_{U_n}(\cdot)$ is also a LPO in the matrix sense.*

Proof. The linearity follows from the linearity of $\mathcal{P}_{U_n}(\cdot)$ (see Lemma 2.3.1). Fix $f \geq 0$, then $T_n(f)$ is nonnegative definite (cf. Lemma 1.3.3) and therefore any Rayleigh quotient is nonnegative and particularly the one giving rise to $L_n[U_n](f)$.

The second part is straightforward if we notice that the nonnegative definiteness of $T_n(f)$ implies the nonnegativity of the diagonal entries of $U_n^\star T_n(f)U_n$ and therefore the nonnegativity of all the eigenvalues of the Hermitian matrix $\mathcal{P}_{U_n}(T_n(f))$ (see Lemma 2.3.2). □

Now we resort to the well-known Korovkin theorem to establish whether the eigenvalues of $\{\mathcal{P}_{U_n}(T_n(f))\}_n$ tend to $\{\{f(x_j^{(n)})\}_j\}_n$ for n going to infinity. Observe that this also implies that the spectra of $\mathcal{P}_{U_n}(T_n(f))$ and $T_n(f)$ are equally distributed in the sense of Weyl-Tyrtyshnikov [77], [164].

So the convergence of the j-th eigenvalue of the Frobenius-optimal matrix $\{\mathcal{P}_{U_n}(T_n(f))\}_n$ to $\{f(x_j^{(n)})\}_n$ would be trivially implied by the uniform convergence of $\{L_n[U_n](f)\}_n$ to f. On the other hand, this general result is implied by the convergence of $\{L_n[U_n](p_i)\}_n$ to p_i for three test functions as stated in the Korovkin Theorems.

We introduce the notion of Chebyshev set [102].

Definition 2.5.3. *A finite sequence of functions $\{b_i\}_{i=1}^m$ defined on a set I is a Chebyshev set if and only if for any choice of m distinct points $\{x_i\}_{i=1}^m$ of I, the associated $m \times m$ Vandermonde matrix $\{b_i(x_j)\}_{i,j=1}^m$ is nonsingular.*

Theorem 2.5.1. *[Korovkin [90]]. Let $\mathcal{G}$ be the linear space of the continuous (periodic) real valued functions on a suitable interval I and let $\{\Phi_n(\cdot)\}_n$ be a sequence of linear positive operators from $\mathcal{G}$ to $\mathcal{G}$. If $\{\Phi_n(p_i)\}_n$ uniformly converges to p_i, for $i = 1, 2, 3$ and n going to infinity, $\{p_i\}_{i=1}^3$ being a Chebyshev set on I, then, for any function $f \in \mathcal{G}$, $\{\Phi_n(f)\}_n$ uniformly converges to f. The same statement holds if the "uniform convergence" is replaced by "pointwise convergence" or if the interval I is replaced by any subinterval J of I.*

Observe that, in the case of 2π-periodic functions, the most classical choice of the three test functions is given by 1, $\cos x$, and $\sin x$. In terms of Hermitian Toeplitz matrices, this means that the Korovkin test should be performed only on the three tridiagonal Toeplitz matrices $I = T_n(1)$, $C = T_n(\cos x)$ and $S = T_n(\sin x)$ where I is the identity matrix. All this is summarized in the following theorem.

Theorem 2.5.2. *Let f be a continuous real valued periodic function and J be a subinterval of $I = [-\pi, \pi]$. Then $L_n[U_n](f)$ converges uniformly to f on J if*

$$v(x)T_n(p)v^\star(x) - p(x)$$

converges uniformly to zero on J for any p being a test function. Here $v(x)$ is given by the formal expression of the j-th row of $U_n^\star$ where the value $x_j^{(n)}$ has been replaced by x.

In addition, the result due to Korovkin can be refined a bit. Let us suppose that the order of the error $\max_{i=1,2,3} \|\Phi_n(p_i) - p_i\|_\infty$ is $O(\theta_n)$ with θ_n going to zero as n tends to infinity. Then the same order of convergence holds for all trigonometric polynomials. This result, stated in the following proposition, is crucial for the strong convergence obtained in Theorem 2.6.4.

Proposition 2.5.1. [123]. *Let $\mathcal{G}$ be the linear space of the continuous (periodic) functions on a suitable interval $[-\pi, \pi]$ and let $\{\Phi_n(\cdot)\}_n$ be a sequence of linear positive operators from $\mathcal{G}$ to $\mathcal{G}$. Let $\|\cdot\|_{\infty,J}$ be the usual infinity norm on a set $J \subset [-\pi, \pi]$. If $\max_{i=1,2,3} \|\Phi_n(p_i) - p_i\|_{\infty,J} = O(\theta_n)$ for $\{p_i\}_{i=1}^3$ being the Chebyshev set: 1, $\sin x$ and $\cos x$ (cf. [102]) on $[-\pi, \pi]$, then, for any trigonometric polynomial p of a fixed degree (independent of n) we find*

$$\|\Phi_n(p) - p\|_{\infty,J} = O(\theta_n).$$

The same statement holds true if $\mathcal{G}$ is the linear space of the continuous even periodic functions, if p is an even trigonometric polynomial (with a finite cosine expansion) and if the Chebyshev set $\{p_i\}_{i=1}^3$ is given by 1, $\cos x$ and $\cos 2x$ over $I = [0, \pi]$ with $J \subset I$.

2.5.1 Some special cases

We start with the operator associated with the circulant algebra. In this case, there is nothing to check, since the eigenvalue function

$$L_n[U_n](\cdot)$$

is the Cesaro sum $[C_n(\cdot)](x)$ (for the details of this derivation see [45] and for a practical application see [133]). Therefore, as it is well known, this operator converges uniformly to the identity operator over $\mathbb{C}_{2\pi}(I, \mathbb{R}, \|\cdot\|_\infty)$ with $I = [-\pi, \pi]$ and has a rate of convergence of order n^{-1} on a class of functions (see [176] at pages $122-123$) which contain the polynomials (for a consequence of this property see Theorem 2.6.4).

Now we consider the τ and Hartley classes which are the algebras of all the matrices simultaneously diagonalized by the matrices $U_n = S_n$ and $U_n = H_n$ respectively, given in section 2.3.

Before going on, we need an explicit expression of the eigenvalues of the Frobenius-optimal approximation $\mathcal{P}_{U_n}(T_n(f))$.

Theorem 2.5.3. [54]. *Let f be the generating function of the Toeplitz matrix $T_n(f)$. Then the eigenvalues of $\mathcal{P}_{U_n}(T_n(f))$ are given by the values taken on the grid $\left\{\frac{i\pi}{n+1}\right\}$ by the function $[L_n[U_n](f)](x)$ defined as*

$$\begin{aligned} [L_n[U_n](f)](x) &= [K_n(f)](x) - \frac{2}{n+1}h(x), \\ h(x) &= s^{'}(x) - s(x)cot(x), \end{aligned}$$

where $s(x) = \sum_{j=1}^{n-1} a_j \sin(jx)$. Here $K_n(f)$ denotes the n-th Fourier sum of f.

Theorem 2.5.4. *The operator $L_n[U_n](\cdot)$ can be written as*

$$[L_n[U_n](f)](x) = [C_n(f)](x) - \frac{\cos(x)}{n+1}\sum_{j=0}^{n-2} a_{j+1}U_j(\cos(x))$$

where U_j denotes the j-th Chebyshev polynomial of second kind.

Concerning the representation of the eigenvalues of $\mathcal{P}_{U_n}(T_n(f))$ in the τ class, the vector $v(x)$ has the form

$$\sqrt{\frac{2}{n+1}}\,(\sin(x), \sin(2x), ..., \sin(nx))$$

where each function acts on $I = [0, \pi]$. Consequently the operator $L_n[U_n](f)$ is a combination of products of functions $\sin(jx)$ and so it is impossible to approximate the constant function $f \equiv 1$ at $x = 0$ and $x = \pi$.

We directly analyze the behavior of $L_n[U_n](f)$ on the test functions in the light of the Korovkin Theorem. Since the τ class is inherently symmetric, we have to consider symmetric Toeplitz matrices. In terms of functions, this means that the generating functions have to be even and so we consider the domain as the interval $I = [0, \pi]$. By direct calculation we find that $\{L_n[U_n](p)\}_n$ converges uniformly to p on each closed set $[a, b]$ contained in $(0, \pi)$ while this convergence is pointwise on $(0, \pi)$. As expected at $x = 0$ and $x = \pi$, there is no convergence. Finally, we point out that the calculation of $L_n[U_n](p)$ on the grid points of the algebra $x_i = \frac{i\pi}{n+1}$ leads to convergence to $p(x_i)$: this is a little surprising since the first points x_i and the last points x_{n-i}, for i fixed with respect to n and n going to infinity, tend to the critical points 0 and π, respectively. More precisely, the evaluation of $[L_n[U_n](1)](x) = v(x)Iv^{\star}(x)$ leads to

$$\begin{aligned} [L_n[U_n](1)](x) &= \frac{2}{n+1}\sum_{j=1}^{n} \sin^2(jx) \\ &= \frac{2}{n+1}\sum_{j=1}^{n} \left(\frac{e^{\hat{i}jx} - e^{-\hat{i}jx}}{2\hat{i}}\right)^2 \\ &= \frac{2n}{2(n+1)} - \frac{2}{4(n+1)}\sum_{j=1}^{n}(e^{2\hat{i}jx} + e^{-2\hat{i}jx}) \\ &= 1 - \frac{1}{n+1} - \frac{1}{2(n+1)}H(x). \end{aligned}$$

Here

$$H(x) = e^{2\hat{i}x}\frac{e^{2n\hat{i}x}-1}{e^{2\hat{i}x}-1} + e^{-2\hat{i}x}\frac{e^{-2n\hat{i}x}-1}{e^{-2\hat{i}x}-1}.$$

Then, for any closed interval in the open set $(0,\pi)$, we find that $H(x)$ is uniformly bounded and therefore we have uniform convergence to the constant 1. The convergence is pointwise in the open set $(0,\pi)$. On the other hand, for x going to 0, the expression $H(x)$ tends to $2n$ and, as expected, $[L_n[U_n](1)](0) = 0$. In a similar way we calculate $L_n[U_n](\cos(y))$ which is related to the tridiagonal Toeplitz matrix C and $L_n[U_n](\cos(2y))$ which is related to the pentadiagonal Toeplitz matrix P having $a_2 = a_{-2} = 1$ and $a_k = 0$ elsewhere. By making the same kind of check and by expanding the $\sin(jx)$ functions in terms of complex valued exponentials, we obtain that

$$[L_n[U_n](\cos(y))](x) = \cos(x) + O(n^{-1}) + \frac{1}{n+1}G_1(x)$$

and

$$[L_n[U_n](\cos(2y))](x) = \cos(2x) + O(n^{-1}) + \frac{1}{n+1}G_2(x)$$

where G_i are uniformly bounded functions in any compact set contained in $(0,\pi)$ and are diverging as n if we evaluate them at $x = 0$ and $x = \pi$. Therefore, for any continuous f defined on I, we find that

$\{L_n[U_n](f)\}_n$ converges to f uniformly on $[a,b] \subset (0,\pi)$,

$\{L_n[U_n](f)\}_n$ converges to f pointwise on $[0,\pi]$,

$\{L_n[U_n](f)\}_n$ converges to f pointwise on the grid points.

Concerning the Hartley class, the vector of function which characterizes the algebra is the following

$$\sqrt{\frac{1}{n}}\left(\sin(x)+\cos(x), \sin(2x)+\cos(2x), ..., \sin(nx)+\cos(nx)\right).$$

The eigenvalue function is also known [12, 52]

$$[L_n[U_n](f)](x) = [C_n(f)](x) - \frac{2}{n+1}\sum_k a_k \sin(kx).$$

In the light of the Korovkin Theorem we directly analyze the behavior of $L_n[U_n](f)$ on the test functions. Similar to the τ class, the Hartley one is intrinsically symmetric and so the test functions are $1, \cos x$ and $\cos 2x$ defined on $I = [0,\pi]$. By virtue of the same trivial checks performed for the τ class, we find that $\{L_n[U_n](p)\}_n$ converges uniformly to p on each closed set $[a,b]$ contained in $(0,\pi)$ while this convergence is pointwise on I. Moreover, $\{L_n[U_n](p)\}_n$ is convergent to p on the grid points of the algebra $x_i = \frac{2i\pi}{n}$.

Therefore, for any continuous f defined on I, we obtain that

$\{L_n[U_n](f)\}_n$ converges to f uniformly on $[a,b] \subset (0,\pi)$,

$\{L_n[U_n](f)\}_n$ converges to f pointwise on $(0,\pi)$,

$\{L_n[U_n](f)\}_n$ converges to f pointwise on the grid points.

2.5.2 Some remarks

Concerning the new linear positive operators related to the Frobenius-optimal approximation of Toeplitz matrices by trigonometric matrix algebras (τ class, Hartley class, etc...), it is interesting to observe that in [135] these LPOs have been used in connection with Theorem 2.6.5 in order to solve some nontrivial approximation problems (e.g. rational approximation of f/g with $g \geq 0$ and f/g continuous, "construction" of the essential range of f/g with $g \geq 0$ $f, g \in L^2$, etc...) under the assumption that computable expressions of f and g are unknown (only the Fourier coefficients of f and g are known) and when the generating functions are continuous or in L^2. More precisely, see chapter 9.

2.6 A Korovkin-type matrix theory

In this part, we would like to reduce the matrix approximation problem analyzed in section 2.4 to the very concise problem of checking the approximation only on the Toeplitz matrices generated by the test functions. This can be done with the help of the Lemma 2.4.1.

Before going on, it is useful to recall the theory about the distribution of the spectra of Toeplitz matrices. This kind of results goes back to Szegö [77], but we have very recently observed an impressive sequence of very interesting and enlightening papers on the subject [171], [109], [172], [164], [153].

Theorem 2.6.1. *[Szegö-Tyrtyshnikov [77, 164]. Let $f \in L^2$ and $\{\lambda_i^{(n)}\}_{i=1}^n$ be the eigenvalues of $T_n(f)$ (which are real since f is real valued and then the matrix $T_n(f)$ is Hermitian). Then, for any continuous function F with bounded support, we find the following asymptotic formula (the Szegö relation)*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} F(\lambda_i^{(n)}) = \frac{1}{2\pi} \int_{-\pi}^{\pi} F(f(x))dx. \tag{2.9}$$

In addition, mention has to be made of two very recent extensions accredited to Tilli [153] and Tyrtyshnikov, Zamarashkin [167]. The first one seems to be the most general with respect to the generality of the involved structures: actually this result is Theorem 2.6.1 modified to deal with finite dimensional Toeplitz operators generated by matrix-valued L^2 functions, including also the non-Hermitian case and the nonsquared one. On the other hand, in a complementary direction [167], the authors proved the classical Szegö formula for f ranging in L^1 space: of course, in the non-Hermitian case in formula (2.9), the eigenvalues must be replaced by the singular values.

For our purpose, another needed result is a second-order one concerning the Szegö formula. This refinement is a little simpler to state in terms of the squares of the singular values $\sigma_i^{(n)}$ of $T_n(f)$ (f is now complex valued) rather than the singular values themselves. The restriction on f is due to the assumption that it belongs to the Krein algebra K [91] (the Besov space $B_2^{\frac{1}{2}}$) of all the functions f that are essentially bounded and satisfy

$$\|f\|_K^2 = \sum_{k=-\infty}^{\infty} |k||a_k|^2 < \infty.$$

More precisely, this result is stated in the following theorem [172] (see also [121], [120]).

Theorem 2.6.2. *Let $\{\sigma_i^{(n)}\}$ the singular values of $T_n(f)$, $t_i^{(n)} = (\sigma_i^{(n)})^2$, $f \in K$ and let G be a function belonging to $C^3[m_f^2, M_f^2]$. Then*

$$\lim_{n\to\infty} \left\{ \sum_{i=1}^{n} G(t_i^{(n)}) - \frac{n}{2\pi} \int_{-\pi}^{\pi} G(|f(x)|^2)dx \right\} = c(f, G). \tag{2.10}$$

Here $c(f,G)$ is a known constant characterized in [172], *$M_f = \|f\|_\infty$ and m_f is the distance of the zero in the complex plane from the convex hull of the range of f. Indeed, if f is real valued then m_f is the minimum of $|f|$ and $\{t_i^{(n)}\}$ are the squares of the eigenvalues of $T_n(f)$.*

Moreover we need another technical lemma.

Lemma 2.6.1. *Let $\{S_n\}_n$ be a sequence of quasi-uniformly distributed grid points $x_i^{(n)}$ on I. Then, for any bounded and Riemann integrable function g, we have*

$$\sum_{i=0}^{n-1} g(x_i^{(n)}) = \frac{n}{2\pi}\int_{-\pi}^{\pi} g + o(n).$$

If the distribution is uniform and if g is bounded and Lipschitz continuous except, at most, for a finite number of discontinuity points, then

$$\sum_{i=0}^{n-1} g(x_i^{(n)}) = \frac{n}{2\pi}\int_{-\pi}^{\pi} g + O(1).$$

Proof. Since g is Riemann integrable, it follows that

$$\sum_{i=0}^{n-1} g(x_i^{(n)})(x_i^{(n)} - x_{i-1}^{(n)}) = \int_{-\pi}^{\pi} g + o(1),$$

and therefore, calling

$$S = \left| \sum_{i=0}^{n-1} g(x_i^{(n)}) - \frac{n}{2\pi}\int_{-\pi}^{\pi} g \right|,$$

it follows that

$$\begin{aligned} S &\leq \left| \sum_{i=0}^{n-1} g(x_i^{(n)}) - \frac{n}{2\pi}\sum_{i=0}^{n-1} g(x_i^{(n)})(x_i^{(n)} - x_{i-1}^{(n)}) \right| + o(n) \\ &= \frac{n}{2\pi}\left| \sum_{i=0}^{n-1} g(x_i^{(n)}) \left(\frac{2\pi}{n} - (x_i^{(n)} - x_{i-1}^{(n)}) \right) \right| + o(n) \\ &\leq \frac{n}{2\pi}\|g\|_\infty \sum_{i=0}^{n-1} \left| \frac{2\pi}{n} - (x_i^{(n)} - x_{i-1}^{(n)}) \right| + o(n) \\ &\leq \frac{n}{2\pi}\|g\|_\infty o(1) + o(n) = o(n). \end{aligned}$$

In a very similar way, we prove the other part of the lemma. □

Now we are ready to prove the matrix versions of the Korovkin result. In these theorems the test functions are given by 1, $\sin x$ and $\cos x$ in the 2π-periodic case and by 1, $\cos x$ and $\cos 2x$ in the case where the functions are also even (this situation typically occurs in inherently symmetric problems [62], [34], [132] in which we use inherently symmetric algebras as the τ class, the Hartley class or some cosine algebras). Moreover, the results are stated for f being real-valued because we are interested in the Hermitian case. However a complete generalization, requiring some technical tools and taking into account of the case where f is complex valued or matrix/tensor valued, is made in [124].

Theorem 2.6.3. [123, 124, 134]. *Let f be a continuous periodic function and let p a test function. If $L_n[U_n](p) = p + \epsilon_n(p)$ with ϵ_n going uniformly to zero, then $\{\mathcal{P}_{U_n}(T_n(f))\}_n$ converges to $\{T_n(f)\}_n$ in the weak sense.*

Proof. From identity 7 in Lemma 2.3.1, for any polynomial p we have

$$0 \leq \|T_n(p) - \mathcal{P}_{U_n}(T_n(p))\|_F^2 = \|T_n(p)\|_F^2 - \|\mathcal{P}_{U_n}(T_n(p))\|_F^2.$$

From the uniform convergence of $\{L_n[U_n](p)\}_n$ to p on the test functions we obtain the same convergence property for any polynomial of fixed degree. Therefore,

$$\|T_n(p) - \mathcal{P}_{U_n}(T_n(p))\|_F^2 = \|T_n(p)\|_F^2 - \sum_i \left(p(x_i^{(n)}) + \epsilon_n(p)(x_i^{(n)})\right)^2.$$

Now, from the definition of the Frobenius norm and since $T_n(p)$ is Hermitian, we find that

$$\|T_n(p)\|_F^2 = \sum_{i=1}^{n} \lambda_i^2(T_n(p)).$$

The preceding relation is very interesting because, after division by n, it coincides with the sum appearing in the left-hand side of the famous Szegö relation (see Theorem 2.6.1). Then, by applying the quoted result, we find

$$\|T_n(p)\|_F^2 = n \cdot \frac{1}{2\pi} \int_{-\pi}^{\pi} p^2 + o(n). \tag{2.11}$$

In addition, by exploiting the convergence of $\{L_n[U_n](p)\}_n$ to p, we may conclude that

$$\sum_{i=0}^{n-1} \left(p(x_i^{(n)}) + \epsilon_n(p)(x_i^{(n)})\right)^2 = \sum_i p^2(x_i^{(n)}) + o(n). \tag{2.12}$$

So, by virtue of the quasi-uniform distribution of grid points $\{x_i^{(n)}\}$, we arrive at

$$\sum_{i=0}^{n-1} \left(p(x_i^{(n)}) + \epsilon_n(p)(x_i^{(n)})\right)^2 = n \cdot \frac{1}{2\pi} \int_{-\pi}^{\pi} p^2 + o(n). \tag{2.13}$$

The combination of equations (2.11) and (2.13), in the light of the powerful Lemma 2.4.1, allows one to state the weak convergence of $\{\mathcal{P}_{U_n}(T_n(p))\}_n$ to $\{T_n(p)\}_n$. But, by noticing that this is the assumption of the second Weierstrass-Jackson type Theorem 2.4.2, the proof is proved. □

Theorem 2.6.4. [123, 124, 134]. *Under the same assumption of Theorem* 2.6.3, *if $\epsilon_n(p) = o(n^{-1})$ for the three test functions p and if the grid points of the algebra are uniformly distributed, then the convergence is strong.*

Proof. We follow the same proof given in Theorem 2.6.3. In particular, in all the equations (2.11), (2.12) and (2.13) the terms $o(n)$ are replaced by terms of constant order. In equation (2.11), we notice that all the polynomials are in the Krein algebra and then the second-order result due to Widom [172] can be applied (see Theorem 2.6.2) with $G(t) = t$. For the relation (2.12), the hypothesis on $\epsilon_n(p)$ with p test function and Lemma 2.5.1 are used while, for equation (2.13), we need the uniform distribution instead of the quasi-uniform one. Finally, Lemma 2.4.1 and the first Weierstrass-Jackson type Theorem are invoked. □

The L^2 case

It is worthwhile observing that we have used the last part of Lemma 2.3.1 in all the Korovkin-type Theorems. This is again the key for stating the following theorem which gives a criterion for testing the weak convergence in the L^2 case.

Theorem 2.6.5. *Let f be a function in L^2. If Szegö relation* (2.9)*, referred to the eigenvalues of $\{\mathcal{P}_{U_n}(T_n(f))\}_n$, holds only for $F(t) = t^2$, then the convergence is weak.*

Proof. We make use only of the last part of Lemma 2.3.1, of the ergodic relation stated in the hypothesis and of Szegö-Tyrtyshnikov Theorem 2.6.1. The structure of the proof follows the same steps as in Theorem 2.6.3. □

It should be remarked that Tyrtyshnikov [164] uses Theorem 2.6.1 in order to prove the weak clustering when circulant preconditioners are considered. Here, we find that for general algebras also containing the circulants, a much less restrictive condition is required, i.e., the validity of the Szegö relation only for $F(t) = t^2$. Observe that this is a very natural request because it can be viewed as the convergence of the discrete L^2-norm of $L_n[U_n](f)$ on the grid points of the algebra to the L^2-norm of f :

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n}[L_n[U_n](f)]^2(x_i^{(n)}) = \frac{1}{2\pi}\int_{-\pi}^{\pi} f^2(x)dx. \tag{2.14}$$

2.7 The multilevel case

By following the notations of Tyrtyshnikov, a multilevel Toeplitz matrix of level m and dimension $n_1 \times n_2 \times ... \times n_m$ is defined as the matrix generated by the Fourier coefficients of a multivariate Lebesgue integrable function $f = f(x_1, x_2, ..., x_m)$ according to the law given in equation (6.1) at page 23 in [164]. Similarly, given the unitary matrix U_n related to the transform of a one-level algebra, a corresponding m-level algebra is defined as the set of $n_1 \times n_2 \times ... \times n_m$ matrices simultaneously diagonalized by means of the following tensor product of matrices

$$U_n = U_{n_1} \otimes U_{n_2} \otimes ... \otimes U_{n_m}. \tag{2.15}$$

Now, since we are interested in extending the results proved in the preceding sections to m dimensions, we analyze what is necessary to have and, especially, what is kept when we switch from one dimension to m dimensions: for the first level we used the Weierstrass and Korovkin theorems, the last part of Lemma 2.3.1 (see [55]), Lemma 2.4.1 (see [164]) and the Szegö-Tyrtyshnikov theorem for one level Toeplitz matrices. Surprisingly enough, we find that all these tools hold or have a version in m dimensions: for the Korovkin and the Weierstrass theorems, the multidimensional extensions are classic results. Very recently, Tyrtyshnikov has proved the Szegö relation in any dimension [164] while Lemma 2.4.1 contains a statement not depending on the structure of the matrices and part (7) of Lemma 2.3.1 is valid for any algebra and so for multilevel algebras as well (recall that U_n in (2.15) is unitary).

Therefore, we instantly deduce the validity in m dimensions of the main statements that is Theorems 2.4.1, 2.4.2, 2.6.3, 2.6.4, 2.6.5. However, we remark that we have no examples in which the strong convergence holds.

For instance, in the two-level circulant and τ cases, only the weak convergence has been proved because the number of the outliers is, in both cases, equal to $O(n_1+n_2)$ [40], [53] even if the function f is a bivariate polynomial: more precisely, this means that the hypotheses of Theorems 2.4.1 and 2.6.4, regarding the strong approximation in the polynomial case, are not fulfilled by the two-level circulant and τ algebras and therefore strong convergence cannot be proved in the general case (see also [124]). Very recently, in [143] and [144] it has been proved that any sequence of preconditioners belonging to "partially Equimodular" algebras [144] cannot be superlinear for sequence of multilevel Toeplitz matrices generated by simple positive polynomials. Here, "partially Equimodular" refers to some very weak assumptions on U_n that are instantly fulfilled by all the known multilevel trigonometric algebras.

Conclusion.

In this chapter, we have studied the asymptotic equivalence of circulant and Toeplitz sequences (in the case where the generating function f belongs to the Wiener class of functions defined over the domain $[-\pi, \pi]$). Furthermore, a large part of works was consecrated to a detailed study of the Frobenius-optimal preconditioners for Toeplitz matrices. We will extend these notions in the following chapters for a spectral analysis and the singular value distribution of the g-circulants and g-Toeplitz sequences.

Chapter Three

Singular Values and Eigenvalues of the g-Circulant Matrices

3.1 Introduction

For a given nonnegative integer g, a matrix A_n of size n is called g-circulant if $A_n = [a_{(r-gs) mod\, n}]_{r,s=0}^{n-1}$. As example, if $n = 7$ and $g = 4$ we have

$$A_n \equiv C_{n,g} = \begin{bmatrix} a_0 & a_3 & a_6 & a_2 & a_5 & a_1 & a_4 \\ a_1 & a_4 & a_0 & a_3 & a_6 & a_2 & a_5 \\ a_2 & a_5 & a_1 & a_4 & a_0 & a_3 & a_6 \\ a_3 & a_6 & a_2 & a_5 & a_1 & a_4 & a_0 \\ a_4 & a_0 & a_3 & a_6 & a_2 & a_5 & a_1 \\ a_5 & a_1 & a_4 & a_0 & a_3 & a_6 & a_2 \\ a_6 & a_2 & a_5 & a_1 & a_4 & a_0 & a_3 \end{bmatrix}$$

Such kind of matrices arises in wavelet analysis [50] and subdivision algorithm or, equivalently, in the associated refinement equations, see [58] and references therein. Furthermore, it is interesting to remind that Gilbert Strang [150] has shown rich connections between dilation equations in the wavelets context and multigrid methods [78, 162], when constructing the restriction/prolongation operators [61, 1] with various boundary conditions. It is worth noticing that the use of different boundary conditions is quite natural when dealing with signal/image restoration problems or differential equations, see [129, 126].

This work treats the problem of characterizing of singular values and eigenvalues of the g-circulant matrices in the case where the sequence $\{a_{k(mod\, n)}\}$ defines the entries of the matrices and whose the values $\{a_k\}_k$ can be interpreted as the sequence of Fourier coefficients of an integrable function f over the domain $(-\pi, \pi)$. As special cases and observations, we will show interesting relations with the analysis of convergence of multigrid methods given, e.g., in [141], [1]. Finally we generalized the analysis in a simple example to the block, multilevel case, amounting to choose the symbol f multivariate, i.e., defined on the set $G = (-\pi, \pi)^d$ for some $d > 1$, and matrix valued, i.e., such that $f(x)$ is a matrix of given size $p \times q$.

3.2 General tools

For any $n \times n$ matrix A with eigenvalues $\lambda_j(A)$, $j = 1, 2, ..., n$, and for any $m \times n$ matrix B with singular values $\sigma_j(A)$, $j = 1, 2, ..., l$, $l = \min\{m, n\}$, we set

$$Eig(A) = \{\lambda_j(A) : \ j = 1, 2, ..., n\}, \ \ Sval(B) = \{\sigma_j(B) \ j = 1, 2, ..., l\}.$$

The matrix $B^\star B$ is positive semidefinite, since $x^\star(B^\star B)x = \|Bx\|_2^2 \geq 0$ for all $x \in \mathbb{C}^n$, with "$\star$" denoting the transpose conjugate operator. Moreover, it is clear that the eigenvalues $\lambda_1(B^\star B) \geq \lambda_2(B^\star B) \geq ... \geq \lambda_n(B^\star B) \geq 0$ are nonnegative and can therefore be written in the form

$$\lambda_j(B^\star B) = \sigma_j^2, \tag{3.1}$$

with $\sigma_j \geq 0$, $j = 1, 2, ..., n$. The numbers $\sigma_1 \geq \sigma_2 \geq ... \geq \sigma_l \geq 0$, $l = \min\{m, n\}$ are called "**singular values** of B", i.e., $\sigma_j = \sigma_j(B)$ and if $n > l$ then $\lambda_j(B^\star B) = 0$, $j = l+1, ..., n$. A more general statement is contained in the singular value decomposition theorem (see e.g. [72]).

Theorem 3.2.1. *Let B be an arbitrary (complex) $m \times n$ matrix. Then:*

(a) There exists a unitary $m \times m$ matrix U and a unitary $n \times n$ matrix V such that $U^\star AV = \Sigma$ is an $m \times n$ "diagonal matrix" of the following form:

$$\Sigma = \begin{bmatrix} D & 0 \\ 0 & 0 \end{bmatrix}, \; D := diag(\sigma_1, ..., \sigma_r). \; \sigma_1 \geq \sigma_2 \geq ... \geq \sigma_r \geq 0.$$

Here $\sigma_1, ..., \sigma_r$ are the nonvanishing singular values of B, and r is the rank of B.

*(b) The nonvanishing singular values of $A^\star$ are also precisely the number $\sigma_1, ..., \sigma_r$. The decomposition $B = U\Sigma V^\star$ is called "the **singular value decomposition** of B."*

We are interested in explicit formulae for the singular values and eigenvalues of g-circulant matrices. Following what is known in the standard case $g = 1$ (or $g = e = (1, 1, ..., 1)$ in the multilevel setting), we need to link the coefficients of the g-circulant matrix to a certain symbol.

Let f be a Lebesgue integrable function defined on $G = (-\pi, \pi)^d$ and taking values in $\mathcal{M}_{pq}$, for given positive integers p and q. Then, for d-indices $r = (r_1, r_2, ..., r_d)$, $j = (j_1, j_2, ..., j_d)$, $n = (n_1, n_2, ..., n_d)$, $e = (1, 1, ..., 1)$, $\underline{0} = (0, 0, ..., 0)$, the circulant matrix $C_n(f)$ of size $p\hat{n} \times q\hat{n}$, $\hat{n} = n_1.n_2...n_d$, is defined as follows

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

where $\tilde{f}_k$ are the Fourier coefficients of f defined by equation

$$\tilde{f}_j = \tilde{f}_{(j_1,...,j_d)} = \frac{1}{(2\pi)^d} \int_{[-\pi,\pi]^d} f(t_1, ..., t_d) e^{-\hat{i}(j_1 t_1 + ... + j_d t_d)} dt_1 ... dt_d, \quad \hat{i}^2 = -1, \tag{3.2}$$

for integers j_l such that $-\infty < j_l < \infty$ for $1 \leq l \leq d$. Since f is a matrix-valued function of d variables whose component functions are all integrable, then the $(j_1, j_2, ..., j_d)$-th Fourier coefficient is considered to be the matrix whose (u, v)-th entry is the $(j_1, j_2, ..., j_d)$-th Fourier coefficient of the function $(f(t_1, ..., t_d))_{u,v}$.

According to this multi-index block notation, we can define general multi-level block g-circulant matrices. Of course, in this multidimensional setting, g denotes a d-dimensional vector of nonnegative integers that is, $g = (g_1, g_2, ..., g_d)$. In that case $A_n = [a_{(r-g\circ s) mod\, n}]_{r,s=\underline{0}}^{n-e}$ where the $\circ$ operation is the componentwise product Hadamard between vector or matrices of the same size and where

$$(r - g \circ s) mod\, n = ((r_1 - g_1 \cdot s_1) mod\, n_1, (r_2 - g_2 \cdot s_2) mod\, n_2, ..., (r_d - g_d \cdot s_d) mod\, n_d).$$

3.2.1 The extremal cases where $g = 0$ or $g = e$, and the intermediate cases

We consider a d-level setting and we analyze in detail the case where $\underline{0} \leq g \leq e$ and with " $\leq$ " denoting the componentwise partial ordering between real vectors. When g has at least a zero component, the analysis can be reduced to the positive one as studied in subsection 3.2.2.

$\underline{g = e}$

In the literature the only case deeply studied is the case of $g = e$ (standard shift in every level). Here for multilevel block circulants $A_n = [a_{(r-g\circ s) mod\, n}]_{r,s=\underline{0}}^{n-e}$ the singular values are given by those of

$$\sigma_k(A_n) = \sum_{j=\underline{0}}^{n-e} a_j e^{\hat{i}2\pi(j_1k_1/n_1 + j_2k_2/n_2 + ... + j_dk_d/n_d)}, \quad k = (k_1, k_2, ..., k_d),$$

for any k_l such that $0 \leq k_l \leq n_l - 1$; $l = 1, 2, ..., d$. Of course when the coefficients a_j comes from the Fourier coefficients of a given Lebesgue integrable function f, i.e., $\tilde{f}_j = a_{j\, mod\, n}$, $j = -n/2, ..., n/2$ (where $n/2 = (n_1/2, ..., n_d/2)$), the singular values are those of $(n/2)$-th Fourier sum of f evaluated at the grid points

$$2\pi k/n = 2\pi(k_1/n_1, ..., k_d/n_d),$$

$0 \leq k_j \leq n_j - 1$; $j = 1, 2, ..., d$. Moreover the explicit Schur decomposition is known. For $d = p = q = 1$, according to chapter 1 any standard circulant can be written in the form

$$A_n \equiv C_n = F_n D_n F_n^{\star}, \tag{3.3}$$

where

$$D_n = diag(\sqrt{n} F_n^{\star} \underline{a}), \tag{3.4}$$

$$F_n = \frac{1}{\sqrt{n}} \left[e^{-\hat{i}\frac{2\pi jk}{n}} \right]_{j,k=0}^{n-1}, \text{ Fourier matrix}, \tag{3.5}$$

$$\underline{a} = [a_0, a_1, ..., a_{n-1}]^T, \text{ first column of the matrix } A_n. \tag{3.6}$$

Of course for general d, p, q the formula generalizes as

$$A_n = (F_n \otimes I_p) D_n (F_n^{\star} \otimes I_q),$$

with $F_n = F_{n_1} \otimes F_{n_2} \otimes ... \otimes F_{n_d}$, $D_n = diag(\sqrt{\hat{n}}(F_n^{\star} \otimes I_p)\underline{a})$, where $\hat{n} = n_1 n_2 ... n_d$ and $\underline{a}$ being the first "column" of A_n whose entries a_j, $j = (j_1, j_2, ..., j_d)$, ordered lexicographically, are blocks of sizes $p \times q$,

$\underline{g = \underline{0}}$

The other extreme is represented by the case where g is the zero vector. Here the multilevel block g-circulant is given by

$$A_n = [a_{(r-0\circ s) mod\, n}]_{r,s=\underline{0}}^{n-e} = [a_{r\, mod\, n}]_{r,s=\underline{0}}^{n-e} = [a_r]_{r,s=\underline{0}}^{n-e} = \begin{bmatrix} a_{\underline{0}} & \dots & a_{\underline{0}} \\ \vdots & & \vdots \\ a_{n-e} & \dots & a_{n-e} \end{bmatrix}.$$

A simple computation shows that all the singular values are zero except for few of them given by $\sqrt{\hat{n}}\sigma$, where $\hat{n} = n_1.n_2...n_d$ and σ is any singular value of the matrix $\left(\sum_{j=\underline{0}}^{n-e} a_j^{\star} a_j \right)^{1/2}$

and in the case where $p = q$, all the eigenvalues are also equal to zero except few of them which are the eigenvalues of the matrix $\sum_{j=\underline{0}}^{n-e} a_j$. In that case it is evident that

$$\{A_n\}_n \sim_\sigma (0, G) \text{ and } \{A_n\}_n \sim_\lambda (0, G)$$

where $G = (-\pi, \pi)^d$.

NB: the relation $\{A_n\}_n \sim_\sigma (0, G)$ means that the matrix sequence $\{A_n\}_n$ distributes in the sense of the singular values as the null function over the domain G. For more detail about the distribution, see chapter 4.

3.2.2 When some of the entries of g vanish

The content of this subsection reduces to the following remark: the case of a nonnegative g can be reduced to the case of a positive vector so that we are motivated to treat in detail the latter in section 3.3. Let g be a d-dimensional vector of nonnegative integers and let $\mathcal{N} \subset \{1, ..., d\}$ be the set of indices such that $j \in \mathcal{N}$ if and only if $g_j = 0$. Assume that $\mathcal{N}$ is nonempty, let $t \geq 1$ be its cardinality and $d^+ = d - t$. Then a simple calculation shows that the singular values of the corresponding g-circulant matrix $A_n = [a_{(r-g\circ s) mod\, n}]_{r,s=\underline{0}}^{n-e}$ are zero except for few of them given by $\sqrt{\hat{n}[0]}\sigma$ where

$$\hat{n}[0] = \prod_{j\in\mathcal{N}} n_j, \quad \hat{n}[0] = (n_{j_1}, n_{j_2}, ..., n_{j_t}), \quad \mathcal{N} = \{j_1, ..., j_t\},$$

and σ is any singular value of the matrix

$$\left(\sum_{j=\underline{0}}^{\hat{n}[0]-e} C_j^\star C_j \right)^{1/2}. \tag{3.7}$$

Also, the eigenvalues are equal to zero except few of them that are the eigenvalues of the matrix

$$\sum_{j=\underline{0}}^{\hat{n}[0]-e} C_j.$$

Here C_j is a d^+-level g^+-circulant matrix with $g^+ = (g_{k_1}, g_{k_2}, ..., g_{k_{d^+}})$ and of partial sizes $n[>0] = (n_{k_1}, n_{k_2}, ..., n_{k_{d^+}})$, $\mathcal{N}^C = \{k_1, k_2, ..., k_{d^+}\}$, and whose expression is

$$C_j = [a_{(r-g\circ s) mod}]_{r',s'=\underline{0}}^{n[>0]-e},$$

where $(r - g \circ s)_k = j_k$ for $g_k = 0$ and $r_i' = r_{k_i}$, $s_i' = s_{k_i}$, $i = 1, ..., d^+$. Also in this case, since most of the singular values are identically zero, we infer that

$$\{A_n\} \sim_\sigma (0, G).$$

3.3 Singular values of g-circulant matrices

Of course the aim of this chapter is to give the general picture for any nonnegative vector g. Since the notations can become quite heavy, for the sake of simplicity, we start with the case $d = p = q = 1$. Several generalizations, including also the degenerate case in which g has some zero entries is treated in section 3.5 via the observations in subsection 3.2.2, which

imply that the general analysis can be reduced to the case where all the entries of g are positive, that is $g_j > 0$, $j = 1, 2, ..., d$.

In the following, we denote by (n, g) the greatest common divisor of n and g, i.e., (n, g)=gcd(n, g), by $n_g = \frac{n}{(n,g)}$, by $\check{g} = \frac{g}{(n,g)}$, and by I_t the identity matrix of order t.

If we denote by C_n the classical circulant matrix (i.e. with $g = 1$) and by $C_{n,g}$ the g-circulant matrix generated by its elements, for generic n and g one immediately verifies that

$$C_{n,g} = C_n Z_{n,g}, \tag{3.8}$$

where

$$Z_{n,g} = [\delta_{r-gs}]_{r,s=0}^{n-1}; \qquad \delta_k = \begin{cases} 1 & \text{if } k \equiv 0 (mod\, n); \\ 0 & \text{otherwise.} \end{cases} \tag{3.9}$$

Lemma 3.3.1. *Let n be an integer greater than 2 such that*

$$Z_{n,g} = [\underbrace{\tilde{Z}_{n,g}|\tilde{Z}_{n,g}|...|\tilde{Z}_{n,g}}_{(n,g)\ times}] \tag{3.10}$$

where $Z_{n,g}$ is the matrix defined in (3.9) *and $\widetilde{Z}_{n,g} \in \mathbb{C}^{n \times n_g}$ is the submatrix of $Z_{n,g}$ obtained by considering only its first n_g columns, that is*

$$\widetilde{Z}_{n,g} = Z_{n,g} \left[\frac{I_{n_g}}{0} \right]. \tag{3.11}$$

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

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

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

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

and

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

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

Another useful fact is represented by the following equation

$$\widetilde{Z}_{n,g} = \widetilde{Z}_{n,(n,g)} Z_{n_g,\check{g}}, \tag{3.12}$$

where $Z_{n_g,\check{g}}$ is the matrix defined in (3.9) of dimension $n_g \times n_g$. Therefore

$$Z_{n_g,\check{g}} = \left[\widehat{\delta}_{r-\check{g}s} \right]_{r,s=0}^{n_g-1}, \quad \widehat{\delta}_k = \begin{cases} 1 & \text{if } k \equiv 0 (mod\, n_g), \\ 0 & \text{otherwise.} \end{cases} \tag{3.13}$$

Relation (3.12) will be used later.

Proof. (of relation (3.12)). For $r = 0, 1, ..., n-1$ and $s = 0, 1, ..., n_g - 1$, we find

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

and

$$\begin{aligned}(\widetilde{Z}_{n,(n,g)} Z_{n_g,\check{g}})_{r,s} &= \sum_{l=0}^{n_g-1} (\widetilde{Z}_{n,(n,g)})_{r,l} (Z_{n_g,\check{g}})_{l,s} \\ &= \sum_{l=0}^{n_g-1} \delta_{r-(n,g)l} \widehat{\delta}_{l-\check{g}s} \\ &\underset{(a)}{=} \delta_{r-(n,g)(\check{g}s) mod\, n_g} \\ &= \delta_{r-(n,g)\left(\frac{g}{(n,g)}s\right) mod\, n_g} \\ &\underset{(b)}{=} \delta_{r-(gs) mod\, n} \\ &= \delta_{(r-(gs) mod\, n) mod\, n} \\ &= \delta_{(r-gs) mod\, n},\end{aligned}$$

□

where

(a) holds true since there exists a unique $l \in \{0, 1, ..., n_g - 1\}$ such that $l - \check{g}s \equiv 0 (mod\, n_g)$, that is, $l \equiv \check{g}s (mod\, n_g)$ and hence $\delta_{r-(n,g)l} = \delta_{r-(n,g)\cdot(\check{g}s) mod\, n_g}$,

(b) is due to the following property: if we have three integer numbers ρ, θ and γ, then

$$\rho(\theta\, mod\, \gamma) = (\rho\theta)\, mod\, \rho\gamma.$$

Lemma 3.3.2. *If $g \geq n$ then $Z_{n,g} = Z_{n,g^o}$ where g^o is the unique integer which satisfies $g = tn + g^o$ where $0 \leq g^o < n$ and $t \in \mathbb{N}$; $Z_{n,g}$ is defined in* (3.9).

Remark 3.3.1. *One can define g^o as $g^o := g\, mod\, n$.*

Proof. From (3.9) we know that

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

For $r, s = 0, 1, ..., n-1$, one has

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

since $tns \equiv 0 (mod\, n)$. Whence $Z_{n,g} = Z_{n,g^o}$. □

The previous lemma tells us that, for g-circulant matrices, we can consider only the case where $0 \leq g < n$. In fact, if $g \geq n$, from (3.8) we infer that

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

Finally, it is worth noticing that the use of (3.3) and (3.8) implies that

$$C_{n,g} = F_n D_n F_n^\star Z_{n,g}. \tag{3.14}$$

Formula (3.14) plays an important role for studying the singular values and the eigenvalues of the g-circulant matrices.

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

Lemma 3.3.3. *Let F_n be the Fourier matrix of order n defined in (3.5) and $\widetilde{Z}_{n,g} \in \mathbb{C}^{n\times n_g}$ be the matrix represented in (3.11). Then*

$$F_n\widetilde{Z}_{n,g} = \frac{1}{\sqrt{(n,g)}} I_{n,g} F_{n_g} Z_{n_g,\check{g}}, \tag{3.15}$$

where $I_{n,g} \in \mathbb{C}^{n\times n_g}$ 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} being the identity matrix of size n_g and $Z_{n_g,\check{g}}$ as in (3.13).

Remark 3.3.2. $n = n_g \times (n,g)$.

Proof. (of Lemma 3.3.3). Rewrite the Fourier matrix as

$$F_n = \frac{1}{\sqrt{n}}[f_0|f_1|f_2|...|f_{n-1}],$$

where f_k, $k = 0,1,2,...,n-1$, is the k-th column of the Fourier matrix of order n:

$$f_k = [e^{-\frac{2\pi\hat{i}jk}{n}}]_{j=0}^{n-1} = \begin{bmatrix} e^{-\frac{2\pi\hat{i}k.0}{n}} \\ \hline e^{-\frac{2\pi\hat{i}k.1}{n}} \\ \hline \vdots \\ \hline e^{-\frac{2\pi\hat{i}k(n-1)}{n}} \end{bmatrix} \tag{3.16}$$

From (3.12)

$$F_n\widetilde{Z}_{n,g} = F_n\widetilde{Z}_{n,(n,g)}Z_{n_g,\check{g}} = \frac{1}{\sqrt{n}}[f_0|f_{1.(n,g)}|f_{2.(n,g)}|...|f_{(n_g-1).(n,g)}]Z_{n_g,\check{g}} \in \mathbb{C}^{n\times n_g}. \tag{3.17}$$

Indeed, For $k = 0,1,...,n_g-1$ and $j = 0,1,...,n-1$, one has

$$(F_n\widetilde{Z}_{n,(n,g)})_{jk} = \sum_{l=0}^{n-1}(F_n)_{jl}(\widetilde{Z}_{n,(n,g)})_{lk} = \sum_{l=0}^{n-1}\delta_{l-(n,g)k}e^{-\frac{2\pi\hat{i}kl}{n}}, \tag{3.18}$$

and since $0 \le (n,g)k \le n-(n,g)$, there exists a unique $l_k \in \{0,1,...,n-1\}$ such that $l_k - (n,g)k \equiv 0(mod\, n)$, so $l_k = (n,g)k$. Consequently relation (3.18) implies

$$(F_n\widetilde{Z}_{n,(n,g)})_{jk} = \delta_{l_k-(n,g)k}e^{-\frac{2\pi\hat{i}jl_k}{n}} = e^{-\frac{2\pi\hat{i}j(n,g)k}{n}} = (f_{(n,g)k})_j,$$

for all $0 \le j \le n-1$ and $0 \le k \le n_g-1$, and hence

$$F_n\widetilde{Z}_{n,(n,g)} = \frac{1}{\sqrt{n}}[f_0|f_{1.(n,g)}|f_{2.(n,g)}|...|f_{(n_g-1)\cdot(n,g)}].$$

For $k = 0,1,2,...,n_g-1$, one has

$$f_{(n,g)k} = [e^{-\frac{2\pi\hat{i}j(n,g)k}{n}}]_{j=0}^{n-1} = [e^{-\frac{2\pi\hat{i}jk}{n_g}}]_{j=0}^{n-1},$$

and then, taking into account the equalities $n = (n,g)\frac{n}{(n,g)} = (n,g)n_g$, we can write

$$f_{(n,g)k} = \begin{bmatrix} \left[e^{-\frac{2\pi\hat{i}jk}{n_g}}\right]_{j=0}^{n_g-1} \\ \hline \left[e^{-\frac{2\pi\hat{i}jk}{n_g}}\right]_{j=n_g}^{2n_g-1} \\ \hline \vdots \\ \hline \left[e^{-\frac{2\pi\hat{i}jk}{n_g}}\right]_{j=((n,g)-1)n_g}^{(n,g)n_g-1} \end{bmatrix}, \tag{3.19}$$

where

$$\left[e^{-\frac{2\pi\hat{i}jk}{n_g}}\right]_{j=0}^{n_g-1} = \begin{bmatrix} e^{-\frac{2\pi\hat{i}k.0}{n_g}} \\ \hline e^{-\frac{2\pi\hat{i}k.1}{n_g}} \\ \hline \vdots \\ \hline e^{-\frac{2\pi\hat{i}k.(n_g-1)}{n_g}} \end{bmatrix}, \tag{3.20}$$

According to formula (3.16), one observes that the vector in (3.20) is the k-th column of the Fourier matrix F_{n_g}. Furthermore, for $l = 0, 1, ..., (n,g)-1$, we find

$$\left[e^{-\frac{2\pi\hat{i}jk}{n_g}}\right]_{j=ln_g}^{(l+1)n_g-1} = \begin{bmatrix} e^{-\frac{2\pi\hat{i}kln_g}{n_g}} \\ \hline e^{-\frac{2\pi\hat{i}k(ln_g+1)}{n_g}} \\ \hline \vdots \\ \hline e^{-\frac{2\pi\hat{i}k.((l+1)n_g-1)}{n_g}} \end{bmatrix} = e^{-2\pi\hat{i}kl} \begin{bmatrix} e^{-\frac{2\pi\hat{i}k.0}{n_g}} \\ \hline e^{-\frac{2\pi\hat{i}k.1}{n_g}} \\ \hline \vdots \\ \hline e^{-\frac{2\pi\hat{i}k.(n_g-1)}{n_g}} \end{bmatrix} = \left[e^{-\frac{2\pi\hat{i}jk}{n_g}}\right]_{j=0}^{n_g-1}. \tag{3.21}$$

using (3.21), the expression of the vector in (3.20) becomes

$$f_{(n,g)k} = \left.\begin{bmatrix} \left[e^{-\frac{2\pi\hat{i}jk}{n_g}}\right]_{j=0}^{n_g-1} \\ \hline \left[e^{-\frac{2\pi\hat{i}jk}{n_g}}\right]_{j=0}^{n_g-1} \\ \hline \vdots \\ \hline \left[e^{-\frac{2\pi\hat{i}jk}{n_g}}\right]_{j=0}^{n_g-1} \end{bmatrix}\right\} (n,g)\ times. \tag{3.22}$$

Setting $\widetilde{f}_r = \left[e^{-\frac{2\pi\hat{i}jr}{n_g}}\right]_{j=0}^{n_g-1}$, for $0 \leq r \leq n_g - 1$, the Fourier matrix F_{n_g} of size n_g takes the form

$$F_{n_g} = \frac{1}{\sqrt{n_g}}[\widetilde{f}_0|\widetilde{f}_1|\widetilde{f}_2|...|\widetilde{f}_{n_g-1}] \tag{3.23}$$

From formula (3.20), the relation (3.22) can be expressed as

$$f_{(n,g)k} = \left.\begin{bmatrix} \widetilde{f}_k \\ \hline \widetilde{f}_k \\ \hline \vdots \\ \hline \widetilde{f}_k \end{bmatrix}\right\} (n,g)\ times.$$

and, as a consequence, formula (3.17) can be rewritten as

$$
\begin{aligned}
F_n\widetilde{Z}_{n,g} = F_n\widetilde{Z}_{n,(n,g)}Z_{n_g,\check{g}} &= \frac{1}{\sqrt{n}}\left[\begin{array}{c|c|c|c|c} \widetilde{f}_0 & \widetilde{f}_1 & \widetilde{f}_2 & \cdots & \widetilde{f}_{n_g-1} \\ \widetilde{f}_0 & \widetilde{f}_1 & \widetilde{f}_2 & \cdots & \widetilde{f}_{n_g-1} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \widetilde{f}_0 & \widetilde{f}_1 & \widetilde{f}_2 & \cdots & \widetilde{f}_{n_g-1} \end{array}\right] Z_{n_g,\check{g}} \\
&= \frac{1}{\sqrt{(n,g)n_g}}\left[\begin{array}{c} \sqrt{n_g}F_{n_g} \\ \hline \sqrt{n_g}F_{n_g} \\ \hline \vdots \\ \hline \sqrt{n_g}F_{n_g} \end{array}\right] Z_{n_g,\check{g}} \\
&= \frac{1}{\sqrt{(n,g)}}\left[\begin{array}{c} F_{n_g} \\ \hline F_{n_g} \\ \hline \vdots \\ \hline F_{n_g} \end{array}\right] Z_{n_g,\check{g}} \\
&= \frac{1}{\sqrt{(n,g)}}\left[\begin{array}{c} I_{n_g} \\ \hline I_{n_g} \\ \hline \vdots \\ \hline I_{n_g} \end{array}\right] F_{n_g}Z_{n_g,\check{g}} \\
&= \frac{1}{\sqrt{(n,g)}}I_{n,g}F_{n_g}Z_{n_g,\check{g}}.
\end{aligned}
$$

□

In the subsequent subsection, we will exploit Lemma 3.3.3 in order to characterize the singular values of g-circulant matrices $C_{n,g}$. Here we conclude the subsection with the following simple observations.

Remark 3.3.3. *In Lemma* 3.3.3, *if* $(n,g) = g$, *we have* $n_g = \frac{n}{(n,g)} = \frac{n}{g}$ *and* $\check{g} = \frac{g}{(n,g)} = 1$; *so the matrix* $Z_{n_g,\check{g}} = Z_{n_g,1}$, *appearing in* (3.15), *is the identity matrix of dimension* $n_g \times n_g$. *The relation* (3.15) *becomes*

$$F_n\widetilde{Z}_{n,g} = F_n\widetilde{Z}_{n,(n,g)}Z_{n_g,\check{g}} = \frac{1}{\sqrt{g}}I_{n,g}F_{n_g}.$$

The latter equation with $g = 2$ and even n appear (and is crucial) in the multigrid literature; see [141], equation (3.2), page 59 and, in slightly different form for the sine algebra of type I, see [60], section 2.1.

Remark 3.3.4. *If* $(n,g) = 1$, *Lemma* 3.3.3 *is trivial, because* $n_g = \frac{n}{(n,g)} = n$, $\check{g} = \frac{g}{(n,g)} = g$, *and so* $\widetilde{Z}_{n,g} = Z_{n,g}$. *The relation* (3.15) *becomes*

$$
\begin{aligned}
F_n\widetilde{Z}_{n,g} &= I_{n,g}F_{n_g}Z_{n_g,\check{g}} \\
&= F_nZ_{n,g},
\end{aligned}
$$

since the matrix $I_{n,g}$ *reduces by its definition to the identity matrix of order* n.

Remark 3.3.5. *Lemma* 3.3.3 *is true also if, instead of* F_n *and* F_{n_g}, *we put* $F_n^\star$ *and* $F_{n_g}^\star$, *respectively, because* $F_n^\star = \overline{F}_n$. *In fact there is no transposition, but only conjugation.*

3.3.2 Characterization of the singular values of the g-circulant matrices

Now we link the singular values of g-circulant matrices with the eigenvalues of its circulant counterpart C_n. This is nontrivial given the multiplicative relation $C_{n,g} = C_n Z_{n,g}$.

Having in mind the definition of the diagonal matrix D_n given in (3.4), we start by setting

$$D_n^\star D_n = diag(|D_n|_{s,s}^2;\ \text{s=0,1,...,n-1}) = diag(d_s;\ s = 0, 1, ..., n-1) = \bigoplus_{l=1}^{(n,g)} \Delta_l,$$

$$J_{(n,g)} \otimes I_{n_g} = \underbrace{[I_{n,g}|I_{n,g}|...|I_{n,g}]}_{(n,g)\ times} = \left.\begin{bmatrix} I_{n_g} & I_{n_g} & \dots & I_{n_g} \\ I_{n_g} & I_{n_g} & \dots & I_{n_g} \\ \vdots & \vdots & \vdots & \vdots \\ I_{n_g} & I_{n_g} & \dots & I_{n_g} \end{bmatrix}\right\} (n,g)\ times, \tag{3.24}$$

where

$$d_s = |D_n|_{s,s}^2 = (D_n)_{s,s}\overline{(D_n)_{s,s}},\quad s = 0, 1, ..., n-1, \tag{3.25}$$

$$\Delta_l = \begin{bmatrix} d_{(l-1)n_g} & & & \\ & d_{(l-1)n_g+1} & & \\ & & \ddots & \\ & & & d_{(l-1)n_g+n_g-1} \end{bmatrix} \in \mathbb{C}^{n_g \times n_g};\ \ l = 1, 2, ..., (n,g),$$

$$J_{(n,g)} = \left.\begin{bmatrix} 1 & 1 & \dots & 1 \\ 1 & 1 & \dots & 1 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & \dots & 1 \end{bmatrix}\right\} (n,g)\ \ times. \tag{3.26}$$

We now exploit relation (3.10) and Lemma 3.3.3, and we obtain that

$$\begin{aligned} F_n Z_{n,g} &= F_n \left[\widetilde{Z}_{n,g}|\widetilde{Z}_{n,g}|...|\widetilde{Z}_{n,g}\right] \\ &= \left[F_n\widetilde{Z}_{n,g}|F_n\widetilde{Z}_{n,g}|...|F_n\widetilde{Z}_{n,g}\right] \\ &= \frac{1}{\sqrt{(n,g)}}\left[I_{n,g}F_{n_g}Z_{n_g,\check{g}}|I_{n,g}F_{n_g}Z_{n_g,\check{g}}|...|I_{n,g}F_{n_g}Z_{n_g,\check{g}}\right] \\ &= \frac{1}{\sqrt{(n,g)}}\left[I_{n,g}|I_{n,g}|...|I_{n,g}\right] \left.\begin{bmatrix} F_{n_g}Z_{n_g,\check{g}} & & & \\ & F_{n_g}Z_{n_g,\check{g}} & & \\ & & \ddots & \\ & & & F_{n_g}Z_{n_g,\check{g}} \end{bmatrix}\right\} (n,g)\ times \\ &= \frac{1}{\sqrt{(n,g)}}\left[I_{n,g}|I_{n,g}|...|I_{n,g}\right](I_{(n,g)} \otimes F_{n_g}Z_{n_g,\check{g}}), \end{aligned}$$

so

$$F_n Z_{n,g} = \frac{1}{\sqrt{(n,g)}}\left[I_{n,g}|I_{n,g}|...|I_{n,g}\right](I_{(n,g)} \otimes F_{n_g}Z_{n_g,\check{g}}), \tag{3.27}$$

where $I_{(n,g)}$ is the identity matrix of order (n,g). Furthermore,

$$\begin{aligned} C_{n,g}^\star C_{n,g} &= (F_n D_n F_n^\star Z_{n,g})^\star (F_n D_n F_n^\star Z_{n,g}) \\ &= Z_{n,g}^\star F_n D_n^\star F_n^\star F_n D_n F_n^\star Z_{n,g} \\ &= Z_{n,g}^\star F_n D_n^\star D_n F_n^\star Z_{n,g} \\ &= (F_n^\star Z_{n,g})^\star D_n^\star D_n F_n^\star Z_{n,g}, \end{aligned}$$

so

$$C^{\star}_{n,g}C_{n,g} = (F^{\star}_n Z_{n,g})^{\star} D^{\star}_n D_n F^{\star}_n Z_{n,g}. \tag{3.28}$$

From (3.27) and (3.24), we plainly infer the following relations

$$\begin{aligned}
(F^{\star}_n Z_{n,g})^{\star} &= \left(\frac{1}{\sqrt{(n,g)}} \left[I_{n,g} | I_{n,g} | ... | I_{n,g} \right] \left(I_{(n,g)} \otimes F^{\star}_{n_g} Z_{n_g,\check{g}} \right) \right)^{\star} \\
&= \frac{1}{\sqrt{(n,g)}} \left(I_{(n,g)} \otimes F^{\star}_{n_g} Z_{n_g,\check{g}} \right)^{\star} \left(J_{(n,g)} \otimes I_{n_g} \right) \\
&= \frac{1}{\sqrt{(n,g)}} \left(I_{(n,g)} \otimes Z^{\star}_{n_g,\check{g}} F_{n_g} \right) \left(J_{(n,g)} \otimes I_{n_g} \right)
\end{aligned}$$

$$\begin{aligned}
F^{\star}_n Z_{n,g} &= \frac{1}{\sqrt{(n,g)}} \left[I_{n,g} | I_{n,g} | ... | I_{n,g} \right] \left(I_{(n,g)} \otimes F^{\star}_{n_g} Z_{n_g,\check{g}} \right) \\
&= \frac{1}{\sqrt{(n,g)}} \left(J_{(n,g)} \otimes I_{n_g} \right) \left(I_{(n,g)} \otimes F^{\star}_{n_g} Z_{n_g,\check{g}} \right).
\end{aligned}$$

Hence

$$C^{\star}_{n,g}C_{n,g} = \left(I_{(n,g)} \otimes Z^{\star}_{n_g,\check{g}} F_{n_g} \right) \left(J_{(n,g)} \otimes I_{n_g} \right) \frac{1}{(n,g)} D^{\star}_n D_n \left(J_{(n,g)} \otimes I_{n_g} \right) \left(I_{(n,g)} \otimes F^{\star}_{n_g} Z_{n_g,\check{g}} \right).$$

Now using the properties of the tensorial product

$$\begin{aligned}
\left(I_{(n,g)} \otimes Z^{\star}_{n_g,\check{g}} F_{n_g} \right) \left(I_{(n,g)} \otimes F^{\star}_{n_g} Z_{n_g,\check{g}} \right) &= I_{(n,g)} I_{(n,g)} \otimes Z^{\star}_{n_g,\check{g}} F_{n_g} F^{\star}_{n_g} Z_{n_g,\check{g}} \\
&= I_{(n,g)} \otimes Z^{\star}_{n_g,\check{g}} Z_{n_g,\check{g}} \\
&= I_{(n,g)} \otimes I_{n_g} = I_n,
\end{aligned}$$

and from a similarity argument, one deduces that the eigenvalues of $C^{\star}_{n,g}C_{n,g}$ are the eigenvalues of the matrix

$$
\begin{aligned}
&\left(J_{(n,g)}\otimes I_{n_g}\right)\tfrac{1}{(n,g)}D_n^{\star}D_n\left(J_{(n,g)}\otimes I_{n_g}\right)\\
&= \frac{1}{(n,g)}\left[\begin{array}{c|c|c|c} I_{n_g} & I_{n_g} & \dots & I_{n_g}\\ \hline I_{n_g} & I_{n_g} & \dots & I_{n_g}\\ \hline \vdots & \vdots & \vdots & \vdots\\ \hline I_{n_g} & I_{n_g} & \dots & I_{n_g}\end{array}\right]\left[\begin{array}{cccc}\Delta_1 & & & \\ & \Delta_2 & & \\ & & \ddots & \\ & & & \Delta_{(n,g)}\end{array}\right]\left[\begin{array}{c|c|c|c} I_{n_g} & I_{n_g} & \dots & I_{n_g}\\ \hline I_{n_g} & I_{n_g} & \dots & I_{n_g}\\ \hline \vdots & \vdots & \vdots & \vdots\\ \hline I_{n_g} & I_{n_g} & \dots & I_{n_g}\end{array}\right]\\
&= \frac{1}{(n,g)}\left[\begin{array}{c|c|c|c} I_{n_g} & I_{n_g} & \dots & I_{n_g}\\ \hline I_{n_g} & I_{n_g} & \dots & I_{n_g}\\ \hline \vdots & \vdots & \vdots & \vdots\\ \hline I_{n_g} & I_{n_g} & \dots & I_{n_g}\end{array}\right]\left[\begin{array}{c|c|c|c} \Delta_1 & \Delta_1 & \dots & \Delta_1\\ \hline \Delta_2 & \Delta_2 & \dots & \Delta_2\\ \hline \vdots & \vdots & \vdots & \vdots\\ \hline \Delta_{(n,g)} & \Delta_{(n,g)} & \dots & \Delta_{(n,g)}\end{array}\right]\\
&= \frac{1}{(n,g)}\left[\begin{array}{c|c|c|c} \sum_{l=1}^{(n,g)}\Delta_l & \sum_{l=1}^{(n,g)}\Delta_l & \dots & \sum_{l=1}^{(n,g)}\Delta_l\\ \hline \sum_{l=1}^{(n,g)}\Delta_l & \sum_{l=1}^{(n,g)}\Delta_l & \dots & \sum_{l=1}^{(n,g)}\Delta_l\\ \hline \vdots & \vdots & \vdots & \vdots\\ \hline \sum_{l=1}^{(n,g)}\Delta_l & \sum_{l=1}^{(n,g)}\Delta_l & \dots & \sum_{l=1}^{(n,g)}\Delta_l\end{array}\right]\\
&= \frac{1}{(n,g)}\underbrace{\left[\begin{array}{cccc}1 & 1 & \dots & 1\\ 1 & 1 & \dots & 1\\ \vdots & \vdots & \vdots & \vdots\\ 1 & 1 & \dots & 1\end{array}\right]}_{(n,g)\,times}\otimes\left(\sum_{l=1}^{(n,g)}\Delta_l\right).
\end{aligned}
$$

Therefore, from (3.26), we infer that

$$
Eig(C_{n,g}^{\star}C_{n,g}) = \frac{1}{(n,g)}Eig\left(J_{(n,g)}\otimes\sum_{l=1}^{(n,g)}\Delta_l\right), \tag{3.29}
$$

where

$$
\frac{1}{(n,g)}Eig\left(J_{(n,g)}\right) = \{0;1\}. \tag{3.30}
$$

Here we must observe that $\frac{1}{(n,g)}J_{(n,g)}$ is a matrix of rank 1, so it has all eigenvalues equal to zero except one eigenvalue equal to 1. In fact note that the trace of a matrix is, by definition, the sum of its eigenvalues: in our case the trace is $(n,g)\cdot\frac{1}{(n,g)}=1$ and hence the only nonzero eigenvalue is necessarily equal to 1. Moreover

$$
\begin{aligned}
\sum_{l=1}^{(n,g)}\Delta_l &= \sum_{l=1}^{(n,g)}diag(d_{(l-1)n_g+j}: 0\le j\le n_g-1)\\
&= diag\left(\sum_{l=1}^{(n,g)}d_{(l-1)n_g+j}: 0\le j\le n_g-1\right).
\end{aligned}
$$

Consequently, since $\sum_{l=1}^{(n,g)}\Delta_l$ is a diagonal matrix, we have

$$
Eig\left(\sum_{l=1}^{(n,g)}\Delta_l\right) = \left\{\sum_{l=1}^{(n,g)}d_{(l-1)n_g+j}: 0\le j\le n_g-1\right\}, \tag{3.31}
$$

where d_k are defined in (3.25).

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 $j = 1, 2, ..., p$ and $k = 1, 2, ..., q$. Therefore, by taking into consideration (3.29), (3.30) and (3.31) we find

$$\lambda_j(C_{n,g}^{\star} C_{n,g}) = \sum_{l=1}^{(n,g)} d_{(l-1)n_g+j}, \quad 0 \le j \le n_g - 1, \tag{3.32}$$

$$\lambda_j(C_{n,g}^{\star} C_{n,g}) = 0, \quad j = n_g, ..., n-1. \tag{3.33}$$

From (3.32), (2.33) and (3.1), one obtains that the singular values of an g-circulant matrix $C_{n,g}$ are given by

$$\sigma_j(C_{n,g}) = \sqrt{\sum_{l=1}^{(n,g)} d_{(l-1)n_g+j}}, \quad 0 \le j \le n_g - 1, \tag{3.34}$$

$$\sigma_j(C_{n,g}) = 0, \quad j = n_g, ..., n-1. \tag{3.35}$$

where the values d_k, $k = 0, 1, ..., n-1$ are defined in (3.25).

3.3.3 Special cases and observations

In this subsection we consider some special cases and we furnish a further link between the eigenvalues of circulant matrices and the singular values of g-circulants. In the case where $(n, g) = 1$, we have $n_g = \frac{n}{(n,g)} = n$. Hence the formula (3.34) becomes

$$\sigma_j(C_{n,g}) = \sqrt{d_j}, \quad 0 \le j \le n-1.$$

In order words, the singular values of $C_{n,g}$ coincide with those of C_n (this is excepted since $Z_{n,g}$ is a permutation matrix) and in particular with the modulus of the eigenvalues of C_n.

Concerning the eigenvalues of the circulant matrices it should be observed that formula (3.4) can be interpreted in function terms as the evaluation of a polynomial at the grid points given by the n-th roots of the unity. This is a standard observation because the Fourier matrix is a special instance of the classical Vandermonde matrices when the knots are exactly all the n-th roots of the unity.

Therefore, defining the polynomial $p(t) = \sum_{k=1}^{n-1} a_k e^{\hat{\imath} k t}$, it is trivial to observe that the eigenvalues of $C_n = F_n D_n F_{n^\star}$ are given by

$$\lambda_j(C_n) = p\left(\frac{2\pi j}{n}\right), \quad 0 \le j \le n-1.$$

The question that naturally arises is how to connect the expression in (3.34) of the nontrivial singular values of $C_{n,g}$ 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 section 4.4 in chapter 4)

$$\sigma_j(C_{n,g}) = \sqrt{\sum_{l=0}^{(n,g)-1} |p|^2 \left(\frac{x_j + 2\pi l}{(n,g)}\right)}, \quad x_j = \frac{2\pi j}{n_g}, \quad j = 0, 1, ..., n_g - 1. \tag{3.36}$$

In addition if g is fixed and a sequence of integers n is chosen so that $(n, g) > 1$ for n large enough, then $\{C_{n,g}\}_n \sim_\sigma (0, G)$ for a proper set G. If the sequence of n is chosen so that n and g are coprime for all n large enough, then the existence of the distribution is related to the smoothness properties of a function f such that $\{a_k\}_k$ can be interpreted as the sequence of its Fourier coefficients (see e.g. [131]). From the above reasoning it is clear that, if n is allowed to be vary among all the positive integer numbers, then the sequence $\{C_{n,g}\}_n$ does not possess a joint singular value distribution.

3.4 Eigenvalues of the g-circulant matrices

In perfect analogy with the singular values of the g-circulant matrix, we observe that $C_{n,g}$ can be written in the form: $C_{n,g} = F_n D_n M_{n,g} F_n^\star$, where D_n is the matrix defined in (3.4) and $M_{n,g}$ is a product of three matrices F_n, $Z_{n,g}$, and $F_n^\star$. Here $Z_{n,g}$ is the matrix defined in (3.9) and F_n is the Fourier matrix. However, the study of the eigenvalues of the matrix $C_{n,g}$ asks additional difficulties with respect to those of singular values, because it is not possible to find a direct method of determination of these eigenvalues. Despite all these difficulties, and since D_n is a diagonal matrix and $M_{n,g}$ is a sparse matrix whose entries are 0 and 1, one of the best techniques consists to construct a finite sequence of matrices $\left\{ M^{(k-1)}_{n_{g_{(k-1)}},g_k} \cdot \Delta^{(k-1)}_{n_{g_{(k-1)}}} \right\}_{k=0}^{s}$ of decreasing size satisfying $M^{(-1)}_{n_{g_{(-1)}},g_0} \cdot \Delta^{(-1)}_{n_{g_{(-1)}}} = M_{n,g} D_n$. Find the eigenvalues of $D_n M_{n,g}$, and because the matrices $C_{n,g}$ and $D_n M_{n,g}$ are similar, deduce those of $C_{n,g}$.

3.4.1 Some preliminary results

In the following we denote by $Eig(A)$ the spectrum of a matrix A.

Lemma 3.4.1. *Let a, b, k be three positive integers, then*

$$(a \bmod k)(b \bmod k) \bmod k \;=\; ab \bmod k \tag{3.37}$$

$$(a \bmod k \pm b \bmod k) \bmod k = (a \pm b) \bmod k \tag{3.38}$$

$$a \bmod k + b < k \Leftrightarrow a \bmod k + b = (a + b) \bmod k \tag{3.39}$$

Proof. Setting $a = a_0 + r_1 k$, $b = b_0 + r_2 k$ with $0 \le a_0, b_0 < k$. one has

$$\begin{aligned} (a \bmod k)(b \bmod k) \bmod k &= a_0 b_0 \bmod k \quad \text{and} \\ ab \bmod k &= a_0 b_0 \bmod k, \end{aligned}$$

furthermore

$$\begin{aligned} (a \bmod k \pm b \bmod k) \bmod k &= (a_0 \pm b_0) \bmod k \quad \text{and} \\ (a \pm b) \bmod k &= (a_0 \pm b_0) \bmod k \end{aligned}$$

finally

$$\begin{aligned} a \bmod k + b &= a_0 + b \;\text{ and }\; (a + b) \bmod k = (a_0 + b) \bmod k, \quad \text{hence} \\ a \bmod k + b < k &\Leftrightarrow a \bmod k + b = (a + b) \bmod k \end{aligned}$$

□

Proposition 3.4.1. *(Euler-Fermat Theorem). Let $a, b \in \mathbb{N}^\star$ (with $a < b$): If $(a, b) = 1$ then*

$$a^{\varphi(b)} \equiv 1 \, (mod \, b)$$

and

1. *$\varphi(b) = b - 1$ if b is coprime,*

2. *$\varphi(b) | b$ otherwise.*

Lemma 3.4.2. *If $(n, g) = 1$, then the map*

$$\begin{array}{rcl} l : \{0, 1, ..., n-1\} & \to & \{0, 1, ..., n-1\} \\ k & \mapsto & l(k) = l_k = gk \, mod \, n \end{array}$$

is a bijection and there exists $\varphi(n) \in \mathbb{N}^*$ such that for $j \in \{0, 1, ..., n-1\}$ fixed, and for every $q \in \{0, 1, ..., \varphi(n) - 1\}$ one has

$$l^q(j) = jg^q \, mod \, n \tag{3.40}$$

where $l^q = \underbrace{l \circ l \circ ... \circ l}_{q \text{ times}}$.

Proof. **Hypothesis**: $(n, g) = 1$
Injection. Let $k_1, k_2 \in \{0, 1, ..., n-1\}$ such that $l_{k_1} = l_{k_2}$ then

$$gk_1 - l_{k_1} \equiv 0 (mod \, n) \text{ and } gk_2 - l_{k_1} \equiv 0 (mod \, n)$$

then

$$g(k_1 - k_2) \equiv 0 (mod \, n)$$

then

$$k_1 - k_2 \equiv 0 (mod \, n) \text{ because } (n, g) = 1$$

then

$$k_1 = k_2.$$

So, the map l is an injection. Since the number of elements of the set $\{0, 1, ..., n-1\}$ is finite one deduces that the map l is a bijection.
Furthermore, since $(n, g) = 1$ then for $j \in \{0, 1, ..., n-1\}$ fixed, and for every $q \in \{0, 1, ..., \varphi(n) - 1\}$ one has

$$l^q(j) = jg^q \, mod \, n \text{ and } l^{\varphi(n)}(j) = j \, mod \, n \tag{3.41}$$

□

Lemma 3.4.3. *There exists an g-matrix $M_{n,g}$ such that*

$$Z_{n,g} = F_n M_{n,g} F_n^\star \tag{3.42}$$

with

$$M_{n,g} = [\delta^{(0)}_{gi-j}]^{n-1}_{i,j=0} \text{ with } \delta^{(0)}_k = \left\{ \begin{array}{ll} 1 \, if & k \equiv 0 (mod \, n); \\ 0 & otherwise. \end{array} \right.$$

Proof. For $j,k = 0,1,...,n-1$; one has

$$(F_n^\star Z_{n,g})_{jk} = \sum_{l=0}^{n-1}(F_n^\star)_{jl}(Z_{n,g})_{lk} = \sum_{l=0}^{n-1}\delta_{l-gk}(F_n^\star)_{jl} = (F_n^\star)_{jl_k} \tag{3.43}$$

because there exists a unique $(q_k, l_k) \in \mathbb{Z}^2$ with $0 \leq l_k < n$ such that $l_k - gk = q_k n$. According to (3.43), it follows that

$$(M_{n,g})_{ij} = (F_n^\star Z_{n,g} F_n)_{ij} = \sum_{k=0}^{n-1}(F_n^\star Z_{n,g})_{ik}(F_n)_{kj} = \sum_{k=0}^{n-1}(F_n^\star)_{il_k}(F_n)_{kj} = \frac{1}{n}\sum_{k=0}^{n-1} e^{\frac{\hat{i}2\pi i.l_k}{n}} e^{\frac{-\hat{i}2\pi k.j}{n}} =$$

$$\frac{1}{n}\sum_{k=0}^{n-1} e^{\frac{\hat{i}2\pi(i.l_k - k.j)}{n}} = \frac{1}{n}\sum_{k=0}^{n-1} e^{\frac{\hat{i}2\pi[i(g.k+nq_k)-k.j]}{n}} = \frac{1}{n}\sum_{k=0}^{n-1} e^{\hat{i}2\pi i q_k} e^{\frac{\hat{i}2\pi k(gi-j)}{n}} =$$

$$\frac{1}{n}\sum_{k=0}^{n-1} e^{\frac{\hat{i}2\pi k(gi-j)}{n}} = \begin{cases} 1 & \text{if } gi - j \equiv 0 (mod\, n); \\ 0 & \text{otherwise.} \end{cases}$$

according to Lemma 3.4.2. So,

$$M_{n,g} = [\delta^{(0)}_{gi-j}]_{i,j=0}^{n-1}$$

□

Remark 3.4.1. *One deduces from* (3.8), (3.3) *and* (3.42) *that*

$$\begin{aligned} C_{n,g} &= F_n D_n F_n^\star F_n M_{n,g} F_n^\star \\ &= F_n D_n M_{n,g} F_n^\star. \end{aligned}$$

Therefore

$$Eig(C_{n,g}) = Eig(D_n M_{n,g}) \tag{3.44}$$

3.4.2 Some preparatory tools

In the following we denote by $\delta^0 = (n,g)$ the greater common divisor of n and g and by $\widetilde{Z}_{n,g} \in \mathbb{C}^{n \times n_g}$ the matrix $Z_{n,g}$ defined in (3.9) by considering only the $n_g = \frac{n}{\delta^0}$ first columns.

Lemma 3.4.4. *Let n and g be two integers such that $1 < g < n$ and $M_{n,g}$ be the g-matrix defined by*

$$M_{n,g} = [\delta^{(0)}_{gi-j}]_{i,j=0}^{n-1} \text{ with } \delta_k^{(0)} = \begin{cases} 1 & \text{if } k \equiv 0 (mod\, n); \\ 0 & \text{otherwise.} \end{cases} \tag{3.45}$$

Then

$$M_{n,g} = \left.\begin{bmatrix} M_0 & M_1 & ... & M_{\delta^0-1} \\ M_0 & M_1 & ... & M_{\delta^0-1} \\ \vdots & \vdots & \vdots & \vdots \\ M_0 & M_1 & ... & M_{\delta^0-1} \end{bmatrix}\right\} \delta^0 \text{ times}$$

with for $j = 0,1,...,\delta^0 - 1$, $M_j \in \mathcal{M}_{n_g}(\mathbb{C})$.

Proof. According to (3.9) and (3.45), one has $M_{n,g} = Z_{n,g}^T$. Or

$$Z_{n,g}^T = \begin{bmatrix} \tilde{Z}_{n,g}^T \\ \tilde{Z}_{n,g}^T \\ \vdots \\ \vdots \\ \tilde{Z}_{n,g}^T \end{bmatrix} \text{ and } \tilde{Z}_{n,g}^T = [M_0|M_1|...|M_{\delta^0-1}],$$

because $\tilde{Z}_{n,g}^T \in \mathcal{M}_{n_g \times n}(\mathbb{C})$ and $n = n_g.\delta^0$, hence

$$M_{n,g} = \left.\begin{bmatrix} M_0 & M_1 & ... & M_{\delta^0-1} \\ M_0 & M_1 & ... & M_{\delta^0-1} \\ \vdots & \vdots & \vdots & \vdots \\ M_0 & M_1 & ... & M_{\delta^0-1} \end{bmatrix}\right\} \delta^0 \; times$$

with for $j = 0, 1, ..., \delta^0 - 1$, $M_j \in \mathcal{M}_{n_g}(\mathbb{C})$. □

Lemma 3.4.5.

$$Eig(C_{n,g}) = Eig\left(\sum_{j=0}^{\delta^0-1} M_j \Delta_j\right) \cup \{0 : \; mult. = n - n_g\},$$

where M_j is a matrix of order n_g and Δ_j a diagonal matrix also of order n_g.

Proof. According to Lemma 3.4.4, one has

$$\begin{aligned} M_{n,g}D_n &= \begin{bmatrix} M_0 & M_1 & ... & M_{\delta^0-1} \\ M_0 & M_1 & ... & M_{\delta^0-1} \\ \vdots & \vdots & \vdots & \vdots \\ M_0 & M_1 & ... & M_{\delta^0-1} \end{bmatrix} \begin{bmatrix} \Delta_0 & & & & \\ & \Delta_1 & & & \\ & & \ddots & & \\ & & & \ddots & \\ & & & & \Delta_{\delta^0-1} \end{bmatrix} \\ &= \begin{bmatrix} M_0\Delta_0 & M_1\Delta_1 & ... & M_{\delta^0-1}\Delta_{\delta^0-1} \\ M_0\Delta_0 & M_1\Delta_1 & ... & M_{\delta^0-1}\Delta_{\delta^0-1} \\ \vdots & \vdots & \vdots & \vdots \\ M_0\Delta_0 & M_1\Delta_1 & ... & M_{\delta^0-1}\Delta_{\delta^0-1} \end{bmatrix} \end{aligned}$$

then

$$\begin{aligned}
\det(M_{n,g}D_n - \lambda I_n) &= \begin{vmatrix} M_0\Delta_0 - \lambda I_{n_g} & M_1\Delta_1 & ... & M_{\delta^0-1}\Delta_{\delta^0-1} \\ M_0\Delta_0 & M_1\Delta_1 - \lambda I_{n_g} & M_2\Delta_2 & ... & M_{\delta^0-1}\Delta_{\delta^0-1} \\ \vdots & \vdots & \vdots & & \vdots \\ M_0\Delta_0 & M_1\Delta_1 & ... & & M_{\delta^0-1}\Delta_{\delta^0-1} - \lambda I_{n_g} \end{vmatrix} \\
&= \begin{vmatrix} -\lambda I_{n_g} + \sum_{j=0}^{\delta^0-1} M_j\Delta_j & M_1\Delta_1 & ... & M_{\delta^0-1}\Delta_{\delta^0-1} \\ -\lambda I_{n_g} + \sum_{j=0}^{\delta^0-1} M_j\Delta_j & -\lambda I + M_1\Delta_1 & M_2\Delta_2 & ... & M_{\delta^0-1}\Delta_{\delta^0-1} \\ \vdots & \vdots & \vdots & & \vdots \\ -\lambda I_{n_g} + \sum_{j=0}^{\delta^0-1} M_j\Delta_j & M_1\Delta_1 & ... & & -\lambda I + M_{\delta^0-1}\Delta_{\delta^0-1} \end{vmatrix} \\
&= \begin{vmatrix} -\lambda I_{n_g} + \sum_{j=0}^{\delta^0-1} M_j\Delta_j & M_1\Delta_1 & ... & & M_{\delta^0-1}\Delta_{\delta^0-1} \\ 0 & -\lambda I_{n_g} & 0 & ... & 0 \\ \vdots & 0 & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ 0 & \dots & \dots & 0 & -\lambda I_{n_g} \end{vmatrix} \\
&= (-\lambda)^{n-n_g} \left| -\lambda I_{n_g} + \sum_{j=0}^{\delta^0-1} M_j\Delta_j \right|
\end{aligned}$$

where I_{n_g} is the identity matrix of dimension $n_g \times n_g$, then

$$Eig(M_{n,g}D_n) = Eig\left(\sum_{j=0}^{\delta^0-1} M_j\Delta_j\right) \cup \{0 : mult. = n - n_g\} \tag{3.46}$$

From (3.44) it follows that

$$\begin{aligned}
Eig(C_{n,g}) &= Eig(D_nM_{n,g}) \\
&= Eig(M_{n,g}D_{n,g}) \\
&= Eig\left(\sum_{j=0}^{\delta^0-1} M_j\Delta_j\right) \cup \{0 : mult. = n - n_g\}.
\end{aligned}$$

□

Lemma 3.4.6. *Setting $g_1 = g \,(mod\, n_g)$ and defining $\delta_k^{(n_g)} = \begin{cases} 1 & \text{if } k \equiv 0 \,(mod\, n_g), \\ 0 & \text{otherwise.} \end{cases}$ then the equality* (3.47) *holds true:*

$$\sum_{k=0}^{\delta^0-1} M_k\Delta_k = \Delta_{n_g}^{(0)} M_{n_g,g_1}^{(0)} \tag{3.47}$$

where $\Delta_{n_g}^{(0)}$ is a diagonal matrix of order n_g, $M_{n_g,g_1}^{(0)}$ is a matrix of order n_g, and such that for $i, j = 0, 1, ..., n_g - 1$,

(3.48) $$\left(\Delta_{n_g}^{(0)}\right)_{jj} = d_{gj\,(mod\,n)} \text{ and } \left(M_{n_g,g_1}^{(0)}\right)_{ij} = \delta_{g_1 i-j}^{(n_g)}$$

with $d_k := d_{kk} = (D_n)_{k,k}$, D_n *is the diagonal matrix defined in* (3.4).

Proof. For $k = 0, 1, ..., \delta^0 - 1$; one has

$$M_k = \left[(M_{n,g})_{kn_g+i,kn_g+j}\right]_{i,j=0}^{n_g-1} \text{ and } \Delta_k = \left[(D_n)_{kn_g+i,kn_g+j}\right]_{i,j=0}^{n_g-1},$$

for $i, j = 0, 1, ..., n_g - 1$

$$\begin{aligned}
(M_k\Delta_k)_{i,j} &= \sum_{p=0}^{n_g-1} (M_k)_{ip}(\Delta_k)_{pj} \\
&= (M_k)_{ij}(\Delta_k)_{jj} \\
&= \delta_{g(kn_g+i)-kn_g-j}^{(n)} d_{kn_g+j,kn_g+j} \\
&= \delta_{gi-(kn_g+j)}^{(n)} d_{kn_g+j}
\end{aligned}$$

Since $g = \tilde{g}n_g + g_1$ then

$$gi = n_g\tilde{g}i + g_1 i = n_g(q_i\delta^0 + r_i) + g_1 i = q_i n + r_i n_g + g_1 i, \;\; 0 \le r_i < \delta^0.$$

Hence, $r_i n_g = (g - g_1)i \; mod\, n$ and

$$\begin{aligned}
\left(\sum_{k=0}^{\delta^0-1} M_k\Delta_k\right)_{i,j} &= \sum_{k=0}^{\delta^0-1} \delta_{gi-kn_g-j}^{(n)} d_{kn_g+j} \\
&= \sum_{k=0}^{\delta^0-1} \delta_{(r_i-k)n_g+g_1 i-j}^{(n)} d_{kn_g+j} \\
&\overset{(a)}{=} \delta_{g_1 i-j}^{(n)} d_{r_i n_g+j} \\
&= \delta_{g_1 i-j}^{(n)} d_{(g-g_1)i \bmod n+j} \\
&\overset{(b)}{=} \begin{cases} d_{gi \bmod n} & \text{if } j = g_1 i \; mod\, n \\ 0 & \text{otherwise} \end{cases} \\
&\overset{(c)}{=} \begin{cases} d_{gi \bmod n} & \text{if } j = g_1 i \; mod\, n_g \\ 0 & \text{otherwise} \end{cases} \\
&= \delta_{g_1 i-j}^{(n_g)} d_{gi \bmod n} \\
&\overset{(d)}{=} \left(\Delta_{n_g}^{(0)} M_{n_g,g_1}^{(0)}\right)_{ij}
\end{aligned}$$

- (a) is due to the fact that there exists a unique $k_i \in \{0, 1, ..., \delta^0 - 1\}$ such that $k_i = r_i$
- (b) follows from Lemma 3.4.1
- (c) is due to the fact that $j < n_g$
- (d) is a straightforward calculation of the entries of $\Delta_{n_g}^{(0)} M_{n_g,g_1}^{(0)}$

□

Remark 3.4.2. *If* $g_1 \neq 0$ *then* $(n_g, g) = (n_g, g_1)$.

Armed with the above tools we can start the study of the eigenvalues of the g-circulant matrices $C_{n,g}$.

3.4.3 Characterization of eigenvalues

In this subsection, we start our study in the case where the positive integers n and g are coprime. Furthermore, we reduce every time the size n of the matrix $C_{n,g}$ and the parameter g by discussing following the greater common divisor of the reduced quantities and we provide an algorithm that computes of recursive way the eigenvalues of $C_{n,g}$. Finally, we use the notation $d_s := d_{ss} = (D_n)_{s,s}$ where D_n is the diagonal matrix defined in (3.4).

Lemma 3.4.7. *If $(n,g) = 1$, then*

$$(M_{n,g}D_n)^{\varphi(n)} = \begin{bmatrix} d_0 d_{l_0}...d_{l_0^{\varphi(n)-1}} & & & \\ & d_1 d_{l_1}...d_{l_1^{\varphi(n)-1}} & & \\ & & \ddots & \\ & & & d_{n-1}d_{l_{n-1}}...d_{l_{n-1}^{\varphi(n)-1}} \end{bmatrix}$$

where $l : \{0,1,...,n-1\} \to \{0,1,...,n-1\}$ is the bijection defined in Lemma 3.4.2 satisfying $l_j^q = l^q(j)$, for $1 \leq q \leq \varphi(n)-1$ and $0 \leq j \leq n-1$ ($\varphi(n)$ is the Euler indicator). Furthermore $l^0 = l^{\varphi(n)} = id$.

Proof. First of all, $g^{\varphi(n)} \equiv 1 \, mod \, n$. For $i,k = 0,1,...,n-1$, one has

$$(M_{n,g}D_n)_{ik} = \sum_{l=0}^{n-1}(M_{n,g})_{il}(D_n)_{lk} = \sum_{l=0}^{n-1} d_{lk}\delta_{gi-l}^{(0)} = d_{l_i k} \tag{3.49}$$

according to Lemma 3.4.2.
also

$$\begin{aligned}(M_{n,g}D_nM_{n,g}D_n)_{ik} &= \sum_{p=0}^{n-1}(M_{n,g}D_n)_{ip}(M_{n,g}D_n)_{pk} \\ &= \sum_{p=0}^{n-1} d_{l_i p}d_{l_p k} \\ &= d_{l_i^2 k}d_{l_i l_i}\end{aligned}$$

according to (3.49).
Let us suppose that for all $q \in \{2,3,...,\varphi(n)-1\}$,

$$(M_{n,g}D_n)_{ik}^q = d_{l_i^q k}d_{l_i l_i}...d_{l_i^{q-1}l_i^{q-1}} \tag{3.50}$$

then

$$\begin{aligned}(M_{n,g}D_n)_{ik}^{\varphi(n)} &= \sum_{p=0}^{n-1}(M_{n,g}D_n)_{ip}^{\varphi(n)-1}(M_{n,g}D_n)_{pk} \\ &= \sum_{p=0}^{n-1} d_{l_i^{\varphi(n)-1}p}d_{l_i l_i}...d_{l_i^{\varphi(n)-2}l_i^{\varphi(n)-2}}d_{l_p k} \\ &= d_{l_i^{\varphi(n)-1}l_i^{\varphi(n)-1}}d_{l_i l_i}...d_{l_i^{\varphi(n)-2}l_i^{\varphi(n)-2}}d_{l_i^{\varphi(n)}k} \\ &= d_{ik}d_{l_i l_i}d_{l_i^2 l_i^2}...d_{l_i^{\varphi(n)-1}l_i^{\varphi(n)-1}}\end{aligned}$$

according to (3.50), (3.40) and Lemma 3.4.2.
Then

$$(3.51) \qquad (M_{n,g}D_n)^{\varphi(n)} \;=\; diag\left(\prod_{p=0}^{\varphi(n)-1} d_{l_j^p l_j^p} \,:\, j=0,1,...,n-1\right)$$

□

Lemma 3.4.8. *If* $(n,g)=1$, *then*

$$Eig(C_{n,g}) = Eig(M_{n,g}D_n) = \left\{ e^{\hat{i}\frac{2k(j)\pi}{\varphi(n)}} \prod_{m=0}^{\varphi(n)-1} f_{l^m_{p(j)}};\ j=0,1,...,n-1 \right\}$$

where

$$l^m_{p(j)} = g^m p(j) \bmod n \quad \text{and} \quad f_{l^m_{p(j)}} = L^{\frac{1}{\varphi(n)}}_{l^m_{p(j)}} e^{\hat{i}\frac{\theta_{l^m_{p(j)}}}{\varphi(n)}}$$

with $d_k = (D_n)_{k,k} = L_k e^{\hat{i}\theta_k}$, D_n *is the diagonal matrix defined in* (3.4).

Proof. Now, let us set $Eig(M_{n,g}D_n) = diag(\beta_0,\beta_1,...,\beta_{n-1})$ then, there exists a unitary matrix U such that

$$U^\star(M_{n,g}D_n)U = R = \begin{pmatrix} \beta_0 & \star & \dots & \star \\ & \beta_1 & \star\dots & \vdots \\ & & \ddots & \star \\ 0 & & & \beta_{n-1} \end{pmatrix}$$

then

$$(U^\star(M_{n,g}D_n)U)^{\varphi(n)} = U^\star(M_{n,g}D_n)^{\varphi(n)}U = R^{\varphi(n)} = \begin{pmatrix} \beta_0^{\varphi(n)} & \star & \dots & \star \\ & \beta_1^{\varphi(n)} & \star\dots & \vdots \\ & & \ddots & \star \\ 0 & & & \beta_{n-1}^{\varphi(n)} \end{pmatrix}.$$

So

$$Eig\left((M_{n,g}D_n)^{\varphi(n)}\right) = Eig(R^{\varphi(n)}),$$

so

$$\left\{ \prod_{p=0}^{\varphi(n)-1} d_{l_j^p} \,:\, j=0,1,...,n-1 \right\} = \left\{ \beta_j^{\varphi(n)} \,:\, j=0,1,...,n-1 \right\}.$$

Then, for every $j\in\{0,1,...n-1\}$ there exists an element $i_j\in\{0,1,...,n-1\}$ ($i_j=p(j)$ where p is a map from $\{0,1,...,n-1\}$) to $\{0,1,...,n-1\}$) such that

$$\beta_j^{\varphi(n)} = \prod_{m=0}^{\varphi(n)-1} d_{l^m_{p(j)}}.$$

Setting

$$\beta_j = a_j e^{\hat{i}\alpha_j} = [a_j,\alpha_j],\ \ d_t = L_t e^{\hat{i}\theta_t} = [L_t,\theta_t],$$

one obtains

$$\beta_j^{\varphi(n)} = \prod_{m=0}^{\varphi(n)-1} d_{l^m_{p(j)}} = \prod_{m=0}^{\varphi(n)-1} L_{l^m_{p(j)}} e^{\hat{i}\theta_{l^m_{p(j)}}}$$

then

$$[a_j^{\varphi(n)}, \varphi(n)\alpha_j] = \left[\prod_{m=0}^{\varphi(n)-1} L_{l^m_{p(j)}}, \sum_{m=0}^{\varphi(n)-1} \theta_{l^m_{p(j)}} \right]$$

then

$$a_j = \left(\prod_{m=0}^{\varphi(n)-1} L_{l^m_{p(j)}} \right)^{\frac{1}{\varphi(n)}} \tag{3.52}$$

and

$$\alpha_j \in \left\{ \frac{1}{\varphi(n)} \left(2k\pi + \sum_{m=0}^{\varphi(n)-1} \theta_{l^m_{p(j)}} \right) : \ k = 0, 1, ..., \varphi(n) - 1 \right\}.$$

Also, let us define by $k_j = k(j)$ an element of the set $\{0, 1, ..., \varphi(n)-1\}$ which corresponds to index of α_j, so

$$\alpha_j = \frac{1}{\varphi(n)} \left(2k(j)\pi + \sum_{m=0}^{\varphi(n)-1} \theta_{l^m_{p(j)}} \right) \tag{3.53}$$

It follows from (3.52) and (3.53) that

$$\beta_j = e^{\hat{i}\frac{2k(j)\pi}{\varphi(n)}} \prod_{m=0}^{\varphi(n)-1} L_{l^m_{p(j)}}^{\frac{1}{\varphi(n)}} e^{\hat{i}\frac{\theta_{l^m_{p(j)}}}{\varphi(n)}} = e^{\hat{i}\frac{2k(j)\pi}{\varphi(n)}} \prod_{m=0}^{\varphi(n)-1} f_{l^m_{p(j)}} \tag{3.54}$$

where

$$l^m_{p(j)} = g^m p(j) \, mod \, n \ \text{ and } \ f_{l^m_{p(j)}} = L_{l^m_{p(j)}}^{\frac{1}{\varphi(n)}} e^{\hat{i}\frac{\theta_{l^m_{p(j)}}}{\varphi(n)}}.$$

Whence

$$Eig(M_{n,g}D_n) = \left\{ e^{\hat{i}\frac{2k(j)\pi}{\varphi(n)}} \prod_{m=0}^{\varphi(n)-1} f_{l^m_{p(j)}}; \ j = 0, 1, ..., n-1 \right\}$$

Since

$$Eig(M_{n,g}D_n) = Eig(D_n M_{n,g})$$

it follows from (3.44) that

$$Eig(C_{n,g}) = \left\{ e^{\hat{i}\frac{2k(j)\pi}{\varphi(n)}} \prod_{m=0}^{\varphi(n)-1} f_{l^m_{p(j)}}; \ j = 0, 1, ..., n-1 \right\}.$$

□

Lemma 3.4.9. *If $n = g^p.n_0$ with $p \geq 1$, $n_0 \geq 1$ and $(n_0, g) = 1$. Then*

$$Eig(C_{n,g}) = \left\{ e^{\hat{i}\frac{2k(j)\pi}{\varphi(n_0)}} \prod_{q=0}^{\varphi(n_0)-1} f_{l^{p+q}_{v(j)}}; j = 0, 1, ..., n_0 - 1 \right\} \cup \{0 : \ mult. = n - n_0\}$$

where l is the bijection defined in Lemma 3.4.2 *and*

$$l^{p+q}_{v(j)} = g^{p+q} v(j) \, mod \, n \, , \quad f_{l^{p+q}_{v(j)}} = L^{\frac{1}{\varphi(n_0)}}_{l^{p+q}_{v(j)}} e^{\hat{i}\frac{\theta_{l^{p+q}_{v(j)}}}{\varphi(n_0)}}$$

with $d_j = L_j e^{\hat{i}\theta_j}$, $v(j)$ belongs to the set $\{0, 1, ..., n_0 - 1\}$ and $k(j) \in \{0, 1, ..., \varphi(n_0) - 1\}$.

Proof. Here $\delta^0 = (n, g) = g$. From (3.46) $-$ (3.47) it follows that

$$\begin{aligned} Eig(M_{n,g} D_{n,g}) &= Eig\left(\sum_{j=0}^{g-1} M_j \Delta_j \right) \cup \{0 : \ mult. = n - g^{p-1} n_0\}. \\ &= Eig\left(M^{(0)}_{g^{p-1}n_0, g} \Delta^{(0)}_{g^{p-1}n_0} \right) \cup \{0 : \ mult. = n - g^{p-1} n_0\} \end{aligned}$$

with

$$(\Delta^{(0)}_{g^{p-1}n_0})_{ii} = d_{gi \, mod \, n} = d_{gi}; \ \ (M^{(0)}_{g^{p-1}n_0, g})_{ij} = \delta^{(g^{p-1}n_0)}_{gi-j}; \quad i, j = 0, 1, ..., g^{p-1} n_0 - 1.$$

If $p > 1$, when working of similar way with the matrix $M^{(0)}_{g^{p-1}n_0, g} \Delta^{(0)}_{g^{p-1}n_0}$, one obtains

$$\begin{aligned} Eig(M^{(0)}_{g^{p-1}n_0, g} \Delta^{(0)}_{g^{p-1}n_0}) &= Eig\left(\sum_{j=0}^{g-1} M_j \Delta_j \right) \cup \{0 : \ mult. = g^{p-1} n_0 - g^{p-2} n_0\}. \\ &= Eig\left(M^{(1)}_{g^{p-2}n_0, g} \Delta^{(1)}_{g^{p-2}n_0} \right) \cup \{0 : \ mult. = g^{p-1} n_0 - g^{p-2} n_0\} \end{aligned}$$

where

$$(\Delta^{(1)}_{g^{p-2}n_0})_{ii} = d_{g(gi) \, mod \, n} = d_{g^2 i}; \ \ (M^{(1)}_{g^{p-2}n_0, g})_{ij} = \delta^{(g^{p-2}n_0)}_{gi-j}; \quad i, j = 0, 1, ..., g^{p-2} n_0 - 1.$$

So, one constructs by mathematical induction a finite matrix sequence $\left\{ M^{(k)}_{g^{p-k-1}n_0, g} \Delta^{(k)}_{g^{p-k-1}n_0} \right\}_{k=0}^{p-1}$ of decreasing order $g^{p-k-1} n_0$ such that for $k = 1, 2, ..., p - 1$

$$\begin{aligned} Eig\left(M^{(k-1)}_{g^{p-k}n_0, g} \Delta^{(k-1)}_{g^{p-k}n_0} \right) &= Eig\left(\sum_{j=0}^{g-1} M_j \Delta_j \right) \cup \{0 : \ mult. = g^{p-k} n_0 - g^{p-k-1} n_0\}. \\ &= Eig\left(M^{(k)}_{g^{p-k-1}n_0, g} \Delta^{(k)}_{g^{p-k-1}n_0} \right) \cup \{0 : \ \text{mult.} = (g^{p-k} - g^{p-k-1}) n_0\} \end{aligned}$$

then

(3.55)

$$Eig\left(M^{(k-1)}_{g^{p-k}n_0, g} \Delta^{(k-1)}_{g^{p-k}n_0} \right) = Eig\left(M^{(k)}_{g^{p-k-1}n_0, g} \Delta^{(k)}_{g^{p-k-1}n_0} \right) \cup \{0 : \ \text{mult.} = (g^{p-k} - g^{p-k-1}) n_0\}$$

with

$$(\Delta^{(k)}_{g^{p-k-1}n_0})_{ii} = d_{g^{k+1} i}; \ \ (M^{(k)}_{g^{p-k-1}n_0, g})_{ij} = \delta^{(g^{p-k-1}n_0)}_{gi-j}; \quad i, j = 0, 1, ..., g^{p-k-1} n_0 - 1.$$

From (3.44) and (3.55) it follows that

$$\begin{aligned} Eig(C_{n,g}) &= Eig(M_{n,g}D_n) \\ &= Eig\left(M^{(0)}_{g^{p-1}n_0,g}\Delta^{(0)}_{g^{p-1}n_0}\right) \cup \{0 : \text{ mult. } = g^p n_0 - g^{p-1}n_0\} \\ &\vdots \\ &= Eig\left(M^{(p-1)}_{n_0,g}\Delta^{(p-1)}_{n_0}\right) \cup \{0 : \text{ mult. } = g^p n_0 - n_0\} \\ &= Eig\left(M^{(p-1)}_{n_0,g_1}\Delta^{(p-1)}_{n_0}\right) \cup \{0 : \text{ mult. } = g^p n_0 - n_0\} \end{aligned} \tag{3.56}$$

where $g_1 = g \, mod \, n_0$. For $i, j = 0, 1, ..., n_0 - 1$

$$(\Delta^{(p-1)}_{n_0})_{jj} = d_{g^p j \, mod \, n} = d_{g^p j}; \;\; (M^{(p-1)}_{n_0,g_1})_{ij} = \delta^{(n_0)}_{g_1 i - j} = \begin{cases} 1 & \text{if } j \equiv g_1 i \, mod \, n_0 \\ 0 & \text{otherwise.} \end{cases}$$

• **Determination of** $Eig\left(M^{(p-1)}_{n_0,g_1}\Delta^{(p-1)}_{n_0}\right)$

Setting $\tilde{d}_j = d_{g^p j}$, since $(n_0, g_1) = 1$ then $g_1^{\varphi(n_0)} \equiv 1 \, mod \, n_0$. As in **case** $(n, g) = 1$, when replacing in relation (3.51) n, g and $\varphi(n)$ by n_0, g_1 and $\varphi(n_0)$, respectively, we obtain:

$$\left(M^{(p-1)}_{n_0,g_1}\Delta^{(p-1)}_{n_0}\right)^{\varphi(n_0)} = diag\left\{\prod_{q=0}^{\varphi(n_0)-1} \tilde{d}_{\tilde{l}^q_j} : j = 0, 1, ..., n_0 - 1\right\} \tag{3.57}$$

where

$$\tilde{l}^q_j = g_1^q j \, mod \, n_0 = g^q j \, mod \, n_0$$

and

$$\begin{aligned} \tilde{l} : \{0, 1, ..., n_0 - 1\} &\to \{0, 1, ..., n_0 - 1\} \\ j &\mapsto \tilde{l}(j) = \tilde{l}_j = g_1 j \, mod \, n_0 \end{aligned}$$

is a bijection (this is proved as in Lemma 3.4.2). Furthermore,

$$\begin{aligned} \tilde{d}_{\tilde{l}^q_j} &= \tilde{d}_{g^q j \, mod \, n_0} \\ &= d_{(g^q j \, mod \, n_0) g^p} \\ &= d_{g^{p+q} j \, mod \, n} = d_{l^q_{g^q j}} = d_{l^{p+q}_j} \end{aligned} \tag{3.58}$$

Always as in case $(n, g) = 1$ and according to (3.57) – (3.58), one deduces that the eigenvalues β_j of $M^{(p-1)}_{n_0,g_1}\Delta^{(p-1)}_{n_0}$ are given by

$$\beta_j = e^{\hat{i}\frac{2k(j)\pi}{\varphi(n_0)}} \prod_{m=0}^{\varphi(n_0)-1} L^{\frac{1}{\varphi(n_0)}}_{l^{p+m}_{v(j)}} e^{\hat{i}\frac{\theta_{l^{p+m}_{v(j)}}}{\varphi(n_0)}} = e^{\hat{i}\frac{2k(j)\pi}{\varphi(n_0)}} \prod_{m=0}^{\varphi(n_0)-1} f_{l^{p+m}_{v(j)}} \tag{3.59}$$

where

$$l^{p+m}_{v(j)} = g^{p+m} v(j) \, mod \, n \;\text{ and }\; f_{l^{p+m}_{v(j)}} = L^{\frac{1}{\varphi(n_0)}}_{l^{p+m}_{v(j)}} e^{\hat{i}\frac{\theta_{l^{p+m}_{v(j)}}}{\varphi(n_0)}}$$

with $d_j = L_j e^{\hat{i}\theta_j}$, $v(j) \in \{0, 1, ..., n_0 - 1\}$ and $k(j) \in \{0, 1, ..., \varphi(n_0) - 1\}$. So

$$Eig\left(M^{(p-1)}_{n_0,g}\Delta^{(p-1)}_{n_0}\right) = \left\{e^{\hat{i}\frac{2k(j)\pi}{\varphi(n_0)}} \prod_{m=0}^{\varphi(n_0)-1} f_{l^{p+m}_{v(j)}} : j = 0, 1, ..., n_0 - 1\right\}$$

from (3.56) we have that

$$Eig(C_{n,g}) = \left\{ e^{\hat{i}\frac{2k(j)\pi}{\varphi(n_0)}} \prod_{m=0}^{\varphi(n_0)-1} f_{l^{p+m}_{v(j)}} : j = 0, 1, ..., n_0 - 1 \right\} \cup \{0 : \text{ mult. } = n - n_0\}$$

□

Lemma 3.4.10. *If $(n_g, g_1) = 1$, then*

$$Eig(C_{n,g}) = \left\{ e^{\hat{i}\frac{2k(j)\pi}{\varphi(n_g)}} \prod_{p=1}^{\varphi(n_g)} f_{l^p_{v(j)}}; j = 0, 1, ..., n_g - 1 \right\} \cup \{0 : \textit{ mult. } = n - n_g\}$$

where l is the bijection defined in Lemma 3.4.2 *and*

$$l^p_{v(j)} = g^p v(j) \, mod \, n \quad \textit{and} \quad f_{l^p_{v(j)}} = L_{l^{p+m}_{v(j)}}^{\frac{1}{\varphi(n_g)}} e^{\hat{i}\frac{\theta_{l^p_{v(j)}}}{\varphi(n_g)}}$$

with $d_j = (D_n)_{j,j} = L_j e^{\hat{i}\theta_j}$, $v(j) \in \{0, 1, ..., n_g - 1\}$ and $k(j) \in \{0, 1, ..., \varphi(n_g) - 1\}$.

Proof. According to Lemma 3.4.5 and relation (3.47),

$$Eig(C_{n,g}) = Eig(M^{(0)}_{n_g,g_1} \Delta^{(0)}_{n_g}) \cup \{0 : \text{ mult. } = n - n_g\}$$

Since $(n_g, g_1) = 1$, there exists $\varphi(n_g) \in \mathbb{N}^*$ such that $g_1^{\varphi(n_g)} \equiv 1 \, mod \, n_g$.
Furthermore,

$$\begin{aligned}
(M^{(0)}_{n_g,g_1} \Delta^{(0)}_{n_g})^2_{ij} &= \sum_{l=0}^{n_g-1} (M^{(0)}_{n_g,g_1} \Delta^{(0)}_{n_g})_{il} (M^{(0)}_{n_g,g_1} \Delta^{(0)}_{n_g})_{lj} \\
&= \sum_{l=0}^{n_g-1} \delta^{(n_g)}_{g_1 i - l} d_{gi \, mod \, n} \delta^{(n_g)}_{g_1 l - j} d_{gl \, mod \, n} \\
&= d_{gi \, mod \, n} \delta^{(n_g)}_{g(g_1 i \, mod \, n_g) - j} d_{g(g_1 i \, mod \, n_g) \, mod \, n} \\
&\overset{(\gamma)}{=} \delta^{(n_g)}_{g(g_1 i \, mod \, n_g) - j} d_{gi \, mod \, n} d_{g^2 i \, mod \, n}
\end{aligned}$$

(γ) follows from $g(g_1 i \, mod \, n_g) \, mod \, n = (g^2 i \, mod \, \tilde{g} n) \, mod \, n = g^2 i \, mod \, n$, where $g = \tilde{g}\delta^0$.

Setting $\beta^{(h)}(j) = j g_1^h \, mod \, n_g$, one shows by mathematical induction that

$$(M^{(0)}_{n_g,g_1} \Delta^{(0)}_{n_g})^{\varphi(n_g)}_{ij} = \delta^{(n_g)}_{g\beta^{(\varphi(n_g)-1)}(i) - j} \prod_{p=1}^{\varphi(n_g)} d_{g^p i \, mod \, n}$$

It follows from $(3.38) - (3.39)$ that

$$\begin{aligned}
\left(g\beta^{(\varphi(n_g)-1)}(i) - j\right) \, mod \, n_g &= \left(g\beta^{(\varphi(n_g)-1)}(i) \, mod \, n_g - j \, mod \, n_g\right) \, mod \, n_g \\
&= \left((g_1 \beta^{(\varphi(n_g)-1)}(i) \, mod \, n_g) \, mod \, n_g - j\right) \, mod \, n_g \\
&= \left((g_1 \cdot g_1^{\varphi(n_g)-1} i \, mod \, n_g) \, mod \, n_g - j\right) \, mod \, n_g \\
&= \left(g_1^{\varphi(n_g)} i \, mod \, n_g - j\right) \, mod \, n_g \\
&= (i - j) \, mod \, n_g = 0 \Leftrightarrow i = j.
\end{aligned}$$

(because $g_1 = g \, mod \, n_g$ and $g_1^{\varphi(n_g)} \equiv 1 \, mod \, n_g$). Then

$$(M^{(0)}_{n_g,g_1}\Delta^{(0)}_{n_g})^{\varphi(n_g)}_{ij} \;=\; \begin{cases} \prod\limits_{p=1}^{\varphi(n_g)} d_{g^p i \, mod \, n} & \text{if j=i} \\ 0 & \text{otherwise.} \end{cases}$$

Setting $d_t = L_t e^{\hat{i}\theta_t} = [L_t, \theta_t]$ $(t \in \mathbb{N})$ then

$$(M^{(0)}_{n_g,g_1}\Delta^{(0)}_{n_g})^{\varphi(n_g)}_{jj} = \left[\prod_{p=1}^{\varphi(n_g)} L_{g^p j \; mod \, n}; \sum_{p=1}^{\varphi(n_g)} \theta_{g^p j \, mod \, n} \right]$$

Let $Eig(M^{(0)}_{n_g,g_1}\Delta^{(0)}_{n_g}) = \{\lambda_j : \; j = 0, 1, ..., n_g - 1\}$. As in the case $(n, g) = 1$ we have that

$$\lambda_j \;=\; e^{\hat{i}\frac{2k(j)\pi}{\varphi(n_g)}} \prod_{p=1}^{\varphi(n_g)} L^{\frac{1}{\varphi(n_g)}}_{g^p v(j) \, mod \, n} e^{\hat{i}\frac{\theta_{g^p v(j) \, mod \, n}}{\varphi(n_g)}} = e^{\hat{i}\frac{2k(j)\pi}{\varphi(n_g)}} \prod_{p=1}^{\varphi(n_g)} f_{g^p v(j) \, mod \, n}$$

where $v(j) \in \{0, 1, ..., n_g - 1\}$, $k(j)$ belongs to the set $\{0, 1, ..., \varphi(n_g) - 1\}$ and

$$f_{g^p v(j) \, mod \, n} = L^{\frac{1}{\varphi(n_g)}}_{g^p v(j) \, mod \, n} e^{\hat{i}\frac{\theta_{g^p v(j) \, mod \, n}}{\varphi(n_g)}}$$

Therefore

$$Eig(C_{n,g}) \;=\; \left\{ e^{\hat{i}\frac{2k(j)\pi}{\varphi(n_g)}} \prod_{p=1}^{\varphi(n_g)} f_{g^p v(j) \, mod \, n} : \; j = 0, 1, ..., n_g - 1 \right\} \cup \{0 : \text{mult.} = n - n_g\}$$

□

Moreover, we have the following fundamental lemmas.

Lemma 3.4.11. *Setting $\delta^{(0)} = (n_g, g_1) > 1$, there exists a decreasing (componentwise) finite sequence $\left\{(g_{k+1}, \delta^{(k)}, n_{g_{(k)}})\right\}_k$ of elements of $\mathbb{N}^3$ satisfying*

$$g_0 = g; \;\; n_{g_{(0)}} = n_g; \;\; g_1 = g \; mod \, n_g;$$

and for $k \in \mathbb{N}^\star$

$$g_k = g_{k-1} \; mod \, n_{g_{(k-1)}}; \;\; \delta^{(k-1)} = (n_{g_{(k-1)}}, g_k); \;\; n_{g_{(k)}} = \frac{n_{g_{(k-1)}}}{\delta^{(k-1)}};$$

and a decreasing matrix sequence $\left\{M^{(k)}_{n_{g_{(k)}},g_{k+1}}\Delta^{(k)}_{n_{g_{(k)}}}\right\}_k$ of order $n_{g_{(k)}}$ (this sequence is finite since the matrix $M_{n,g}D_n$ is of order n) such that

a) $M^{(0)}_{n_{g_{(0)}},g_1}\Delta^{(0)}_{n_{g_{(0)}}} = M^{(0)}_{n_g,g_1}\Delta^{(0)}_{n_g}$,

b) $\Delta^{(k)}_{n_{g_{(k)}}} = diag\left(d_{g^{k+1} j \, mod \, n}) : j = 0, 1, ..., n_{g_{(k)}} - 1\right)$,

c) $M^{(k)}_{n_{g_{(k)}},g_{k+1}} = \left[\delta^{(n_{g_{(k)}})}_{g_{k+1}i-j}\right]^{n_{g_{(k)}}-1}_{i,j=0}$; *where* $\delta^{(n_{g_{(k)}})}_q = \begin{cases} 1 & \text{if } q \equiv 0 \, mod \, n_{g_{(k)}} \\ 0 & \text{otherwise} \end{cases}$

Proof. First of all, let us set:

$$g_0 = g;\ n_{g_{(0)}} = n_g;\ g_1 = g\ mod\, n_g;\ \delta^{(0)} = (n_g, g_1);\ n_{g_{(1)}} = \frac{n_g}{\delta^{(0)}};$$

As it was shown in Lemmas $3.4.5 - 3.4.6$, one immediately verifies that

$$Eig\left(M^{(0)}_{n_g,g_1}\Delta^{(0)}_{n_g}\right) \;=\; Eig\left(\sum_{k=0}^{\delta^{(0)}-1} M^{(1)}_k\Delta^{(1)}_k\right) \cup \{0 :\ \text{mult.} = n_g - n_{g_{(1)}}\}$$

where for $k = 0, 1, ..., \delta^{(0)} - 1$ and $i, j = 0, 1, ..., n_{g_{(1)}} - 1$

$$\left(M^{(1)}_k\right)_{ij} = \left(M^{(0)}_{n_g,g_1}\right)_{kn_{g_{(1)}}+i,kn_{g_{(1)}}+j} = \delta^{(n_g)}_{g_1(kn_{g_{(1)}}+i)-(kn_{g_{(1)}}+j)} \overset{(a)}{=} \delta^{(n_g)}_{(g_1i-j)-kn_{g_{(1)}}}$$

and

$$\left(\Delta^{(1)}_k\right)_{jj} = \left(\Delta^{(0)}_{n_g}\right)_{kn_{g_{(1)}}+j,kn_{g_{(1)}}+j} = d_{g(kn_{g_{(1)}}+j)mod\, n}$$

(a) follows from $g_1 n_{g_{(1)}} = \frac{g_1}{\delta^{(0)}}\delta^{(0)} n_{g_{(1)}} = \frac{g_1}{\delta^{(0)}} n_g \equiv 0\, mod\, n_g$. Furthermore, setting $g_2 = g_1\, mod\, n_{g_{(1)}}$, for $i, j = 0, 1, ..., n_{g_{(1)}} - 1$

$$\begin{aligned}
(M^{(1)}_k\Delta^{(1)}_k)_{i,j} &= \sum_{p=0}^{n_{g_{(1)}}-1} (M^{(1)}_k)_{ip}(\Delta^{(1)}_k)_{pj} \\
&= (M^{(1)}_k)_{ij}(\Delta^{(1)}_k)_{jj} \\
&= \delta^{(n_g)}_{g_1i-(kn_{g_{(1)}}+j)} d_{g(kn_{g_{(1)}}+j)mod\, n}
\end{aligned}$$

Since $g_1 = \tilde{g}_1 n_{g_{(1)}} + g_2$ then

$$g_1 i = n_{g_{(1)}}\tilde{g}_1 i + g_2 i = n_{g_{(1)}}(q_i\delta^{(0)} + r_i) + g_2 i = q_i n_g + r_i n_{g_{(1)}} + g_2 i,\ \ 0 \le r_i < \delta^{(0)}.$$

Because $r_i n_{g_{(1)}} = (g_1 - g_2)i\ mod\, n_g$, we have that

$$\begin{aligned}
\left(\sum_{k=0}^{\delta^{(0)}-1} M^{(1)}_k\Delta^{(1)}_k\right)_{i,j} &= \sum_{k=0}^{\delta^{(0)}-1} \delta^{(n_g)}_{g_1i-kn_{g_{(1)}}-j} d_{g(kn_{g_{(1)}}+j)mod\, n} \\
&= \sum_{k=0}^{\delta^{(0)}-1} \delta^{(n_g)}_{(r_i-k)n_{g_{(1)}}+g_2i-j} d_{g(kn_{g_{(1)}}+j)mod\, n} \\
&\overset{(e)}{=} \delta^{(n_g)}_{g_2i-j} d_{g(r_in_{g_{(1)}}+j)mod\, n} \\
&= \delta^{(n_g)}_{g_2i-j} d_{g((g_1-g_2)i\, mod\, n_g+j)mod\, n} \\
&\overset{(f)}{=} \begin{cases} d_{g(g_1i\, mod\, n_g)\, mod\, n} & \text{if } j \equiv g_2 i\ mod\, n_g \\ 0 & \text{otherwise} \end{cases} \\
&\overset{(h)}{=} \begin{cases} d_{g^2i\, mod\, n} & \text{if } j \equiv g_2 i\ mod\, n_{g_{(1)}} \\ 0 & \text{otherwise} \end{cases} \\
&= \delta^{(n_{g_{(1)}})}_{g_2i-j}\, d_{g^2i\, mod\, n} \\
&\overset{(l)}{=} \left(\Delta^{(1)}_{n_{g_{(1)}}} M^{(1)}_{n_{g_{(1)}},g_2}\right)_{ij}
\end{aligned}$$

• (e) holds true because there exists a unique $k_i \in \{0, 1, ..., \delta^{(0)} - 1\}$ such that $k_i = r_i$,
• (f) follows from Lemma 3.4.1,
• (h) follows from $j < n_{g_{(1)}}$,
• (l) is a straightforward calculation of the entries of $\Delta^{(1)}_{n_{g_{(1)}}} M^{(1)}_{n_{g_{(1)}},g_2}$. Here g_2 can be equal to zero.

Finally, for $k \in \mathbb{N}^\star$, let us define

$$g_k = g_{k-1} mod\, n_{g_{(k-1)}};\ \delta^{(k-1)} = (n_{g_{(k-1)}}, g_k);\ n_{g_{(k)}} = \frac{n_{g_{(k-1)}}}{\delta^{(k-1)}}. \tag{3.60}$$

From (3.60) one constructs by mathematical induction a decreasing (componentwise) finite sequence $\left\{(g_{k+1}, \delta^{(k)}, n_{g_{(k)}})\right\}_{k\in\mathbb{N}}$ of elements of $\mathbb{N}^3$ and a matrix sequence $\left\{M^{(k)}_{n_{g_{(k)}},g_{k+1}} \Delta^{(k)}_{n_{g_{(k)}}}\right\}_{k\in\mathbb{N}}$ of decreasing order $n_{g_{(k)}}$ such that for $i, j = 0, 1, ..., n_{g_{(k)}} - 1$,

$$\begin{aligned}
\left(M^{(k)}_{n_{g_{(k)}},g_{k+1}}\right)_{ij} &= \delta^{(n_{g_{(k)}})}_{g_{k+1}i-j} \\
\left(\Delta^{(k)}_{n_{g_{(k)}}}\right)_{ii} &= d_{g[g_1(g_2(...(g_{k-1}(g_k i \,mod\, n_{g_{(k-1)}}) \,mod\, n_{g_{(k-2)}})...) \,mod\, n_{g_{(1)}}) \,mod\, n_{g_0}] \,mod\, n} \\
&\overset{(\beta)}{=} d_{g^{k+1} i \,mod\, n}
\end{aligned} \tag{3.61}$$

Indeed: for equality (β):

$$\begin{aligned}
g[g_1(...(g_{k-1}(g_k i \,mod\, n_{g_{(k-1)}}) \,mod\, n_{g_{(k-2)}})...) \,mod\, n_{g_{(0)}}] \,mod\, n &= \\
g[g_1(g_2(...(g^2_{k-1} i \,mod\, n_{g_{(k-2)}})...) \,mod\, n_{g_{(1)}}) \,mod\, n_{g_{(0)}}] \,mod\, n &= \\
g[g_1(g_2(...(g^3_{k-2} i \,mod\, n_{g_{(k-3)}})...) \,mod\, n_{g_{(1)}}) \,mod\, n_{g_{(0)}}] \,mod\, n &= \\
\vdots & \\
= g[g^k_1 i \,mod\, n_{g_{(0)}}] \,mod\, n & \\
= g^{k+1} i \,mod\, n &
\end{aligned}$$

since, for $j = 1, ..., k$, $g_j = g_{j-1} - m_j . n_{g_{(j-1)}}$, $m_j \in \mathbb{Z}$.

□

Lemma 3.4.12. *Let $s \in \mathbb{N}$ be the first index associated with the sequences constructed in Lemma 3.4.11 such that $\delta^{(s)} \in \{1, n_{g_{(s)}}\}$. Then*

1. *If $\delta^{(s)} = n_{g_{(s)}}$, then $g_{s+1} = 0$. Hence*

$$\begin{aligned}
Eig\,(C_{n,g}) &= Eig\left(M^{(s)}_{n_{g_{(s)}},0} \Delta^{(s)}_{n_{g_{(s)}}}\right) \cup \{0 :\ mult. = n - n_{g_{(s)}}\} \\
&= \{d_0,\, 0 :\ mult. = n - 1\}
\end{aligned}$$

2. *For $\delta^{(s)} = 1$, it holds*

$$\begin{aligned}
Eig\,(C_{n,g}) &= Eig\left(M^{(s)}_{n_{g_{(s)}},g_{s+1}} \Delta^{(s-1)}_{n_{g_{(s)}}}\right) \cup \{0 :\ mult. = n - n_{g_{(s)}}\} \\
&= \left\{0,\, e^{\hat{i}\frac{2k(j)\pi}{\varphi\left(n_{g_{(s)}}\right)}} \prod_{h=0}^{\varphi\left(n_{g_{(s)}}\right)-1} f_{g^h_{s+1} k_s(p(j)) \,mod\, n};\ \ j = 0, 1, ..., n_{g_{(s)}} - 1\right\},
\end{aligned}$$

"0" *is of multiplicity*$=n-n_{g_{(s)}}$. *where* $k_s(j)=g^{s+1}j \bmod n$, $p(j)\in\{0,1,...,n_{g_{(s)}}-1\}$, $k(j)$ *is an element of* $\left\{0,1,...,\varphi\left(n_{g_{(s)}}\right)-1\right\}$, *and*

$$f_{g_{s+1}^h k_s(p(j)) \bmod n}=L_{g_{s+1}^h k_s(p(j)) \bmod n}^{\frac{1}{\varphi\left(n_{g_{(s)}}\right)}} e^{\hat{i}\frac{\theta_{g_{s+1}^h k_s(p(j)) \bmod n}}{\varphi\left(n_{g_{(s)}}\right)}}$$

with $d_j=(D_n)_{j,j}=L_j e^{\hat{i}\theta_j}$.

Proof. **Case 1:** $\delta^{(s)}=n_{g_{(s)}}$.

Since $\delta^{(s)}=n_{g_{(s)}}$ then $g_{s+1}=0$, so

$$M^{(s)}_{n_{g_{(s)}},0}\Delta^{(s)}_{n_{g_{(s)}}}=\begin{bmatrix} d_0 & 0 & \dots & 0\\ d_0 & \vdots & & \vdots\\ \vdots & \vdots & & \vdots\\ d_0 & 0 & \dots & 0\end{bmatrix}$$

It follows from Lemmas 3.4.5-3.4.6 that

$$\begin{aligned} Eig\,(C_{n,g}) &= Eig\left(M^{(0)}_{n_{g_{(0)}},g_1}\Delta^{(0)}_{n_{g_{(0)}}}\right)\cup\{0:\ mult.=n-n_g\}\\ &= Eig\left(M^{(1)}_{n_{g_{(1)}},g_2}\Delta^{(1)}_{n_{g_{(1)}}}\right)\cup\{0:\ mult.=n_g-n_{g_{(1)}}\}\cup\{0:\ \text{mult.}=n-n_g\}\\ &= Eig\left(M^{(1)}_{n_{g_{(1)}},g_2}\Delta^{(1)}_{n_{g_{(1)}}}\right)\cup\{0:\ mult.=n-n_{g_{(1)}}\}\\ &\vdots\\ &= Eig\left(M^{(s)}_{n_{g_{(s)}},g_{s+1}}\Delta^{(s)}_{n_{g_{(s)}}}\right)\cup\{0:mult.:n_{g_{(s-1)}}-n_{g_{(s)}}\}\cup\{0:\ n-n_{g_{(s-1)}}\}\\ &= \{d_0,\,0:\ \text{mult.}=n-1\}\end{aligned}$$

Case 2: $\delta^{(s)}=1$.

There exists a positive integer $\varphi(n_{g_{(s)}})$ such that $g_{s+1}^{\varphi(n_{g_{(s)}})}=1\,mod\,n_{g_{(s)}}$. For $i,j=0,1,...,n_{g_{(s)}}-1$

$$\begin{aligned}\left(\Delta^{(s)}_{n_{g_{(s)}}}M^{(s)}_{n_{g_{(s)}},g_{s+1}}\right)_{ij} &= \delta^{(n_{g_{(s)}})}_{g_{s+1}i-j}d_{g[g_1(g_2(...(g_{s-1}(g_s i \bmod n_{g_{(s-1)}}) \bmod n_{g_{(s-2)}})...) \bmod n_{g_{(1)}}) \bmod n_{g_0}] \bmod n}\\ &= \delta^{(n_{g_{(s)}})}_{g_{s+1}i-j}d_{k_s(i)}.\end{aligned}$$

where

$$k_s(i)=g^{s+1}i\,mod\,n.$$

Then

$$Eig\,(C_{n,g}) = Eig\left(M^{(s)}_{n_{g_{(s)}},g_{s+1}}\Delta^{(s)}_{n_{g_{(s)}}}\right)\cup\{0:\ \text{mult.}=n-n_{g_{(s)}}\}$$

Since $(n_{g_{(s)}},g_{s+1})=1$, according to Lemma 3.4.10, one deduces by replacing n_g by $n_{g_{(s)}}$, g_1 by g_{s+1}, $p(j)$ by $k_s(p(j))$, and $\varphi(n_g)$ by $\varphi(n_{g_{(s)}})$ that

$$Eig(C_{n,g})=\left\{e^{\hat{i}\frac{2k(j)\pi}{\varphi(n_{g_{(s)}})}}\prod_{h=0}^{\varphi(n_{g_{(s)}})-1}f_{g_{s+1}^h k_s(p(j)) \bmod n}:\ j=0,1,...,n_{g_{(s)}}-1\right\}\cup\{0:\text{mult.}=n-n_{g_{(s)}}\}$$

with

$$f_{g_{s+1}^h k_s(p(j)) \bmod n} = L_{g_{s+1}^h k_s(p(j)) \bmod n}^{\frac{1}{\varphi(n_{g_{(s)}})}} \cdot e^{\hat{i}\frac{\theta_{g_{s+1}^h k_s(p(j)) \bmod n}}{\varphi(n_{g_{(s)}})}}$$

Furthermore, $p(j)$ is an element of the set $\{0,1,...,n_{g_{(s)}}-1\}$ and $k(j)$ belongs to the set $\{0,1,...,\varphi(n_{g_{(s)}})-1\}$. □

In this way, we obtain a much simplified iterative method to determine the eigenvalues of the g-circulant matrices $C_{n,g}$.

Algorithm

Initialization: Given a positive integer g, the Fourier matrix F_n and the matrix $Z_{n,g} = [\delta_{r-gs}]_{r,s=0}^{n-1}$, determine the matrix $M_{n,g} = F_n Z_{n,g} F_n^\star := [\delta_{gr-s}]_{r,s=0}^{n-1}$. Set $n_{g_{(-1)}} := n$; $\delta^{(-1)} := gcd(n,g)$; $g_{(0)} = g_0 := g$; $n_g := n_{g_{(0)}} := \frac{n}{\delta^{(-1)}}$; $M_{n_{g_{(-1)}},g_0}^{(-1)} := M_{n,g}$; $\Delta_{n_{g_{(-1)}}}^{(-1)} := D_n = diag(d_j : j = 0,1,2,...,n-1)$; $\delta_s^{(n)} := \delta_s$. Put $k := 0$.

(1) If $\delta^{(k-1)} = 1$,

i. compute as in Lemma 3.4.7 the matrix

$$\left(M_{n_{g_{(k-1)}},g_k}^{(k-1)} \cdot \Delta_{n_{g_{(k-1)}}}^{(k-1)}\right)^{\varphi\left(n_{g_{(k-1)}}\right)} := diag\left(\prod_{p=0}^{\varphi\left(n_{g_{(k-1)}}\right)-1} d_{l_j^p},\ j = 0,1,...,n_{g_{(k-1)}}-1\right)$$

where $l_j^p := g^p j \bmod n_{g_{(k-1)}}$,

ii. For fixed $j \in \{0,1,...,n_{g_{(k-1)}}-1\}$, solve the equation

$$z^{\varphi\left(n_{g_{(k-1)}}\right)} = \prod_{p=0}^{\varphi\left(n_{g_{(k-1)}}\right)-1} d_{l_j^p}$$

iii. Then the spectrum of $C_{n,g}$ is

$$\begin{aligned} Eig(C_{n,g}) &:= Eig\left(M_{n_{g_{(k-1)}},g_k}^{(k-1)} \cdot \Delta_{n_{g_{(k-1)}}}^{(k-1)}\right) \cup \{0\} \\ &:= \left\{0,\ e^{\hat{i}\frac{2k(j)\pi}{\varphi(n_{g_{(k-1)}})}} \prod_{h=0}^{\varphi(n_{g_{(k-1)}})-1} f_{g_k^h R_{k-1}(p(j)) mod\, n_{g_{(k-1)}}} : j = 0,1,...,n_{g_{(k-1)}}-1\right\} \end{aligned}$$

"0" is of multiplicity$= n - n_{g_{(k-1)}}$; $R_{k-1}(j) := g^k j \bmod n$; $d_j := L_j e^{\hat{i}\theta_j}$ with $|d_j| = L_j$;

$$f_{g_k^h R_{k-1}(p(j))(\bmod n_{g_{(k-1)}})} := L_{g_k^h R_{k-1}(p(j))(\bmod n_{g_{(k-1)}})}^{\frac{1}{\varphi\left(n_{g_{(k-1)}}\right)}} \cdot e^{\hat{i}\frac{\theta_{g_k^h R_{k-1}(p(j)) \bmod n_{g_{(k-1)}}}}{\varphi\left(n_{g_{(k-1)}}\right)}}$$

where $\varphi(a)$ denotes the Euler indicator associated with the positive integer a; $p(j) \in \{0,1,...,n_{g_{(k-1)}}-1\}$ and $k(j) \in \left\{0,1,...,\varphi\left(n_{g_{(k-1)}}\right)-1\right\}$; **stop**.

Otherwise,

(2) If $\delta^{(k-1)} := n_{g_{(k-1)}}$,

i. Compute:
$g_k := 0$;

$$Eig\left(M^{(k-1)}_{n_{g_{(k-1)}},0} \cdot \Delta^{(k-1)}_{n_{g_{(k-1)}}}\right) := \{d_0, 0 : mult. = n_{g_{(k-1)}} - 1\}$$

ii. Hence the spectrum of $C_{n,g}$ is

$$\begin{aligned} Eig(C_{n,g}) &:= Eig\left(M^{(k-1)}_{n_{g_{(k-1)}},0} \cdot \Delta^{(k-1)}_{n_{g_{(k-1)}}}\right) \cup \{0 : mult. = n - n_{g_{(k-1)}}\} \\ &:= \{d_0, 0 : mult. = n - 1\} \end{aligned}$$

stop.

Otherwise,

(3) Put $k := k + 1$, compute:

$$n_{g_{(k-1)}} := \frac{n_{g_{(k-2)}}}{\delta(k-2)};\ \ g_k := g_{k-1}\, mod\, n_{g_{(k-1)}};\ \ \delta^{(k-1)} := gcd(n_{g_{(k-1)}}, g_k);$$

and

$$\Delta^{(k-1)}_{n_{g_{(k-1)}}} := diag\left(d_{g^k j \bmod n} :\ j = 0, 1, ..., n_{g_{(k-1)}} - 1\right);$$

$$M^{(k-1)}_{n_{g_{(k-1)}},g_k} := \left[\delta^{(n_{g_{(k-1)}})}_{g_k r - s}\right]^{n_{g_{(k-1)}}-1}_{r,s=0},\ \ \delta^{(n_{g_{(k-1)}})}_q := \begin{cases} 1 & \text{if } q \equiv 0\, mod\, n_{g_{(k-1)}} \\ 0 & \text{otherwise} \end{cases}$$

Remark 3.4.3. *The above algorithm determines of recursive way all the eigenvalues of the g-circulant matrices $C_{n,g}$. It can stop in step (1) if the positive integers n and g are coprime, in that we obtain the same values as those determined by William F. Trench in* [161(1)]. *Next, the algorithm can stop in step (2) if $g = 0$.*

In the following we present some examples of eigenvalues of the g-circulant matrices when $g \in \mathbb{N}^d$ and some of the entries of g vanish.

3.5 Examples of g-circulant matrices when some of the entries of g vanish

We begin this section with a brief digression on multilevel matrices. A d-level matrix A of dimension $\hat{n} \times \hat{n}$ with $n = (n_1, n_2, ..., n_d)$ and $\hat{n} = n_1 n_2 ... n_d$ can be viewed as a matrix of dimension $n_1 \times n_1$ in which each element is a block of dimension $n_2 n_3 ... n_d \times n_2 n_3 ... n_d$; in turn, each block of dimension $n_2 n_3 ... n_d \times n_2 n_3 ... n_d$ can be viewed as a matrix of dimension $n_2 \times n_2$ in which each element is a block of dimension $n_3 n_4 ... n_d \times n_3 n_4 ... n_d$, and so on. So we can say that n_1 is the most "outer" dimension of the matrix A and n_d is the "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 $P^T AP$ becomes a matrix again of dimension $\hat{n} \times \hat{n}$ but with $n = (n_{p(1)}, n_{p(2)}, ..., n_{p(d)})$ and $\hat{n} = n_{p(1)} n_{p(2)} ... n_{p(d)} = n_1 n_2 ... n_d$ (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 d-level matrices, especially when some of the entries of the vector g are zero; indeed: as we observed in subsection 3.2.1, if $g = \underline{0}$, the d-level g-circulant matrix A 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, $g = (0, ..., 0, g_k, ..., g_d)$, the matrix $P^T AP$ (where P is the permutation matrix associated with p) becomes a block matrix with constant blocks on each row and with blocks of dimension $n_k n_{k+1} ... n_d \times n_k n_{k+1} ... n_d$; with this "new" structure, formula (3.7) is even more intuitively understandable, as we shall see later in the examples.

Lemma 3.5.1. *Let A be a 2-level circulant matrix of dimension $\hat{n} \times \hat{n}$ with $n = (n_1, n_2)$ and $\hat{n} = n_1 n_2$,*

$$A = \left[[a_{(j_1 - k_1, j_2 - k_2) mod\, n}]_{j_2,k_2=0}^{n_2-1} \right]_{j_1,k_1=0}^{n_1-1}.$$

There exists a permutation matrix P such that

$$P^T AP = \left[[a_{(j_1 - k_1, j_2 - k_2) mod\, n}]_{j_1,k_1=0}^{n_1-1} \right]_{j_2,k_2=0}^{n_2-1}$$

Moreover, one has the following Corollary

Corollary 3.5.1. *Let A be a d-level circulant matrix of dimension $\hat{n} \times \hat{n}$ with $n = (n_1, n_2, ..., n_d)$ and $\hat{n} = n_1 n_2 ... n_d$,*

$$A = \left[\left[... [a_{(j_1 - k_1, j_2 - k_2, ..., j_d - k_d) mod\, n}]_{j_d,k_d=0}^{n_d-1} \cdots \right]_{j_2,k_2=0}^{n_2-1} \right]_{j_1,k_1=0}^{n_1-1}.$$

For every permutation p of d elements, there exists a permutation matrix P such that

$$P^T AP = \left[\left[... [a_{(j_1 - k_1, j_2 - k_2, ..., j_d - k_d) mod\, n}]_{j_{p(d)},k_{p(d)}=0}^{n_{p(d)}-1} \cdots \right]_{j_{p(2)},k_{p(2)}=0}^{n_{p(2)}-1} \right]_{j_{p(1)},k_{p(1)}=0}^{n_{p(1)}-1}.$$

Remark 3.5.1. *Lemma* 3.5.1 *and Corollary* 3.5.1 *also apply to d-level g-circulant matrices.*

Now, let $g = (g_1, g_2, ...g_d)$ be a d-dimensional vector of nonnegative 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 $g_{p(1)} = g_{p(2)} = ... = g_{p(t)} = 0$, (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 formula (3.7) remains the same for the matrix $P^T AP$ (where P is the permutation matrix associated with p) but with $n[0] = (n_{p(1)}, n_{p(2)}, ..., n_{p(t)})$, and where C_j is a d^+-level g^+-circulant matrix, with $g^+ = (g_{p(t+1)}, g_{p(t+2)}, ..., g_{p(d)})$, of partial size $n[> 0] = (n_{p(t+1)}, n_{p(t+2)}, ..., n_{p(d)})$, and whose expression is

$$C_j = \left[\left[... [a_{(r - g \circ s) mod\, n}]_{r_{p(d)},s_{p(d)}=0}^{n_{p(d)}-1} \cdots \right]_{r_{p(t+2)},s_{p(t+2)}=0}^{n_{p(t+2)}-1} \right]_{r_{p(t+1)},s_{p(t+1)}=0}^{n_{p(t+1)}-1},$$

with $(r_{p(1)}, r_{p(2)}, ..., r_{p(t)}) = j$. Obviously $Sval(A) = Sval(P^T AP)$.

We recall that if B is a matrix of size $n \times n$ positive semidefinite, that is $B^\star = B$ and $x^\star Bx \geq 0 \ \forall x \neq 0$, then $Eig(B) = Sval(B)$. Moreover, if $B = U\Sigma U^\star$ is a SVD for B (which coincides with the Schur decomposition of B) with $\Sigma = \underset{j=1,...,n}{diag}\ (\sigma_j)$, then

$$B^{1/2} = U\Sigma^{1/2}U^\star, \tag{3.62}$$

where $\Sigma^{1/2} = \underset{j=1,...,n}{diag}\ (\sqrt{\sigma_j})$.

We proceed with a detailed example: a 3-level g-circulant matrix with $g = (g_1, g_2, g_3) = (1, 2, 0)$ 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 3.5.1. *Consider a 3-level g-circulant matrix A where $g = (g_1, g_2, g_3) = (1, 2, 0)$*

$$\begin{aligned} A &= \left[\left[\left[a_{((r_1-1\cdot s_1) mod\, n_1,(r_2-2\cdot s_2) mod\, n_2,(r_3-0.s_3) mod\, n_3)}\right]_{r_3,s_3=0}^{n_3-1}\right]_{r_2,s_2=0}^{n_2-1}\right]_{r_1,s_1=0}^{n_1-1} \\ &= \left[\left[\left[a_{(r_1-s_1) mod\, n_1,(r_2-2s_2) mod\, n_2,r_3}\right]_{r_3=0}^{n_3-1}\right]_{r_2,s_2=0}^{n_2-1}\right]_{r_1,s_1=0}^{n_1-1}. \end{aligned}$$

If we choose the permutation p of 3 elements such that

$$\begin{aligned} (p(1), p(2), p(3)) &= (3, 2, 1), \\ (g_{p(1)}, g_{p(2)}, g_{p(3)}) &= (0, 2, 1), \\ (n_{p(1)}, n_{p(2)}, n_{p(3)}) &= (n_3, n_2, n_1), \end{aligned}$$

and if we take the permutation matrix P related to p, then

$$P^T A P \equiv \hat{A} = \left[\left[\left[a_{(r_1-s_1) mod\, n_1,(r_2-2s_2) mod\, n_2,r_3}\right]_{r_1,s_1=0}^{n_1-1}\right]_{r_2,s_2=0}^{n_2-1}\right]_{r_3=0}^{n_3-1}.$$

Now, for $r_3 = 0, 1, ..., n_3 - 1$, let us set

$$C_{r_3} = \left[\left[a_{(r_1-s_1) mod\, n_1,(r_2-2s_2) mod\, n_2,r_3}\right]_{r_1,s_1=0}^{n_1-1}\right]_{r_2,s_2=0}^{n_2-1}.$$

As a consequence, 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

$$\hat{A} = \begin{bmatrix} C_0 & C_0 & \dots & C_0 \\ C_1 & C_1 & \dots & C_1 \\ \vdots & \vdots & \vdots & \vdots \\ C_{n_3-1} & C_{n_3-1} & \dots & C_{n_3-1} \end{bmatrix}.$$

and this is a block matrix with constant blocks on each row. From formula (3.1), the singular values of A are the square root of the eigenvalues of $A^\star A$:

$$\begin{aligned} \hat{A}_n^\star \hat{A}_n &= \begin{bmatrix} C_0^\star & C_1^\star & \dots & C_{n_3-1}^\star \\ C_0^\star & C_1^\star & \dots & C_{n_3-1}^\star \\ \vdots & & \vdots & \vdots \\ C_0^\star & C_1^\star & \dots & C_{n_3-1}^\star \end{bmatrix} \begin{bmatrix} C_0 & C_0 & \dots & C_0 \\ C_1 & C_1 & \dots & C_1 \\ \vdots & & \vdots & \\ C_{n_3-1} & C_{n_3-1} & \dots & C_{n_3-1} \end{bmatrix} \\ &= \begin{bmatrix} \sum_{j=0}^{n_3-1} C_j^\star C_j & \sum_{j=0}^{n_3-1} C_j^\star C_j & \dots & \sum_{j=0}^{n_3-1} C_j^\star C_j \\ \sum_{j=0}^{n_3-1} C_j^\star C_j & \sum_{j=0}^{n_3-1} C_j^\star C_j & \dots & \sum_{j=0}^{n_3-1} C_j^\star C_j \\ \vdots & \vdots & \vdots & \vdots \\ \sum_{j=0}^{n_3-1} C_j^\star C_j & \sum_{j=0}^{n_3-1} C_j^\star C_j & \dots & \sum_{j=0}^{n_3-1} C_j^\star C_j \end{bmatrix} \\ &= \begin{bmatrix} 1 & \dots & 1 \\ \vdots & \vdots & \vdots \\ 1 & \dots & 1 \end{bmatrix} \otimes \sum_{j=0}^{n_3-1} C_j^\star C_j \\ &= J_{n_3} \otimes \sum_{j=0}^{n_3-1} C_j^\star C_j. \end{aligned}$$

Therefore

$$Eig(\hat{A}^{\star}\hat{A}) = Eig\left(J_{n_3} \otimes \sum_{j=0}^{n_3-1} C_j^{\star} C_j\right), \tag{3.63}$$

where

$$Eig(J_{n_3}) = \{0, n_3\}, \tag{3.64}$$

because J_{n_3} is a matrix of rank 1, so it has all eigenvalues equal to zero except one eigenvalue equal to $trace(J_{n_3}) = n_3$ ($trace(J_{n_3})$ is the trace of the matrix J_{n_3}). If we put

$$\lambda_k = \lambda_k\left(\sum_{j=0}^{n_3-1} C_j^{\star} C_j\right), \quad k = 0, 1, ..., n_2 n_1 - 1,$$

by exploiting basic properties of the tensor product and taking into consideration (3.63) *and* (3.64) *we find*

$$\lambda_k(\hat{A}^{\star}\hat{A}) = n_3\lambda_k, \quad k = 0, 1, ..., n_1 n_2 - 1, \tag{3.65}$$

$$\lambda_k(\hat{A}^{\star}\hat{A}) = 0, \quad k = n_1 n_2, n_1 n_2 + 1, ..., n_1 n_2 n_3 - 1. \tag{3.66}$$

From (3.65), (3.66) *and* (3.1), *and recalling that $Sval(\hat{A}) = Sval(A)$, one obtains that the singular values of A are given by*

$$\sigma_k(A) = \sqrt{n_3\lambda_k}, \quad k = 0, 1, ..., n_1 n_2 - 1,$$

$$\sigma_k(A) = 0, \quad k = n_1 n_2, n_1 n_2 + 1, ..., n_3 n_2 n_1 - 1.$$

and, since $\sum_{j=0}^{n_3-1} C_j^{\star} C_j$ is a positive semidefinite matrix, from (3.62) *we can write*

$$\sigma_k(A) = \sqrt{n_3}\widetilde{\sigma}_k, \quad k = 0, 1, ..., n_1 n_2 - 1,$$

$$\sigma_k(A) = 0, \quad k = n_1 n_2, n_1 n_2 + 1, ..., n_3 n_2 n_1 - 1.$$

where $\widetilde{\sigma}_k$ denotes the generic singular values of $\left(\sum_{j=0}^{n_3-1} C_j^{\star} C_j\right)^{1/2}$.

Example 3.5.2. *Let us see what happens when the vector g has only one component different from zero. Let $n = (n_1, n_2, ..., n_d)$ and $g = (0, ..., 0, g_k, 0..., 0)$, $g_k > 0$; 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 $g = (0, ..., 0, g_d)$ (will all zero components in top positions, otherwise we use permutation). From subsection* 3.2.2, *the singular values of $A_n = [a_{(r-g \circ s) mod\, n}]_{r,s=\underline{0}}^{n-e}$ are zero except for few of them given by $\sqrt{\hat{n}[0]}\sigma$ where, in our case, $\hat{n}[0] = n_1 n_2 ... n_{d-1}$, $n[0] = (n_1, n_2, ..., n_{d_{-1}})$, and σ is any singular value of the matrix*

$$\left(\sum_{j=\underline{0}}^{n[0]-e} C_j^{\star} C_j\right)^{1/2},$$

where C_j is an g_d-circulant matrix of dimension $n_d \times n_d$ whose expression is

$$\begin{aligned} C_j = [a_{(r-g\circ s) mod\, n}]_{r_d,s_d=0}^{n_d-1} &= [a_{(r_1,r_2,...,r_{d-1},(r_d-g_d s_d) mod\, n_d)}]_{r_d,s_d=0}^{n_d-1} \\ &= [a_{(j,(r_d-g_d s_d) mod\, n_d)}]_{r_d,s_d=0}^{n_d-1}, \end{aligned}$$

with $j = (r_1, r_2, ..., r_{d-1})$. For $j = \underline{0}, ..., n[0] - e$, if $C_{n_d}^{(j)}$ is the circulant matrix which has as its first column the vector $a^{(j)} = [a_{(j,0)}, a_{(j,1)}, ..., a_{(j,n_d-1)}]^T$ (which is the first column of the matrix C_j,) $C_{n_d}^{(j)} = [a_{(j,(r-s) mod\, n_d)}]_{r,s=0}^{n_d-1} = F_{n_d} D_{n_d}^{(j)} F_{n_d}^{\star}$, with $D_{n_d}^{(j)} = diag(\sqrt{n_d} F_{n_d}^{\star} a^{(j)})$, then, from (3.28), (3.8) *and* (3.14), *it is immediate to verify that*

$$\begin{aligned} \sum_{j=\underline{0}}^{n[0]-e} C_j^{\star} C_j &= \sum_{j=\underline{0}}^{n[0]-e} (F_{n_d} D_{n_d}^{(j)} F_{n_d}^{\star} Z_{n_d,g_d})^{\star} (F_{n_d} D_{n_d}^{(j)} F_{n_d}^{\star} Z_{n_d,g_d}) \\ &= \sum_{j=\underline{0}}^{n[0]-e} (F_{n_d}^{\star} Z_{n_d,g_d})^{\star} (D_{n_d}^{(j)})^{\star} D_{n_d}^{(j)} (F_{n_d}^{\star} Z_{n_d,g_d}) \\ &= (F_{n_d}^{\star} Z_{n_d,g_d})^{\star} \left(\sum_{j=\underline{0}}^{n[0]-e} (D_{n_d}^{(j)})^{\star} D_{n_d}^{(j)} \right) (F_{n_d}^{\star} Z_{n_d,g_d}). \end{aligned}$$

Now, if we put
$n_{d,g} = \frac{n_d}{(n_d,g_d)}$ *and* $q_s^{(j)} = |D_{n_d}^{(j)}|_{s,s}^2 = (D_{n_d}^{(j)})_{s,s} \cdot \overline{(D_{n_d}^{(j)})_{s,s}}$, $s = 0, 1, ..., n_d - 1$,

$$\Delta_l = \begin{bmatrix} \sum_{j=\underline{0}}^{n[0]-e} q_{(l-1)n_{d,g}}^{(j)} & & & \\ & \sum_{j=\underline{0}}^{n[0]-e} q_{(l-1)n_{d,g}+1}^{(j)} & & \\ & & \ddots & \\ & & & \sum_{j=\underline{0}}^{n[0]-e} q_{(l-1)n_{d,g}+n_{d,g}-1}^{(j)} \end{bmatrix} \in \mathbb{C}^{n_{d,g} \times n_{d,g}},$$

for $l = 1, 2, ..., (n_d, g_d)$, then, following the same reasoning employed for proving formula (3.29), *we infer*

$$Eig\left(\sum_{j=\underline{0}}^{n[0]-e} C_j^{\star} C_j \right) = \frac{1}{(n_d, g_d)} Eig\left(J_{(n_d,g_d)} \otimes \sum_{l=1}^{(n_d,g_d)} \Delta_l \right),$$

where

$$J_{(n_d,g_d)} = \underbrace{\begin{bmatrix} 1 & 1 & \dots & 1 \\ 1 & 1 & \dots & 1 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & \dots & 1 \end{bmatrix}}_{(n_d,g_d)\, times},$$

$$\frac{1}{(n_d, g_d)} J_{(n_d,g_d)} = \{0; 1\},$$

and

$$\begin{aligned}\sum_{l=1}^{(n_d,g_d)} \Delta_l &= \sum_{l=1}^{(n_d,g_d)} diag\left(\sum_{j=\underline{0}}^{n[0]-e} q^{(j)}_{(l-1)n_{d,g}+k} : k=0,1,...,n_{d,g}-1\right)\\ &= diag\left(\sum_{l=1}^{(n_d,g_d)n[0]-e}\ \sum_{j=\underline{0}} q^{(j)}_{(l-1)n_{d,g}+k} : k=0,1,...,n_{d,g}-1\right).\end{aligned}$$

Consequently, since $\sum_{l=1}^{(n_d,g_d)} \Delta_l$ *is a diagonal matrix, and by exploiting basic properties of the tensor product, we find*

$$\begin{aligned}\lambda_k\left(\sum_{j=\underline{0}}^{n[0]-e} C_j^\star C_j\right) &= \sum_{l=1}^{(n_d,g_d)n[0]-e}\ \sum_{j=\underline{0}} q^{(j)}_{(l-1)n_{d,g}+k}, \quad k=0,1,...,n_{d,g}-1,\\ \lambda_k\left(\sum_{j=\underline{0}}^{n[0]-e} C_j^\star C_j\right) &= 0, \qquad k=n_{d,g},...,n_d-1.\end{aligned}$$

Now, since $\sum_{j=\underline{0}}^{n[0]-e} C_j^\star C_j$ *is a positive semidefinite matrix, from* (3.62) *we finally have*

$$\begin{aligned}\sigma_k\left(\left(\sum_{j=\underline{0}}^{n[0]-e} C_j^\star C_j\right)^{1/2}\right) &= \sqrt{\sum_{l=1}^{(n_d,g_d)n[0]-e}\ \sum_{j=\underline{0}} q^{(j)}_{(l-1)n_{d,g}+k}}, \quad k=0,1,...,n_{d,g}-1,\\ \sigma_k\left(\left(\sum_{j=\underline{0}}^{n[0]-e} C_j^\star C_j\right)^{1/2}\right) &= 0, \qquad k=n_{d,g},...,n_d-1.\end{aligned}$$

Conclusion

In this chapter We have studied in detail the singular values of g-circulant matrices and have provided a powerful technique for determining the eigenvalues of these matrices. The generalization to the multilevel block setting has been sketched in the case of singular values. The next chapter will treat the asymptotic distribution result of g-Toeplitz sequences associated with a given integrable symbol.

Chapter Four

Singular Value Distribution of g-Toeplitz Sequences

4.1 Introduction

A matrix A_n of size n is called g-Toeplitz if its entries obey the rule $A_n = [a_{r-gs}]_{r,s=0}^{n-1}$, where g is a nonnegative integer. As example, if $n = 5$ and $g = 3$ then

$$A_n \equiv T_{n,g} = \begin{bmatrix} a_0 & a_{-3} & a_{-6} & a_{-9} & a_{-12} \\ a_1 & a_{-2} & a_{-5} & a_{-8} & a_{-11} \\ a_2 & a_{-1} & a_{-4} & a_{-7} & a_{-10} \\ a_3 & a_0 & a_{-3} & a_{-6} & a_{-9} \\ a_4 & a_1 & a_{-2} & a_{-5} & a_{-8} \end{bmatrix}$$

We recall that such kind of matrices arises in wavelet analysis [50] and in the refinement equations associated with the subdivision algorithm, see [58] and references therein. In addition, Gilbert Strang [150] has found interesting relationships between dilation equations in the wavelets context and multigrid methods [78], [162], for the restriction/prolongation operators [61], [1] with various boundary conditions. Moreover, boundary conditions analysis naturally arises when dealing with signal/image restoration problems or differential equations, see [129], [126].

In this chapter we address the problem of characterizing an asymptotic analysis of the distribution results for the singular values of g-Toeplitz sequences, in the case where the sequence of values $\{a_k\}_k$, defining the entries of the matrices, can be interpreted as the sequence of Fourier coefficients of an integrable function f over the domain $(-\pi, \pi)$. As a byproduct, we will show interesting relations with the analysis of convergence of multigrid methods given, e.g., in [141, 1]. Finally we generalized the analysis to the block, multilevel case, amounting to choose the symbol f multivariate, i.e., defined on the set $G = (-\pi, \pi)^d$ for some $d > 1$, and matrix valued, i.e., such that $f(x)$ is a matrix of given size $p \times q$.

4.2 General definitions and tools

For any $n \times n$ matrix A with eigenvalues $\lambda_j(A)$, $j = 1, 2, ..., n$, and for any $m \times n$ matrix B with singular values $\sigma_j(A)$, $j = 1, 2, ..., l$, $l = \min\{m, n\}$, we set

$$Eig(A) = \{\lambda_j(A) : \ j = 1, 2, ..., n\}, \ \ Sval(B) = \{\sigma_j(B) \ j = 1, 2, ..., l\}.$$

The matrix $B^\star B$ is positive semidefinite, since $x^\star(B^\star B)x = \|Bx\|_2^2 \geq 0$ for all $x \in \mathbb{C}^n$, with "$\star$" denoting the transpose conjugate operator. Moreover, it is clear that the eigenvalues $\lambda_1(B^\star B) \geq \lambda_2(B^\star B) \geq ... \geq \lambda_n(B^\star B) \geq 0$ are nonnegative and can therefore be written in the form

$$\lambda_j(B^\star B) = \sigma_j^2, \tag{4.1}$$

with $\sigma_j \geq 0$, $j = 1, 2, ..., n$. The numbers $\sigma_1 \geq \sigma_2 \geq ... \geq \sigma_l \geq 0$, $l = \min\{m, n\}$ are called "**singular values** of B", i.e., $\sigma_j = \sigma_j(B)$ and if $n > l$ then $\lambda_j(B^\star B) = 0$, $j = l+1, ..., n$. A more general statement is contained in the singular value decomposition theorem (see e.g. [11]).

For any function F defined on $\mathbb{R}_0^+$ and for any $m \times n$ matrix A, the symbol $\sum_\sigma(F, A)$ stands for the mean

$$\sum\nolimits_\sigma(F, A) := \frac{1}{\min\{m, n\}} \sum_{j=1}^{\min\{m,n\}} F(\sigma_j(A)) = \frac{1}{\min\{m, n\}} \sum_{\sigma \in Sval(A)} F(\sigma) \tag{4.2}$$

Throughout this chapter we speak also of matrix sequences $\{A_k\}_k$ 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 this chapter, we will not need the extra parameter k and we will consider simply matrix sequences of the form $\{A_n\}_n$.

Concerning the case of matrix-sequences an important notion is that of spectral distribution in eigenvalue or singular value sense, linking the collective behavior of the eigenvalues or singular values of all the matrices in the sequence to a given function (or to a measure). The notion 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 [16] where many classical results by the authors, Szegö, Avram, Parter, Widom Tyrtyshnikov, and many other can be found, and more recent results in [71, 93, 146, 167, 156, 157]). Here we report the definition of spectral distribution only in the singular value sense since our analysis is devoted to singular values. The case of eigenvalues will be the subject of future investigations.

Definition 4.2.1. *Let $\mathcal{C}_0(\mathbb{R}_0^+)$ be the set of continuous functions with bounded support defined over the nonnegative real numbers, d a positive integer, and θ a complex-valued measurable function defined on a set $G \subset \mathbb{R}^d$ of finite and positive Lebesgue measure $\mu(G)$. Here G will be often equal to $(-\pi, \pi)^d$ so that $e^{i\overline{G}} = \mathbb{T}^d$ with $\mathbb{T}$ denoting the complex unit circle. A matrix sequence $\{A_k\}_k$ is said to be distributed (in the sense of the singular values) as the pair (θ, G) or to have the distribution function θ ($\{A_k\}_k \sim_\sigma (\theta, G)$), if, $\forall F \in \mathcal{C}_0(\mathbb{R}_0^+)$, the following limit relation holds*

$$\lim_{k\to\infty} \Sigma_\sigma(F, A_k) = \frac{1}{\mu(G)} \int_G F(|\theta(t)|)dt, \;\; t = (t_1, \ldots, t_d). \tag{4.3}$$

When considering θ taking values in $\mathcal{M}_{pq}$, where $\mathcal{M}_{pq}$ is the space of $p \times q$ matrices with complex entries and a function is considered to be measurable if and only if the component functions are, we say that $\{A_k\}_k \sim_\sigma (\theta, G)$ when for every $F \in \mathcal{C}_0(\mathbb{R}_0^+)$ we have

$$\lim_{k\to\infty} \Sigma_\sigma(F, A_k) = \frac{1}{\mu(G)} \int_G \frac{\sum_{j=1}^{\min\{p,q\}} F(\sigma_j(\theta(t)))}{\min\{p, q\}} dt, \;\; t = (t_1, \ldots, t_d),$$

with $\sigma_j(\theta(t)) = \sqrt{\lambda_j(\theta(t)^\star\theta(t))} = \lambda_j\sqrt{\theta(t)^\star\theta(t)}$. Finally we say that two sequences $\{A_k\}_k$ and $\{B_k\}_k$ are equally distributed in the sense of singular values (σ) if, $\forall F \in \mathcal{C}_0(\mathbb{R}_0^+)$, we have

$$\lim_{k\to\infty} \left[\sum\nolimits_\sigma(F, B_k) - \sum\nolimits_\sigma(F, A_k)\right] = 0.$$

Here we are interested in explicit formula of the distribution results for g-Toeplitz sequences. Following what is known in the standard case of $g = 1$ (or $g = e$ in the multilevel setting), we need to link the coefficients of the g-Toeplitz sequence to a certain symbol.

Let f be a Lebesgue integrable function defined on $(-\pi,\pi)^d$ and taking values in $\mathcal{M}_{pq}$, for given positive integers p and q. Then, for d-indices $r=(r_1,r_2,...,r_d)$, $j=(j_1,j_2,...,j_d)$, $n=(n_1,n_2,...,n_d)$, $e=(1,1,...,1)$, $\underline{0}=(0,0,...,0)$, the Toeplitz matrix $T_n(f)$ of size $p\hat{n}\times q\hat{n}$, $\hat{n}=n_1.n_2...n_d$, is defined as follows

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

where $\tilde{f}_k$ are the Fourier coefficients of f defined by equation

$$\tilde{f}_j=\tilde{f}_{(j_1,...,j_d)}=\frac{1}{(2\pi)^d}\int_{[-\pi,\pi]^d}f(t_1,...,t_d)e^{-\hat{i}(j_1t_1+...+j_dt_d)}dt_1...dt_d,\quad \hat{i}^2=-1, \tag{4.4}$$

for integers j_l such that $-\infty<j_l<\infty$ for $1\le l\le d$. Since f is a matrix-valued function of d variables whose component functions are all integrable, then the $(j_1,j_2,...,j_d)$-th Fourier coefficient is considered to be the matrix whose (u,v)-th entry is the $(j_1,j_2,...,j_d)$-th Fourier coefficient of the function $(f(t_1,...,t_d))_{u,v}$.

According to this multi-index block notation, we can define general multi-level block g-Toeplitz matrices. Of course, in this multidimensional setting, g denotes a d-dimensional vector of nonnegative integers that is, $g=(g_1,g_2,...,g_d)$. In that case $A_n=[a_{r-g\circ s}]_{r,s=\underline{0}}^{n-e}$ where the $\circ$ operation is the componentwise Hadamard product between vector or matrices of the same size.

4.2.1 The extremal cases where $g=0$ or $g=e$, and the intermediate cases

We consider a d-level setting and we analyze in detail the case where $\underline{0}\le g\le e$ and with $\le$ denoting the componentwise partial ordering between real vectors. When g has at least a zero component, the analysis can be reduced to the positive one as studied in section 4.6.

$\underline{g=e}$

In the literature the only case deeply studied is the case of $g=e$ (standard shift in every level). For multilevel block Toeplitz sequences $\{T_n(f)\}$ generated by an integrable d variate and matrix valued symbol f the singular values are not explicitly known but we know the distribution in the sense of Definition 4.2.1; see [156]. More precisely we have

$$\{T_n(f)\}\sim_\sigma (f,Q^d),\quad Q=(-\pi,\pi). \tag{4.5}$$

$\underline{g=\underline{0}}$

The other extreme is represented by the case where g is the zero vector. Here the multilevel block g-Toeplitz is given by

$$A_n=[a_{r-0\circ s}]_{r,s=\underline{0}}^{n-e}=[a_r]_{r,s=\underline{0}}^{n-e}=\begin{bmatrix}a_{\underline{0}} & \dots & a_{\underline{0}}\\ \vdots & & \vdots\\ a_{n-e} & \dots & a_{n-e}\end{bmatrix}.$$

A simple computation shows that all the singular values are zero except for few of them given by $\sqrt{\hat{n}}\sigma$, where $\hat{n}=n_1.n_2...n_d$ and σ is any singular value of the matrix $\left(\sum_{j=\underline{0}}^{n-e}a_j^\star a_j\right)^{1/2}$. 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

$$\{A_n\}\sim_\sigma (0,G),$$

for any domain satisfying the requirements of Definition 4.2.1.

4.2.2 When some of the entries of g vanish

The content of this subsection reduces to the following remark: the case of a nonnegative g can be reduced to the case of a positive vector so that we are motivated to treat in detail the latter in section 4.3. Let g be a d-dimensional vector of nonnegative integers and let $\mathcal{N} \subset \{1, ..., d\}$ be the set of indices such that $j \in \mathcal{N}$ if and only if $g_j = 0$. Assume that $\mathcal{N}$ is nonempty, let $t \geq 1$ be its cardinality and $d^+ = d - t$. Then a simple calculation shows that the singular values of the corresponding g-Toeplitz matrix $A_n = [a_{(r-g\circ s)}]_{r,s=\underline{0}}^{n-e}$ are zero except for few of them given by $\sqrt{\hat{n}[0]}\sigma$ where

$$\hat{n}[0] = \prod_{j\in\mathcal{N}} n_j, \quad \hat{n}[0] = (n_{j_1}, n_{j_2}, ..., n_{j_t}), \quad \mathcal{N} = \{j_1, ..., j_t\},$$

and σ is any singular value of the matrix

$$\left(\sum_{j=\underline{0}}^{\hat{n}[0]-e} T_j^\star T_j \right)^{1/2}. \tag{4.6}$$

Here T_j is a d^+-level g^+-Toeplitz matrix with $g^+ = (g_{k_1}, g_{k_2}, ..., g_{k_{d^+}})$ and of partial sizes $n[>0] = (n_{k_1}, n_{k_2}, ..., n_{k_{d^+}})$, $\mathcal{N}^C = \{k_1, k_2, ..., k_{d^+}\}$, and whose expression is

$$T_j = [a_{(r-g\circ s)}]_{r',s'=\underline{0}}^{n[>0]-e},$$

where $(r - g \circ s)_k = j_k$ for $g_k = 0$ and $r_i^{'} = r_{k_i}$, $s_i^{'} = s_{k_i}$, $i = 1, ..., d^+$. Also in this case, since most of the singular values are identically zero, we infer that

$$\{A_n\} \sim_\sigma (0, G),$$

for any domain satisfying the requirements of Definition 4.2.1.

4.3 Singular values of g-Toeplitz matrices

For $p = q = 1$, we recall that the g-Toeplitz matrices of dimension $n \times n$ are defined as

$$T_{n,g} = [a_{r-gc}]_{r,c=0}^{n-1}, \tag{4.7}$$

where the quantities $r - gs$ are not reduced modulus n. In analogy with the case $g = 1$, the elements a_j are the Fourier coefficients of some function f in $L^1(Q)$, with $Q = (-\pi, \pi)$, i.e., $a_j = \tilde{f}_j$ as in (4.4) with $d = 1$. If we denote by T_n the classical Toeplitz matrix generated by the function $f \in L^1(Q)$, $T_n = [a_{r-c}]_{r,c=0}^{n-1}$, $a_j = \tilde{f}_j$ defined as in (4.4), and by $T_{n,g}$ the g-Toeplitz matrix generated by the same function, one verifies immediately for n and g generic that

$$T_{n,g} = [\hat{T}_{n,g} | \mathcal{T}_{n,g}] = [T_n \hat{Z}_{n,g} | \mathcal{T}_{n,g}], \tag{4.8}$$

where $\hat{T}_{n,g} \in \mathbb{C}^{n\times\mu_g}$ (with $\mu_g = \lceil \frac{n}{g} \rceil$) is the matrix $T_{n,g}$ defined in (4.7) by considering only the μ_g first columns, $\mathcal{T}_{n,g} \in \mathbb{C}^{n\times(n-\mu_g)}$ is the matrix $T_{n,g}$ defined in (4.7) by considering only the $n - \mu_g$ last columns, and $\hat{Z}_{n,g} \in \mathbb{C}^{n\times\mu_g}$ is the matrix defined by

(4.9)

$$\hat{Z}_{n,g} = [\delta_{r-gs}]; \quad r = 0, 1, ..., n-1, \quad s = 0, 1, ..., \mu_g - 1, \text{ where } \quad \delta_k = \begin{cases} 1 & \text{if } k \equiv 0 (mod\, n); \\ 0 & \text{otherwise.} \end{cases}$$

Proof. From relation (4.8). For $r = 0, 1, ..., n-1$ and $s = 0, 1, ..., \mu_g - 1$, one has

$$\begin{aligned}(\hat{T}_{n,g})_{r,s} &= (T_n)_{r,gs},\\ (\hat{Z}_{n,g})_{r,s} &= \delta_{r-gs},\end{aligned}$$

and

$$\begin{aligned}(T_n\hat{Z}_{n,g})_{r,s} &= \sum_{l=0}^{n-1}(T_n)_{r,l}(\hat{Z}_{n,g})_{l,s}\\ &= \sum_{l=0}^{n-1}\delta_{l-gs}(T_n)_{r,l}\\ &\underset{(a)}{=} (T_n)_{r,gs}\\ &= (\hat{T}_{n,g})_{r,s}\end{aligned}$$

where (a) follows from the fact that there exists a unique $l_0 \in \{0, 1, ..., n-1\}$ such that $l_0 - gs \equiv 0 (mod\, n)$, that is, $l_0 \equiv gs (mod\, n)$, and, since $0 \leq gs \leq n-1$, we obtain $l_0 = gs$. □

If we take the matrix $\hat{T}_{n,g}$ of size $n \times (\mu_g + 1)$, the relation (4.8) is no longer true. In reality, looking at the $(\mu_g + 1)$-th column of the g-Toeplitz matrix we observe the Fourier coefficients with indices which are not present (less or equal to $-n$) in the Toeplitz matrix T_n. More precisely,

$$(\hat{T}_{n,g})_{0,\mu_g} = a_{0-g\mu_g} = a_{-g\mu_g}, \quad \text{and} \quad -g\mu_g \leq -n.$$

It follows that μ_g is the maximum number of columns for which relation (4.8) is true.

4.3.1 Some preparatory results

We begin with some preliminary notations and definitions

Definition 4.3.1. *Suppose a sequence of matrices $\{A_n\}_n$ of size d_n is given. We say that $\{\{B_{n,m}\}_n : m \geq 0\}$, $B_{n,m}$ of size d_n, $m \in \mathbb{N}$, is an approximating class of sequences (a.c.s) for $\{A_n\}_n$ if, for all sufficiently large $m \in \mathbb{N}$, the following splitting holds:*

$$A_n = B_{n,m} + R_{n,m} + N_{n,m} \quad \text{for all } n > n_m, \tag{4.10}$$

with

$$Rank(R_{n,m}) \leq d_n c(m), \quad \|N_{n,m}\| \leq \omega(m), \tag{4.11}$$

with $\|\cdot\|$ is the spectral norm (largest singular value), n_m, $c(m)$ and $\omega(m)$ depend only on m and, moreover,

$$\lim_{m\to\infty}\omega(m) = 0, \quad \lim_{m\to\infty} c(m) = 0. \tag{4.12}$$

Proposition 4.3.1. [125]. *Let $\{d_n\}_n$ be an increasing sequence of natural numbers. Suppose a sequence formed by matrices $\{A_n\}_n$ of size d_n is given such that $\{\{B_{n,m}\}_n : m \geq 0\}$, $m \in \hat{\mathbb{N}} \subset \mathbb{N}$, $\sharp\hat{\mathbb{N}} = \infty$, is an a.c.s for $\{A_n\}_n$ in the sense of Definition 4.3.1. Suppose that $\{B_{n,m}\}_n \sim_\sigma (\theta_m, G)$ and that θ_m converges in measure to the measurable function θ over G. Then necessarily*

$$\{A_n\}_n \sim_\sigma (\theta, G), \tag{4.13}$$

(see Definition 4.2.1).

Proposition 4.3.2. [125, 128]. *If $\{A_n\}_n$ and $\{B_n\}_n$ are two sequences of matrices of strictly increasing dimension, such that $\{A_n\}_n \sim_\sigma (\theta, G)$ and $\{B_n\}_n \sim_\sigma (0, G)$, then*

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

Proposition 4.3.3. [125]. *Let $f, g \in L^1(Q^d)$, $Q = (-\pi, \pi)$, and let $\{T_n(f)\}_n$ and $\{T_n(g)\}_n$ be two sequences of Toeplitz matrices generated by f and g, respectively. The following distribution result is true*

$$\{T_n(f)T_n(g)\}_n \sim_\sigma (fg, Q^d).$$

Lemma 4.3.1. *Let f be a measurable complex-valued function on a set K, and consider the measurable function $\sqrt{|f|} : K \to \mathbb{R}^+$. Let $\{A_{n,m}\}_{n,m}$; with $A_{n,m} \in \mathbb{C}^{d_n \times d_n'}$ ($d_n' \leq d_n$) be a sequence of matrices of strictly increasing dimension: $d_n' < d_{n+1}'$, $d_n < d_{n+1}$. If the sequences of matrices $\{A_{n,m}^\star A_{n,m}\}_{n,m}$, with $A_{n,m}^\star A_{n,m} \in C^{d_n' \times d_n'}$, is distributed in the singular value sense as the function f over a suitable set $G \subset K$ in the sense of Definition 4.2.1, then the sequence $\{A_{n,m}\}_{n,m}$ is distributed in the singular value sense as the function $\sqrt{|f|}$ over the same set G.*

Proof. From the singular value decomposition (SVD), we can write $A_{n,m}$ as

$$A_{n,m} = U\Sigma V^\star = U \begin{bmatrix} \sigma_1 & & & \\ & \sigma_2 & & \\ & & \ddots & \\ & & & \sigma_{d_n'} \\ \hline & & 0 & \end{bmatrix} V^\star,$$

with U and V unitary matrices, $U \in \mathbb{C}^{d_n \times d_n}$, $V \in \mathbb{C}^{d_n' \times d_n'}$ and $\Sigma \in \mathbb{R}^{d_n \times d_n'}$, $\sigma_j \geq 0$; by multiplying $A_{n,m}^\star A_{n,m}$ we obtain

$$A_{n,m}^\star A_{n,m} = V\Sigma^T U^\star U\Sigma V^\star = V\Sigma^T\Sigma V^\star = V\Sigma^{(2)}V^\star = V \begin{bmatrix} \sigma_1^2 & & & \\ & \sigma_2^2 & & \\ & & \ddots & \\ & & & \sigma_{d_n'}^2 \\ \hline & & 0 & \end{bmatrix} V^\star, \tag{4.14}$$

with $\Sigma^{(2)} = \Sigma^T\Sigma \in \mathbb{C}^{d_n' \times d_n'}$. We observe that (4.14) is an SVD for $A_{n,m}^\star A_{n,m}$, that is, the singular values $\sigma_j(A_{n,m}^\star A_{n,m})$ of $A_{n,m}^\star A_{n,m}$ are the square of singular values $\sigma_j(A_{n,m})$ of $A_{n,m}$. Since $\{A_{n,m}^\star A_{n,m}\} \sim_\sigma (f, G)$, by definition it holds that for every $F \in \mathcal{C}_0(\mathbb{R}_0^+)$

$$\lim_{n\to\infty} \frac{1}{d_n'} \sum_{i=1}^{d_n'} F(\sigma_i(A_{n,m}^\star A_{n,m})) = \frac{1}{\mu(G)} \int_G F(|f(t)|)dt = \frac{1}{\mu(G)} \int_G H\left(\sqrt{|f(t)|}\right) dt, \tag{4.15}$$

where H is such that $F = H \circ \sqrt{\cdot}$, but, owing to $\sigma_j(A_{n,m}) = \sqrt{\sigma_j(A_{n,m}^\star A_{n,m})}$ we obtain

$$\lim_{n\to\infty} \frac{1}{d_n'} \sum_{i=1}^{d_n'} F(\sigma_i(A_{n,m}^\star A_{n,m})) = \lim_{n\to\infty} \frac{1}{d_n'} \sum_{i=1}^{d_n'} F(\sigma_i^2(A_{n,m})) = \lim_{n\to\infty} \frac{1}{d_n'} \sum_{i=1}^{d_n'} H(\sigma_i(A_{n,m})). \tag{4.16}$$

From (4.15) and (4.16), one deduces that

$$\lim_{n\to\infty} \frac{1}{d_n'} \sum_{i=1}^{d_n'} H(\sigma_i(A_{n,m})) = \frac{1}{\mu(G)} \int_G H\left(\sqrt{|f(t)|}\right) dt, \tag{4.17}$$

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

Lemma 4.3.2. *Let $\{A_n\}_n$ and $\{Q_n\}_n$ be two sequences of matrices of strictly increasing dimension ($A_n, Q_n \in \mathbb{C}^{d_n \times d_n}$, $d_n < d_{n+1}$), where Q_n are all unitary matrices ($Q_n^\star Q_n = I$). If $\{A_n\}_n \sim_\sigma (0, G)$ then $\{A_n Q_n\}_n \sim_\sigma (0, G)$ and $\{Q_n A_n\}_n \sim_\sigma (0, G)$.*

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

$$B_n = A_n Q_n = U_n \Sigma_n V_n Q_n = U_n \Sigma_n \hat{V}_n,$$

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 $\{A_n\}_n$ and $\{B_n\}_n$ are distributed in the same way. □

Lemma 4.3.3. *Let $\{A_n\}_n$ and $\{Q_n\}_n$ be two sequences of matrices of strictly increasing dimension ($A_n, Q_n \in \mathbb{C}^{d_n \times d_n}$, $d_n < d_{n+1}$). If $\{A_n\}_n \sim_\sigma (0, G)$ and $\|Q_n\| \leq M$ for some nonnegative constant M independent of n, then $\{A_n Q_n\}_n \sim_\sigma (0, G)$ and $\{Q_n A_n\}_n \sim_\sigma (0, G)$.*

Proof. Since $\{A_n\}_n \sim_\sigma (0, G)$, then $\{0_n\}_n$ (sequence of null matrices) is an a.c.s for $\{A_n\}_n$, this means (by Definition 4.3.1) that we can write, for every m sufficiently large, $m \in \mathbb{N}$

$$A_n = 0_n + R_{n,m} + N_{n,m} \quad \text{for all } n > n_m, \tag{4.18}$$

with

$$Rank(R_{n,m}) \leq d_n c(m), \quad \|N_{n,m}\| \leq \omega(m),$$

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

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

Now consider the matrix $A_n Q_n$; from (4.18) we obtain

$$A_n Q_n = 0_n + R_{n,m} Q_n + N_{n,m} Q_n \quad \text{for all } n > n_m,$$

with

$$Rank(R_{n,m} Q_n) \leq \min\{Rank(R_{n,m}), Rank(Q_n)\} \leq Rank(R_{n,m}) \leq d_n c(m),$$

$$\|N_{n,m} Q_n\| \leq \|N_{n,m}\| \|Q_n\| \leq M \omega(m),$$

where

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

then $\{0_n\}_n$ is an a.c.s for the sequence $\{A_n Q_n\}_n$ and, by Proposition 4.3.1, $\{A_n Q_n\}_n \sim_\sigma (0, G)$. □

4.3.2 Singular value distribution for the g-Toeplitz sequences

As stated in formula (4.8), the matrix $T_{n,g}$ can be written as

$$T_{n,g} = [T_n \hat{Z}_{n,g} | \mathcal{T}_{n,g}] = [T_n \hat{Z}_{n,g} | 0] + [0 | \mathcal{T}_{n,g}]. \tag{4.19}$$

To find the distribution in the singular value sense of the sequence $\{T_{n,g}\}_n$, the idea is to study separately the distribution of the two sequences $\{[T_n \hat{Z}_{n,g}|0]\}_n$ and $\{[0|\mathcal{T}_{n,g}]\}_n$, to prove $\{[0|\mathcal{T}_{n,g}]\}_n \sim_\sigma (0, G)$, and then to apply Proposition 4.3.2.

Singular value distribution for the sequence $\{[T_n\hat{Z}_{n,g}|0]\}_n$

Since $T_n\hat{Z}_{n,g} \in \mathbb{C}^{n\times\mu_g}$ and $[T_n\hat{Z}_{n,g}|0] \in \mathbb{C}^{n\times n}$, the matrix $[T_n\hat{Z}_{n,g}|0]$ has $n-\mu_g$ singular values equal to zero and the remaining μ_g equal to those of $T_n\hat{Z}_{n,g}$; to study the distribution in the singular value sense of this sequence of non-square matrices, we use Lemma 4.3.1 : consider the g-Toeplitz matrix "truncated" $\hat{T}_n = T_n\hat{Z}_{n,g}$, 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

$$\hat{T}_{n,g}^\star\hat{T}_{n,g} = (T_n(f)\hat{Z}_{n,g})^\star T_n(f)\hat{Z}_{n,g} = \hat{Z}_{n,g}^\star T_n(f)^\star T_n(f)\hat{Z}_{n,g} = \hat{Z}_{n,g}^\star T_n(\bar{f})T_n(f)\hat{Z}_{n,g}. \tag{4.20}$$

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 [125]. From Proposition 4.3.3, if $f \in L^2(Q) \subset L^1(Q)$ (that is, $|f|^2 \in L^1(Q)$), then $\{T_n(\bar{f})T_n(f)\}_n \sim_\sigma (|f|^2, Q)$. Consequently, for every m sufficiently large, $m \in \mathbb{N}$, the use of Proposition 4.3.1 implies

$$T_n(\bar{f})T_n(f) = T_n(|f|^2) + R_{n,m} + N_{n,m} \quad \text{for all } n > n_m,$$

with

$$Rank(R_{n,m}) \le n\cdot c(m), \quad \|N_{n,m}\| \le \omega(m),$$

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

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

Therefore (4.20) becomes

$$\begin{aligned}\hat{T}_{n,g}^\star\hat{T}_{n,g} &= \hat{Z}_{n,g}^\star(T_n(|f|^2) + R_{n,m} + N_{n,m})\hat{Z}_{n,g}\\ &= \hat{Z}_{n,g}^\star T_n(|f|^2)\hat{Z}_{n,g} + \hat{Z}_{n,g}^\star R_{n,m}\hat{Z}_{n,g} + \hat{Z}_{n,g}^\star N_{n,m}\hat{Z}_{n,g}\\ &= \hat{Z}_{n,g}^\star T_n(|f|^2)\hat{Z}_{n,g} + \hat{R}_{n,m} + \hat{N}_{n,m},\end{aligned}$$

that is,

$$\hat{T}_{n,g}^\star\hat{T}_{n,g} = \hat{Z}_{n,g}^\star T_n(|f|^2)\hat{Z}_{n,g} + \hat{R}_{n,m,g} + \hat{N}_{n,m,g} \tag{4.21}$$

$$Rank(\hat{R}_{n,m,g}) \le \min\{Rank(\breve{Z}_{n,m}), Rank(R_{n,m})\} \le Rank(R_{n,m,g}) \le n.c(m), \tag{4.22}$$

$$\|\hat{N}_{n,m,g}\| \le 2\|\breve{Z}_{n,g}\|\|N_{n,m}\| \le 2\omega(m), \tag{4.23}$$

and

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

where in (4.22) and (4.23), $\breve{Z}_{n,m} = [\hat{Z}_{n,m}|0] \in \mathbb{C}^{n\times n}$. In order words, $\breve{Z}_{n,m}$ is the matrix $\hat{Z}_{n,m}$ supplemented by an appropriate number of zero columns in order to make it square. Furthermore, it is worth noticing that $\|\hat{Z}_{n,m}\| = \|\hat{Z}_{n,m}^\star\| = 1$, because $\hat{Z}_{n,m}$ is a submatrix of the identity: we have used the latter relation in (4.23).

Now, consider the matrix $\hat{Z}_{n,g}^\star T_n(|f|^2)\hat{Z}_{n,g} \in \mathbb{C}^{\mu_g\times\mu_g}$, $\mu_g = \lceil\frac{n}{g}\rceil$, $f \in L^2(Q) \subset L^1(Q)$ (so $|f|^2 \in L^1(Q)$). From (4.8), setting $T_n = T_n(|f|^2) = [\tilde{a}_{r-c}]_{r,c=0}^{n-1}$, with $\tilde{a}_j$ being the Fourier

coefficients of $|f|^2$, and setting $T_{n,g}$ the g-Toeplitz generated matrix by the same function $|f|^2$, it is immediate to observe

$$T_n \hat{Z}_{n,g} = \hat{T}_{n,g} \in \mathbb{C}^{n\times\mu_g}, \quad (\hat{T}_{n,g})_{r,c} = \tilde{a}_{r-gc}, \tag{4.24}$$

for $r = 0, 1, ..., n-1$ and $c = 0, 1, ..., \mu_g - 1$. If we compute $\hat{Z}^{\star}_{n,g}\hat{T}_{n,g} \in \mathbb{C}^{\mu_g\times\mu_g}$, where $\hat{Z}^{\star}_{n,g} = [\delta_{c-gr}]$, $r = 0, 1, ..., n-1$ and $c = 0, 1, ..., \mu_g - 1$, (δ_k defined as in (4.9)) and $\hat{Z}^{\star}_{n,g} \in \mathbb{C}^{\mu_g\times n}$ is the submatrix of $Z^{\star}_{n,g}$ obtained by considering only the μ_g first rows. For $r, c = 0, 1, ..., \mu_g - 1$, we obtain

$$\begin{aligned}
(\hat{Z}^{\star}_{n,g}T_n(|f|^2)\hat{Z}_{n,g})_{r,c} &= (\hat{Z}^{\star}_{n,g}\hat{T}_{n,g})_{r,c} \\
&= \sum_{l=0}^{n-1}(\hat{Z}^{\star}_{n,g})_{r,l}(\hat{T}_{n,g})_{l,c} \\
&\underset{(a)}{=} (\hat{T}_{n,g})_{gr,c} \\
&\underset{from\,(4.24)}{=} \tilde{a}_{gr-gc}
\end{aligned}$$

where (a) follows from the existence of a unique $l \in \{0, 1, ..., n-1\}$ such that $l - gr \equiv 0 (mod\, n)$, that is $l \equiv gr (mod\, n)$, and, since $0 \le gr \le n-1$, we find $l = gr$. Therefore

$$\hat{Z}^{\star}_{n,g}T_n(|f|^2)\hat{Z}_{n,g} = [\tilde{a}_{gr-gc}]^{n-1}_{r,c=0} = T_{\mu_g}(\widehat{|f|^{(2)}})$$

where $\widehat{|f|^{(2)}} \in L^1(Q)$ is given by

$$\widehat{|f|^{(2)}}(x) = \frac{1}{g}\sum_{j=0}^{g-1}|f|^2\left(\frac{x+2j\pi}{g}\right), \tag{4.25}$$

$$|f|^2(x) = \sum_{k=-\infty}^{\infty}\tilde{a}_k e^{\hat{i}kx}. \tag{4.26}$$

Proof. (of relation (4.25)). We denote by a_j the Fourier coefficients of $\widehat{|f|^{(2)}}(x)$. We want to show that for $r, c = 0, 1, ..., \mu_g - 1$, $a_{r-c} = \tilde{a}_{gr-gc}$, where $\tilde{a}_k$ are the Fourier coefficients of $|f|^2$. From (3.2), (4.25) and (4.26), we have

$$\begin{aligned}
a_{r-c} &= \frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{1}{g}\sum_{j=0}^{g-1}\sum_{k=-\infty}^{\infty}\tilde{a}_k e^{\hat{i}k(\frac{x+2\pi j}{g})}e^{-\hat{i}(r-c)x}dx \\
&= \frac{1}{2\pi g}\int_{-\pi}^{\pi}\sum_{k=-\infty}^{\infty}\tilde{a}_k\left(\sum_{j=0}^{g-1}e^{\hat{i}\frac{2\pi kj}{g}}\right)e^{\hat{i}\frac{kx}{g}}e^{-\hat{i}(r-c)x}dx.
\end{aligned}$$

Some remarks are in order:

- if k is a multiple of g, i.e., $k = gt$ for some value of t, then we have that

$$\sum_{j=0}^{g-1}e^{\hat{i}\frac{2\pi kj}{g}} = \sum_{j=0}^{g-1}e^{\hat{i}\frac{2\pi gtj}{g}} = \sum_{j=0}^{g-1}e^{\hat{i}2\pi tj} = \sum_{j=0}^{g-1}1 = g,$$

- if k is not a multiple of g, then $e^{\hat{i}\frac{2\pi k}{g}} \neq 1$ and therefore $\sum_{j=0}^{g-1} e^{\hat{i}\frac{2\pi kj}{g}} = \sum_{j=0}^{g-1} \left(e^{\hat{i}\frac{2\pi k}{g}}\right)^j$ is a finite geometric series whose sum is given by

$$\sum_{j=0}^{g-1} \left(e^{\hat{i}\frac{2\pi k}{g}}\right)^j = \frac{1 - e^{\hat{i}\frac{2\pi gk}{g}}}{1 - e^{\hat{i}\frac{2\pi k}{g}}} = 0.$$

Finally, taking into account the latter statements and recalling that $\frac{1}{2\pi}\int_{-\pi}^{\pi} e^{\hat{i}lx}dx = \begin{cases} 1 & \text{if } l = 0 \\ 0 & \text{otherwise} \end{cases}$, we find

$$\begin{aligned} a_{r-c} &= \frac{1}{2\pi g}\int_{-\pi}^{\pi} \sum_{t=-\infty}^{\infty} \tilde{a}_{gt} g e^{\hat{i}\frac{gtx}{g}} e^{-\hat{i}(r-c)x} dx \\ &= \sum_{t=-\infty}^{\infty} \tilde{a}_{gt} \frac{1}{2\pi}\int_{-\pi}^{\pi} e^{\hat{i}x(t-(r-c))} dx \\ &= \tilde{a}_{g(r-c)} \end{aligned}$$

□

In summary, from (4.21), one has

$$\hat{T}_{n,g}^* \hat{T}_{n,g} = T_{\mu_g}(\widehat{|f|^{(2)}}) + \hat{R}_{n,m,g} + \hat{N}_{n,m,g},$$

with $\{T_{\mu_g}(\widehat{|f|^{(2)}})\}_n \sim_\sigma (\widehat{|f|^{(2)}}, Q)$. We recall that, owing (4.25), the relation $|f|^2 \in L^1(Q)$ implies $\widehat{|f|^{(2)}} \in L^1(Q)$. Consequently Proposition 4.3.1 implies that $\{\hat{T}_{n,g}^\star \hat{T}_{n,g}\}_n \sim_\sigma (\widehat{|f|^{(2)}}, Q)$. Clearly $\widehat{|f|^{(2)}} \in L^1(Q)$ is equivalent to write $\sqrt{\widehat{|f|^{(2)}}} \in L^2(Q)$: therefore, from Lemma 4.3.1, we infer $\{\hat{T}_{n,g}\}_n \sim_\sigma (\sqrt{\widehat{|f|^{(2)}}}, Q)$.

Now, as mentioned at the beginning of this subsection, by Definition 4.2.1, we have

$$\begin{aligned} \lim_{n\to\infty} \frac{1}{n}\sum_{j=1}^{n} F\left(\sigma_j([\hat{T}_{n,g}|0])\right) &= \lim_{n\to\infty} \frac{1}{n}\sum_{j=1}^{\mu_g} F\left(\sigma_j([\hat{T}_{n,g}|0])\right) + \lim_{n\to\infty} \frac{1}{n}\sum_{j=\mu_g+1}^{n} F(0) \\ &= \lim_{n\to\infty} \frac{\mu_g}{n}\sum_{j=1}^{\mu_g} \frac{F\left(\sigma_j([\hat{T}_{n,g}|0])\right)}{\mu_g} + \lim_{n\to\infty} \frac{n-\mu_g}{n} F(0) \\ &= \frac{1}{2\pi g}\int_{-\pi}^{\pi} F\left(\sqrt{\widehat{|f|^{(2)}}}(x)\right) dx + \left(1 - \frac{1}{g}\right) F(0), \end{aligned}$$

which results to be equivalent to the following distribution formula

$$\{[T_n \hat{Z}_{n,g}|0]\}_n \sim_\sigma (\theta, Q \times [0,1]), \tag{4.27}$$

where

$$\theta(x,t) = \begin{cases} \sqrt{\widehat{|f|^{(2)}}}(x) & \text{for } t \in [0, \frac{1}{g}] \\ 0 & \text{for } t \in (\frac{1}{g}, 1]. \end{cases} \tag{4.28}$$

Singular value distribution for the sequence $\{[0|\mathcal{T}_{n,g}]\}_n$

In perfect analogy with the case of matrix $[T_n\hat{Z}_{n,g}|0]$, we can observe that $\mathcal{T}_{n,g} \in \mathbb{C}^{n\times(n-\mu_g)}$ and $[0|\mathcal{T}_{n,g}] \in \mathbb{C}^{n\times n}$. Therefore the matrix $[0|\mathcal{T}_{n,g}]$ has μ_g singular values equal to zero and the remaining $n-\mu_g$ equal to those of $\mathcal{T}_{n,g}$. However, in this case we have additional difficulties with respect to the matrix $\hat{T}_{n,g} = T_n\hat{Z}_{n,g}$, because it is not always true that $\mathcal{T}_{n,g}$ can be written as $T_n\mathcal{Z}_{n,g}$, where $\mathcal{Z}_{n,g}$ is the matrix obtained by considering the $n-\mu_g$ last columns of $Z_{n,g}$. Indeed, in $\mathcal{T}_{n,g}$ 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 coefficient of maximum index and $a_{-g(n-1)}$ as 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 last column of $\mathcal{T}_{n,g}$ and the other generated according to the rule $(T_n)_{j,k} = a_{j-k}$, it is not always true that we can write $\mathcal{T}_{n,g} = T_nP$ 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 first column the first column of $\mathcal{T}_{n,g}$, we find that $\beta_0 = (\mathcal{T}_{n,g})_{0,0} = (T_{n,g})_{0,\mu_g} = a_{-g\mu_g}$. As a consequence, T_n has $\beta_{-(n-1)} = a_{-(n-1)-g\mu_g}$ as coefficient of minimum index, while $\mathcal{T}_{n,g}$ has $a_{-g(n-1)}$ as coefficient of minimum index. Therefore

$$\begin{aligned} -(n-1)g-(-(n-1)-g\mu_g) &= (1-g)(n-1)+g\mu_g; \quad n \leq g\mu_g = g\left\lceil\frac{n}{g}\right\rceil \leq (n+g-1) \\ &\leq (1-g)(n-1)+n+g-1 \\ &= (1-g)(n-1)+(n-1)+g \\ &= (2-g)(n-1)+g < 0 \ \text{ for } g>2 \text{ and } n>4. \end{aligned}$$

Thus, if $g > 2$ and $n > 4$ we have $-(n-1)g < -(n-1)-g\mu_g$ and the coefficient of the minimum index $a_{-g(n-1)}$ of $\mathcal{T}_{n,g}$ is not contained in the matrix T_n that has $a_{-(n-1)-g\mu_g}$ as coefficient of minimum index.

Then we proceed in another way: in the first column of $\mathcal{T}_{n,g} \in \mathbb{C}^{n\times(n-\mu_g)}$ (and consequently throughout the matrix) there are only coefficients with index< 0, indeed: coefficient with the largest index of $\mathcal{T}_{n,g}$ is $(\mathcal{T}_{n,g})_{n-1,0} = (T_{n,g})_{n-1,\mu_g} = a_{n-1-g\mu_g}$ and $n-1-g\mu_g \leq n-1-n < 0$ and the coefficient with smallest index is $(\mathcal{T}_{n,g})_{0,n-\mu_g-1} = (T_{n,g})_{0,n-\mu_g-1+\mu_g} = (T_{n,g})_{0,n-1} = a_{-g(n-1)}$. Consider therefore a Toeplitz matrix $T_{d_{n,g}}$ of size $d_{n,g}$ with $d_{n,g} > \frac{g(n-1)}{2}+1$, defined in this way:

$$T_{d_{n,g}} = \begin{bmatrix} a_{-d_{n,g}+1} & a_{-d_{n,g}} & a_{-d_{n,g}-1} & \dots & a_{-2d_{n,g}+2} \\ a_{-d_{n,g}+2} & a_{-d_{n,g}+1} & \ddots & \ddots & a_{-2d_{n,g}+3} \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ a_{-1} & a_{-2} & \ddots & \ddots & a_{-d_{n,g}} \\ a_0 & a_{-1} & a_{-2} & \dots & a_{-d_{n,g}+1} \end{bmatrix} = [a_{r-c-d_{n,g}+1}]_{r,c=0}^{d_{n,g}-1}. \tag{4.29}$$

Since the coefficient with 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 $\mathcal{T}_{n,g}$ 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

$$\mathcal{T}_{n,g} = [0_1|I_n|0_2]T_{d_{n,g}}\mathcal{Z}_{d_{n,g},g} \tag{4.30}$$

where $\mathcal{Z}_{d_{n,g},g} \in \mathbb{C}^{d_{n,g}\times(n-\mu_g)}$ is the matrix of $Z_{d_{n,g},g}$ defined by considering only the $n-\mu_g$ first columns and $[0_1|I_n|0_2] \in \mathbb{C}^{n\times d_{n,g}}$ is a block matrix with $0_1 \in \mathbb{C}^{n\times(d_{n,g}-g\mu_g-1)}$ and $0_2 \in \mathbb{C}^{n\times(g\mu_g-n+1)}$.

Proof. (of relation (4.30)). First we observe that:

for $r = 0, 1, ..., n-1$ and $s = 0, 1, ..., n-\mu_g-1$, we have:

$$(\mathcal{T}_{n,g})_{r,s} = (T_{n,g})_{r,s+\mu_g} = a_{r-gs-g\mu_g}; \tag{4.31}$$

for $r = 0, 1, ..., n-1$ and $s = 0, 1, ..., d_{n,g}-1$, we have:

$$([0_1|I_n|0_2])_{r,s} = \begin{cases} 1 & \text{if } s = r+d_{n,g}-g\mu_g-1, \\ 0 & \text{otherwise.} \end{cases} \tag{4.32}$$

for $r, s = 0, 1, ..., d_{n,g}-1$ we have:

$$(T_{d_{n,g}})_{r,s} = a_{r-s-d_{n,g}+1};$$

for $r = 0, 1, ..., d_{n,g}-1$ and $s = 0, 1, ..., n-\mu_g-1$, we have:

$$(\mathcal{Z}_{d_{n,g},g})_{r,s} = \delta_{r-gs}.$$

Since $T_{d_{n,g}}\mathcal{Z}_{d_{n,g},g} \in \mathbb{C}^{d_{n,g}\times(n-\mu_g)}$, for $r = 0, 1, ..., d_{n,g}-1$ and $s = 0, 1, ..., n-\mu_g-1$, it holds

$$\begin{aligned} (T_{d_{n,g}}\mathcal{Z}_{d_{n,g},g})_{r,s} &= \sum_{l=0}^{d_{n,g}-1} (T_{d_{n,g}})_{r,l}(\mathcal{Z}_{d_{n,g},g})_{l,s} \\ &= \sum_{l=0}^{d_{n,g}-1} \delta_{l-gs}a_{r-l-d_{n,g}+1} \\ &\underset{(a)}{=} a_{r-gs-d_{n,g}+1}, \end{aligned}$$

so,

$$(T_{d_{n,g}}\mathcal{Z}_{d_{n,g},g})_{r,s} = a_{r-gs-d_{n,g}+1}, \tag{4.33}$$

where (a) follows from the existence of a unique $l \in \{0, 1, ..., d_{n,g}-1\}$ such that $l - gs \equiv 0 (mod\, d_{n,g})$, that is, $l \equiv gs (mod\, d_{n,g})$, and, since $0 \le gs \le d_{n,g}-1$, we have $l = gs$. Since $[0_1|I_n|0_2]T_{d_{n,g}}\mathcal{Z}_{d_{n,g},g} \in \mathbb{C}^{n\times(n-\mu_g)}$, for $r = 0, 1, ..., n-1$ and $s = 0, 1, ..., n-\mu_g-1$, we find

$$\begin{aligned} ([0_1|I_n|0_2]T_{d_{n,g}}\mathcal{Z}_{d_{n,g},g})_{r,s} &= \sum_{l=0}^{d_{n,g}-1} ([0_1|I_n|0_2])_{r,l}(T_{d_{n,g}}\mathcal{Z}_{d_{n,g},g})_{l,s} \\ &\underset{(d)}{=} a_{r+d_{n,g}-g\mu_g-1-gs-d_{n,g}+1} \\ &= a_{r-g\mu_g-gs} \\ &\underset{from (4.31)}{=} (\mathcal{T}_{n,g})_{r,s}, \end{aligned}$$

where (d) follows from (4.33), $(T_{d_{n,g}}\mathcal{Z}_{d_{n,g},g})_{l,s} = a_{l-gs-d_{n,g}+1}$, and from the following fact: using (4.32), 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 (4.29) can be written as

$$T_{d_{n,g}} = J \cdot H_{d_{n,g}} \tag{4.34}$$

where J is the "**flip**" permutation matrix of dimension $d_{n,g} \times d_{n,g}$:

$$J = \begin{bmatrix} & & & 1 & 1 \\ & & \dots & & \\ 1 & & & & \end{bmatrix},$$

that is, $(J)_{s,t} = 1$ if and only if $s + t = d_{n,g} + 1$, and $H_{d_{n,g}}$ is the **Hankel matrix** of dimension $d_{n,g} \times d_{n,g}$.

$$H_{d_{n,g}} = \begin{bmatrix} a_0 & a_{-1} & a_{-2} & \dots & a_{-d_{n,g}+1} \\ a_{-1} & a_{-2} & \dots & \dots & a_{-d_{n,g}} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ a_{-d_{n,g}+2} & a_{-d_{n,g}+1} & \dots & \dots & a_{-2d_{n,g}+3} \\ a_{-d_{n,g}+1} & a_{-d_{n,g}} & a_{-d_{n,g}-1} & \dots & a_{-2d_{n,g}+2} \end{bmatrix} = [a_{-r-c}]_{r,c=0}^{d_{n,g}-1}.$$

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 (4.8), 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 Theorem 6, page 161 of [59], the sequence of matrices $\{H_{d_{n,g}}\}_n$ is distributed in the singular value sense as the zero function: $\{H_{d_{n,g}}\}_n \sim_\sigma (0, Q)$. From Lemma 4.3.2, by (4.34), since J is a unitary matrix, we have $\{T_{d_{n,g}}\}_n \sim_\sigma (0, Q)$. as well.

Consider the decomposition in (4.30) :

$$\mathcal{T}_{n,g} = [0_1|I_n|0_2]T_{d_{n,g}}\mathcal{Z}_{d_{n,g},g} = Q_{d_{n,g}}T_{d_{n,g}}\mathcal{Z}_{d_{n,g},g}.$$

If we complete the matrix $Q_{d_{n,g}} \in \mathbb{C}^{n \times d_{n,g}}$ and $\mathcal{Z}_{d_{n,g},g} \in \mathbb{C}^{d_{n,g} \times (n-\mu_g)}$ by adding an appropriate number of zero rows and columns, respectively, in order to make it square

$$\begin{aligned} \mathbf{Q}_{d_{n,g}} &= \left[\begin{array}{c} Q_{d_{n,g}} \\ \hline 0 \end{array}\right] \in \mathbb{C}^{d_{n,g} \times d_{n,g}}, \\ \mathbf{Z}_{d_{n,g}} &= \left[\begin{array}{c|c} \mathcal{Z}_{d_{n,g}} & 0 \end{array}\right] \in \mathbb{C}^{d_{n,g} \times d_{n,g}}, \end{aligned}$$

then it is immediately to note that

$$\mathbf{Q}_{d_{n,g}}T_{d_{n,g}}\mathbf{Z}_{d_{n,g},g} = \left[\begin{array}{c|c} \mathcal{T}_{n,g} & 0 \\ \hline 0 & 0 \end{array}\right] := \mathbf{T}_{d_{n,g}} \in \mathbb{C}^{d_{n,g} \times d_{n,g}}.$$

From Lemma 4.3.3, since $\|\mathbf{Q}_{d_{n,g}}\| = \|\mathbf{Z}_{d_{n,g}}\| = 1$ (indeed, there are both "incomplete" permutation matrices), and since $\{T_{d_{n,g}}\}_n \sim_\sigma (0, Q)$, we infer that $\{\mathbf{T}_{d_{n,g}}\}_n \sim_\sigma (0, Q)$.

Recall that $\mathbf{T}_{d_{n,g}} \in \mathbb{C}^{d_{n,g} \times d_{n,g}}$ 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 $\{\mathbf{T}_{d_{n,g}}\}_n \sim_\sigma (0, Q)$, it holds that the sequence $\{\mathbf{T}_{d_{n,g}}\}_n$ is weakly clustered at zero in the singular value sense, i.e, $\forall \epsilon > 0$,

$$\sharp\{j : \sigma_j(\mathbf{T}_{d_{n,g}}) > \epsilon\} = o(d_{n,g}) = o(gn) = o(n). \tag{4.35}$$

The matrix $\mathbf{T}_{d_{n,g}}$ is a block matrix that can be written as

$$\mathbf{T}_{d_{n,g}} = \left[\begin{array}{c|c} \mathcal{T}_{n,g} & 0 \\ \hline 0 & 0 \end{array}\right] = \left[\begin{array}{c|c} [\mathcal{T}_{n,g}|0] & 0 \\ \hline 0 & 0 \end{array}\right],$$

where $\mathcal{T}_{n,g} \in \mathbb{C}^{n\times(n-\mu_g)}$ and $[\mathcal{T}_{n,g}|0] \in \mathbb{C}^{n\times n}$. By the singular value decomposition we obtain

$$\mathbf{T}_{d_{n,g}} = \left[\begin{array}{c|c} [\mathcal{T}_{n,g}|0] & 0 \\ \hline 0 & 0 \end{array}\right] = \left[\begin{array}{c|c} U_1\Sigma_1V_1^\star & 0 \\ \hline 0 & U_20V_2^\star \end{array}\right] = \left[\begin{array}{c|c} U_1 & 0 \\ \hline 0 & U_2 \end{array}\right] \left[\begin{array}{c|c} \Sigma_1 & 0 \\ \hline 0 & 0 \end{array}\right] \left[\begin{array}{c|c} V_1 & 0 \\ \hline 0 & V_2 \end{array}\right]^\star,$$

that is, the singular values of $\mathbf{T}_{d_{n,g}}$ that are different from zero are the singular values of $[\mathcal{T}_{n,g}|0] \in \mathbb{C}^{n\times n}$. Thus, (4.35) can be written as follows: $\forall \epsilon > 0$,

$$\sharp\{j : \sigma_j([\mathcal{T}_{n,g}|0]) > \epsilon\} = o(d_{n,g}) = o(gn) = o(n).$$

The latter relation means that the sequence $\{[\mathcal{T}_{n,g}|0]\}_n$ is weakly clustered at zero in the singular value sense, and hence $\{[\mathcal{T}_{n,g}|0]\}_n \sim_\sigma (0, Q)$. If we consider the matrix

$$\hat{G} = \left[\begin{array}{c|c} 0 & I_{n-\mu_g} \\ \hline 0 & 0 \end{array}\right] \in \mathbb{C}^{n\times n},$$

where $I_{n-\mu_g}$ is the identity matrix of dimension $(n-\mu_g)\times(n-\mu_g)$, then $[\mathcal{T}_{n,g}|0]\hat{G} = [0|\mathcal{T}_{n,g}]$, and since $\|\hat{G}\| = 1$ and $\{[\mathcal{T}_{n,g}|0]\}_n \sim_\sigma (0, Q)$, from Lemma 4.3.3 we find

$$\{[0|\mathcal{T}_{n,g}]\}_n \quad \sim_\sigma \quad (0, Q). \tag{4.36}$$

In conclusion, from (4.19), (4.27) and (4.36), using Proposition 4.3.2 with $G = Q \times [0, 1]$, we obtain

$$\{T_{n,g}\}_n \sim_\sigma (\theta, G),$$

where θ is defined in (4.28). Notice that for $g = 1$ the symbol $\theta(x, t)$ coincides with $|f|(x)$ on the extended domain $Q \times [0, 1]$. Hence, the Szegö-Tilli-Tyrtyshnikov-Zamarashkin result is found as a particular case. Indeed: $\theta(x, t) = |f|(x)$ does not depend on t and therefore this additional variable can be suppressed i.e. $\{T_{n,g}\}_n \sim_\sigma (f, Q)$ 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 because any measure-preserving exchange function is a distribution function if one representative of the class is.

4.4 Some remarks on multigrid methods

In the design of multigrid methods for large positive definite linear systems one of the key points is to maintain the structure (if any) of the origin matrix in the lower levels. This means that at every recursion level the new projected linear system should retain the main properties of the origin matrix (e.g bandedness, the same level of conditioning, the same algebra/Toeplitz/graph structure etc...). Here for the sake of simplicity the example that has to be considered is the one-level circulant case. Following [1, 141], if $A_n = C_n$ is a positive circulant matrix of size n with n power of 2, then the projected matrix A_k with $k = n/2$ is defined as

$$A_k = \widetilde{Z}_{n,2}^T P_n^\star A_n P_n \widetilde{Z}_{n,2}, \tag{4.37}$$

where P_n is an additional circulant matrix. It is worth noticing that the structure is kept since for every circulant P_n the matrix A_k is a circulant matrix of size $k = n/2$. The features of the specific P_n have to be designed in such a way that the convergence speed of the related multigrid is as high as possible (see [61, 1] for a general strategy). We observe that the eigenvalues of A_k are given by

$$\frac{1}{2}\sum_{l=0}^{1} g\left(\frac{x_j + 2\pi l}{2}\right), \quad x_j = \frac{2\pi j}{k}, \quad j = 0, 1, ..., k-1, \quad k = n/2, \tag{4.38}$$

where g is the polynomial associated with the circulant matrix $P_n^\star A_n P_n$ in the sense of subsection 3.3.3. Therefore the singular values of $(P_n^\star A_n P_n)^{1/2}\widetilde{Z}_{n,2}$ are given by

$$\frac{1}{\sqrt{2}}\sqrt{\sum_{l=0}^{1} g\left(\frac{x_j+2\pi l}{2}\right)},\quad x_j=\frac{2\pi j}{k},\quad j=0,1,...,k-1,\ k=n/2. \tag{4.39}$$

Notice that the latter formula is a special instance of

$$\sigma_j(C_{n,g})=\sqrt{\sum_{l=0}^{(n,g)-1}|p|^2\left(\frac{x_j+2\pi l}{(n,g)}\right)},\quad x_j=\frac{2\pi j}{n_g},\quad j=0,1,...,n_g-1 \tag{4.40}$$

for $|p|^2=g$ (g is necessarily nonnegative since it can be written as $|q|^2 f$ where q is the polynomial associated with P_n and f the nonnegative polynomial associated with A_n), for $g=2$ and n even number so that $(n,2)=2$. Therefore, according to (4.40), the numbers in (4.39) identify the nontrivial singular values of the 2-circulant matrix $(P_n^\star A_n P_n)^{1/2}\widetilde{Z}_{n,2}$ up to a scaling factor. In order words, g-circulant matrices arise naturally in the design of fast multigrid solvers for circulant linear systems and, along the same line g-Toeplitz matrices arise naturally in the design of fast multigrid solvers for Toeplitz linear systems; see [61, 1, 126].

Conversely, we now can see clearly that formula (4.40) furnishes a wide generalization of the spectral analysis of the projected matrices, by allowing a higher degree of freedom: we can choose n divisible by g with $g\neq 2$, we can choose n not divisible by g. Such a degree of freedom is not just academic, but could be exploited for devising optimally convergence multigrid solvers also in critical cases emphasized e.g. in [1, 126]. In particular, if x_0 is an isolated zero of f (the nonnegative polynomial related to $A_n=C_n$) and also $\pi+x_0$ is a zero for the same function, then due to special symmetries, the associated multigrid (or even two-grid) method cannot be optimal. In other words, for reaching a preassigned accuracy, we cannot expect a number of iterations independent of the order n. However these pathological symmetries are due to the choice of $g=2$, so that a choice of a projector as $P_n\widetilde{Z}_{n,g}$ for a different $g\neq 2$ and a different n could completely overcome the latter drawback.

4.5 Generalizations

First of all, we observe that the requirements that the symbol f is square integrable can be removed. In [125] it is proven that the singular value distribution of $\{T_n(f)T_n(g)\}_n$ 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 of relation $\{T_{n,g}\}_n\sim_\sigma(\theta,Q\times[0,1])$ to the case where $\theta(x,t)$ is defined as in (4.28) with the original symbol $f\in L^1$.

Now we consider the general multilevel case. When g is a positive vector, we have

$$\{T_{n,g}\}_n\sim_\sigma(\theta,Q^d\times[0,1]^d) \tag{4.41}$$

where

$$\theta(x,t)\ =\ \begin{cases}\sqrt{\widehat{|f|^{(2)}}(x)} & \text{if } t\in[\underline{0},1/g],\\ 0 & \text{for } t\in(1/g,e],\end{cases} \tag{4.42}$$

with

$$\widehat{|f|^{(2)}}(x) = \frac{1}{\hat{g}} \sum_{j=\underline{0}}^{g-e} |f|^2 \left(\frac{x + 2\pi j}{g} \right) \tag{4.43}$$

and where all the arguments are modulus 2π and all the operations are intended componentwise, that is $t \in [\underline{0}, \frac{1}{g}]$ means that $t_k \in [0, \frac{1}{g_k}]$, $k = 1, 2, ..., d$ and $t \in (\frac{1}{g}, e]$ means that $t_k \in (\frac{1}{g_k}, 1]$, $k = 1, 2, ..., d$. The writing $\frac{x+2\pi j}{g}$ defines the d-dimensional vector whose k-th component is $\frac{x_j + 2\pi j_k}{g_k}$, $k = 1, 2, ..., d$ and $\hat{g} = g_1 g_2 ... g_d$.

4.6 Examples of g-Toeplitz matrices when some of the entries of g vanish

We start this subsection with a brief digression on multilevel matrices. A d-level matrix A of dimension $\hat{n} \times \hat{n}$ with $n = (n_1, n_2, ..., n_d)$ and $\hat{n} = n_1 n_2 ... n_d$ can be viewed as a matrix of dimension $n_1 \times n_1$ in which each element is a block of dimension $n_2 n_3 ... n_d \times n_2 n_3 ... n_d$; in turn, each block of dimension $n_2 n_3 ... n_d \times n_2 n_3 ... n_d$ can be viewed as a matrix of dimension $n_2 \times n_2$ in which each element is a block of dimension $n_3 n_4 ... n_d \times n_3 n_4 ... n_d$, and so on. So we can say that n_1 is the most "outer" dimension of the matrix A and n_d is the "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 $P^T A P$ becomes a matrix again of dimension $\hat{n} \times \hat{n}$ but with $n = (n_{p(1)}, n_{p(2)}, ..., n_{p(d)})$ and $\hat{n} = n_{p(1)} n_{p(2)} ... n_{p(d)} = n_1 n_2 ... n_d$ (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-Toeplitz d-level matrices, especially when some of the entries of the vector g are zero; indeed: as we observed in subsection 4.2.1, if $g = \underline{0}$, the d-level g-Toeplitz matrix A 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, $g = (0, ..., 0, g_k, ..., g_d)$, the matrix $P^T A P$ (where P is the permutation matrix associated with p) becomes a block matrix with constant blocks on each row and with blocks of dimension $n_k n_{k+1} ... n_d \times n_k n_{k+1} ... n_d$; with this "new" structure, formula (4.6) is even more intuitively understandable, as we shall see later in the examples.

Lemma 4.6.1. *Let A be a 2-level Toeplitz matrix of dimension $\hat{n} \times \hat{n}$ with $n = (n_1, n_2)$ and $\hat{n} = n_1 n_2$,*

$$A = \left[[a_{(j_1 - k_1, j_2 - k_2)}]_{j_2, k_2 = 0}^{n_2 - 1} \right]_{j_1, k_1 = 0}^{n_1 - 1}.$$

There exists a permutation matrix P such that

$$P^T A P = \left[[a_{(j_1 - k_1, j_2 - k_2)}]_{j_1, k_1 = 0}^{n_1 - 1} \right]_{j_2, k_2 = 0}^{n_2 - 1}.$$

Example 4.6.1. *Let $(n_1, n_2) = (2, 3)$ and consider the 2-level Toeplitz matrix A of dimension 6×6*

$$A = \left[\begin{array}{ccc|ccc} a_{(0,0)} & a_{(0,-1)} & a_{(0,-2)} & a_{(-1,0)} & a_{(-1,-1)} & a_{(-1,-2)} \\ a_{(0,1)} & a_{(0,0)} & a_{(0,-1)} & a_{(-1,1)} & a_{(-1,0)} & a_{(-1,-1)} \\ a_{(0,2)} & a_{(0,1)} & a_{(0,0)} & a_{(-1,2)} & a_{(-1,1)} & a_{(-1,0)} \\ \hline a_{(1,0)} & a_{(1,-1)} & a_{(1,-2)} & a_{(0,0)} & a_{(0,-1)} & a_{(0,-2)} \\ a_{(1,1)} & a_{(1,0)} & a_{(1,-1)} & a_{(0,1)} & a_{(0,0)} & a_{(0,-1)} \\ a_{(1,2)} & a_{(1,1)} & a_{(1,0)} & a_{(0,2)} & a_{(0,1)} & a_{(0,0)} \end{array}\right].$$

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

$$P^T AP = \left[\begin{array}{cc|cc|cc} a_{(0,0)} & a_{(-1,0)} & a_{(0,-1)} & a_{(-1,-1)} & a_{(0,-2)} & a_{(-1,-2)} \\ a_{(1,0)} & a_{(0,0)} & a_{(1,-1)} & a_{(0,-1)} & a_{(1,-2)} & a_{(0,-2)} \\ \hline a_{(0,1)} & a_{(-1,1)} & a_{(0,0)} & a_{(-1,0)} & a_{(0,-1)} & a_{(-1,-1)} \\ a_{(1,1)} & a_{(0,1)} & a_{(1,0)} & a_{(0,0)} & a_{(1,-1)} & a_{(0,-1)} \\ \hline a_{(0,2)} & a_{(-1,2)} & a_{(0,1)} & a_{(-1,1)} & a_{(0,0)} & a_{(-1,0)} \\ a_{(1,2)} & a_{(0,2)} & a_{(1,1)} & a_{(0,1)} & a_{(1,0)} & a_{(0,0)} \end{array}\right],$$

and now $P^T AP$ *can be naturally viewed as a matrix of dimension* 3×3 *in which each element is a block of dimension* 2×2.

Corollary 4.6.1. *Let* A *be a* d*-level Toeplitz matrix of dimension* $\hat{n} \times \hat{n}$ *with* $n = (n_1, n_2, ..., n_d)$ *and* $\hat{n} = n_1 n_2 ... n_d$,

$$A = \left[\left[... [a_{(j_1 - k_1, j_2 - k_2, ..., j_d - k_d)}]_{j_d, k_d = 0}^{n_d - 1} \cdots \right]_{j_2, k_2 = 0}^{n_2 - 1} \right]_{j_1, k_1 = 0}^{n_1 - 1}.$$

For every permutation p *of* d *elements, there exists a permutation matrix* P *such that*

$$P^T AP = \left[\left[... [a_{(j_1 - k_1, j_2 - k_2, ..., j_d - k_d)}]_{j_{p(d)}, k_{p(d)} = 0}^{n_{p(d)} - 1} \cdots \right]_{j_{p(2)}, k_{p(2)} = 0}^{n_{p(2)} - 1} \right]_{j_{p(1)}, k_{p(1)} = 0}^{n_{p(1)} - 1}.$$

Remark 4.6.1. *Lemma* 4.6.1 *and Corollary* 4.6.1 *also apply to* d*-level* g*-Toeplitz matrices.*

Now, let $g = (g_1, g_2, ...g_d)$ be a d-dimensional vector of nonnegative 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 $g_{p(1)} = g_{p(2)} = ... = g_{p(t)} = 0$, (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 formula (4.6) remains the same for the matrix $P^T AP$ (where P is the permutation matrix associated with p) but with $n[0] = (n_{p(1)}, n_{p(2)}, ..., n_{p(t)})$, and where T_j is a d^+-level g^+-Toeplitz matrix, with $g^+ = (g_{p(t+1)}, g_{p(t+2)}, ..., g_{p(d)})$, of partial size $n[> 0] = (n_{p(t+1)}, n_{p(t+2)}, ..., n_{p(d)})$, and whose expression is

$$T_j = \left[\left[... [a_{(r - g \circ s)}]_{r_{p(d)}, s_{p(d)} = 0}^{n_{p(d)} - 1} \cdots \right]_{r_{p(t+2)}, s_{p(t+2)} = 0}^{n_{p(t+2)} - 1} \right]_{r_{p(t+1)}, s_{p(t+1)} = 0}^{n_{p(t+1)} - 1},$$

with $(r_{p(1)}, r_{p(2)}, ..., r_{p(t)}) = j$. Obviously $Sval(A) = Sval(P^T AP)$.

We recall that if B is a matrix of size $n \times n$ positive semidefinite, that is $B^\star = B$ and $x^\star Bx \geq 0 \ \forall x \neq 0$, then Eig(B)=Sval(B). Moreover, if $B = U \Sigma U^\star$ is a SVD for B (which coincides with the Schur decomposition of B) with $\Sigma = \underset{j=1,...,n}{diag}\ (\sigma_j)$, then

$$B^{1/2} = U \Sigma^{1/2} U^\star, \tag{4.44}$$

where $\Sigma^{1/2} = \underset{j=1,...,n}{diag}\ (\sqrt{\sigma_j})$.

We proceed with a detailed example: a 3-level g-Toeplitz matrix 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.

Example 4.6.2. *Consider a 3-level g-Toeplitz matrix A where* $g = (g_1, g_2, g_3) = (0, 1, 2)$

$$\begin{aligned} A &= \left[\left[\left[a_{(r_1-0\cdot s_1, r_2-1\cdot s_2, r_3-2s_3)}\right]_{r_3,s_3=0}^{n_3-1}\right]_{r_2,s_2=0}^{n_2-1}\right]_{r_1,s_1=0}^{n_1-1} \\ &= \left[\left[\left[a_{(r_1, r_2-s_2, r_3-2s_3)}\right]_{r_3,s_3=0}^{n_3-1}\right]_{r_2,s_2=0}^{n_2-1}\right]_{r_1=0}^{n_1-1}. \end{aligned}$$

The procedure is the same as the example 3.5.1 *in chapter* 3 *for an 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*

$$T_{r_1} = \left[\left[a_{(r_1, r_2-s_2, r_3-2s_3)}\right]_{r_3,s_3=0}^{n_3-1}\right]_{r_2,s_2=0}^{n_2-1},$$

then 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*

$$A = \begin{bmatrix} T_0 & T_0 & \dots & T_0 \\ T_1 & T_1 & \dots & T_1 \\ \vdots & \vdots & \vdots & \vdots \\ T_{n_1-1} & T_{n_1-1} & \dots & T_{n_1-1} \end{bmatrix}.$$

The latter is a block matrix with constant blocks on each row. From formula (4.1), *the singular values of A are the square root of the eigenvalues of* $A^\star A$:

$$\begin{aligned} A_n^\star A_n &= \begin{bmatrix} T_0^\star & T_1^\star & \dots & T_{n_1-1}^\star \\ T_0^\star & T_1^\star & \dots & T_{n_1-1}^\star \\ \vdots & & \vdots & \vdots \\ T_0^\star & T_1^\star & \dots & T_{n_1-1}^\star \end{bmatrix} \begin{bmatrix} T_0 & T_0 & \dots & T_0 \\ T_1 & T_1 & \dots & T_1 \\ \vdots & & \vdots & \\ T_{n_1-1} & T_{n_1-1} & \dots & T_{n_1-1} \end{bmatrix} \\ &= \begin{bmatrix} \sum_{j=0}^{n_1-1} T_j^\star T_j & \sum_{j=0}^{n_1-1} T_j^\star T_j & \dots & \sum_{j=0}^{n_1-1} T_j^\star T_j \\ \sum_{j=0}^{n_1-1} T_j^\star T_j & \sum_{j=0}^{n_1-1} T_j^\star T_j & \dots & \sum_{j=0}^{n_1-1} T_j^\star T_j \\ \vdots & \vdots & \vdots & \vdots \\ \sum_{j=0}^{n_1-1} T_j^\star T_j & \sum_{j=0}^{n_1-1} T_j^\star T_j & \dots & \sum_{j=0}^{n_1-1} T_j^\star T_j \end{bmatrix} \\ &= \begin{bmatrix} 1 & \dots & 1 \\ \vdots & \vdots & \vdots \\ 1 & \dots & 1 \end{bmatrix} \otimes \sum_{j=0}^{n_1-1} T_j^\star T_j \\ &= J_{n_1} \otimes \sum_{j=0}^{n_1-1} T_j^\star T_j. \end{aligned}$$

Therefore

$$Eig(A^\star A) = Eig\left(J_{n_1} \otimes \sum_{j=0}^{n_1-1} T_j^\star T_j\right), \tag{4.45}$$

where

$$Eig(J_{n_1}) = \{0, n_1\}, \tag{4.46}$$

because J_{n_1} is a matrix of rank 1, so it has all eigenvalues equal to zero except one eigenvalue equal to $trace(J_{n_1}) = n_1$. If we put

$$\lambda_k = \lambda_k \left(\sum_{j=0}^{n_1-1} T_j^\star T_j \right), \quad k = 0, 1, ..., n_3 n_2 - 1,$$

by exploiting basic properties of the tensor product and taking into consideration (4.45) *and* (4.46) *we find*

$$\lambda_k(A^\star A) = n_1 \lambda_k, \quad k = 0, 1, ..., n_3 n_2 - 1, \tag{4.47}$$

$$\lambda_k(A^\star A) = 0, \quad k = n_3 n_2, n_3 n_2 + 1, ..., n_3 n_2 n_1 - 1. \tag{4.48}$$

From (4.47), (4.48) *and* (4.1), *one obtains that the singular values of A are given by*

$$\sigma_k(A) = \sqrt{n_1 \lambda_k}, \quad k = 0, 1, ..., n_3 n_2 - 1, \tag{4.49}$$

$$\sigma_k(A) = 0, \quad k = n_3 n_2, n_3 n_2 + 1, ..., n_3 n_2 n_1 - 1. \tag{4.50}$$

and, since $\sum_{j=0}^{n_1-1} T_j^\star T_j$ is a positive semidefinite matrix, from (4.44) *we can write*

$$\sigma_k(A) = \sqrt{n_1} \widetilde{\sigma}_k, \quad k = 0, 1, ..., n_3 n_2 - 1,$$

$$\sigma_k(A) = 0, \quad k = n_3 n_2, n_3 n_2 + 1, ..., n_3 n_2 n_1 - 1.$$

where $\widetilde{\sigma}_k$ denotes the generic singular value of $\left(\sum_{j=0}^{n_1-1} T_j^\star T_j \right)^{1/2}$.

Regarding the distribution in the sense of singular values, let $F \in C_0(\mathbb{R}_0^+)$, continuous over $\mathbb{R}_0^+$ with bounded support, then there exists $a \in \mathbb{R}^+$ such that

$$|F(x)| \leq a \ \ \forall x \in \mathbb{R}_0^+. \tag{4.51}$$

From formula (4.2) *we have*

$$\begin{aligned} \Sigma_\sigma(F, A_n) &= \frac{1}{n_1 n_2 n_3} \sum_{j=0}^{n_1 n_2 n_3 - 1} F(\sqrt{n_1} \widetilde{\sigma}_j) \\ &= \frac{n_2 n_3 (n_1 - 1) F(0)}{n_1 n_2 n_3} + \frac{1}{n_1 n_2 n_3} \sum_{j=0}^{n_2 n_3 - 1} F(\sqrt{n_1} \widetilde{\sigma}_j) \\ &= \left(1 - \frac{1}{n_1}\right) F(0) + \frac{1}{n_1 n_2 n_3} \sum_{j=0}^{n_2 n_3 - 1} F(\sqrt{n_1} \widetilde{\sigma}_j) \end{aligned}$$

According to (4.51), *we find*

$$-a n_2 n_3 \leq \sum_{j=0}^{n_2 n_3 - 1} F(\sqrt{n_1} \widetilde{\sigma}_j) \leq a n_2 n_3.$$

Therefore

$$\frac{-a}{n_1} \leq \frac{1}{n_1 n_2 n_3} \sum_{j=0}^{n_2 n_3 - 1} F(\sqrt{n_1}\widetilde{\sigma}_j) \leq \frac{a}{n_1},$$

so that

$$\left(1 - \frac{1}{n_1}\right) F(0) + \frac{-a}{n_1} \leq \Sigma_\sigma(F, A_n) \leq \left(1 - \frac{1}{n_1}\right) F(0) + \frac{a}{n_1}.$$

Now recalling that the writing $\lim_{n\to\infty} \min_{1\leq j\leq 3} n_j = \infty$, *we obtain*

$$F(0) \leq \lim_{\tilde{n}\to\infty} \Sigma_\sigma(F, A_n) \leq F(0)$$

which implies

$$\lim_{\tilde{n}\to\infty} \Sigma_\sigma(F, A_n) = F(0)$$

Whence

$$\{A_n\} \sim_\sigma (0, G)$$

for any domain G *satisfying the requirements of Definition* 4.2.1.

Conclusion

In this chapter We have studied the distribution in the singular value sense of g-Toeplitz sequences associated with a given integrable symbol. The generalization to the multilevel block setting has been sketched together with some intriguing relationship with the design of multigrid procedures for structured linear systems. In chapter 5, we will recall some preliminary notions of construction of Krylov Space Methods and will study in Chapter 6 powerful methods of solving of large structured systems of linear equations. More precisely, the Krylov Space Methods and a general idea of Multigrid Methods.

Chapter Five

Preliminary Notions of Construction of the Krylov Space Methods

Throughout this chapter, we present general ideas on least-squares problems, by recalling some results of analysis based on the Newton method for the convergence of minimization problems. The Lanczos method (see [95]) for the reduction of a Hermitian matrix to the Tridiagonal form, which are useful for the study of Krylov space methods ends the chapter.

5.1 General idea of least-squares problems

In the following $\|x\|_2$ denotes the Euclidian norm $\|x\|_2 := \sqrt{x^T x}$ where $x \in \mathbb{R}^n$.

5.1.1 Least square problems. The Normal equations

Let a real $m \times n$ matrix A and a vector $y \in \mathbb{R}^m$ be given, and let

$$\|y - Ax\|_2^2 = (y - Ax)^T (y - Ax) \tag{5.1}$$

be minimized as a function of x. We want to show that $x \in \mathbb{R}^n$ is a solution of the normal equations

$$A^T Ax = A^T y \tag{5.2}$$

if and only if $x \in \mathbb{R}^n$ is also a minimum point for (5.1). We have the following result:

Theorem 5.1.1. [27] *The linear least-squares problem*

$$\min_{x \in \mathbb{R}^n} \|y - Ax\|_2$$

has at least one minimum point x_0. If x_1 is another minimum point, then $Ax_0 = Ax_1$. The residual $r := y - Ax_0$ is uniquely determined and satisfies the normal equation $A^T r = 0$. Every minimum point x_0 is also a solution of the normal equations (5.2) *and conversely.*

Proof. Let $L \subset \mathbb{R}^m$ be the linear subspace

$$L = \{Ax | x \in \mathbb{R}^n\}$$

which is spanned by the columns of A, and let L^T be the orthogonal complement

$$L^T := \{r \in \mathbb{R}^m | r^T z = 0 \text{ for all } z \in L\} = \{r | r^T A = 0\}.$$

Because $\mathbb{R}^m = L \oplus L^T$, the vector $y \in \mathbb{R}^m$ can be written uniquely in the form

(5.3) $$y = s + r, \;\; s \in L, \;\; r \in L^T,$$

and there is at least one $x_0 \in \mathbb{R}^n$ with $Ax_0 = s$. Because $A^T r = 0$, x_0 satisfies

$$A^T y = A^T s = A^T A x_0,$$

that is, x_0 is a solution of the normal equations. Further, each solution x_0 of the normal equations is a minimum point for the problem

$$\min_{x \in \mathbb{R}^n} \|y - Ax\|_2.$$

To see this, let $x \in \mathbb{R}^n$ be an arbitrary vector, and set

$$z = Ax - Ax_0, \;\; r := y - Ax_0.$$

Then, since $r^T z = 0$,

$$\|y - Ax\|_2^2 = \|r - z\|_2^2 = \|r\|_2^2 + \|z\|_2^2 \geq \|r\|_2^2 = \|y - Ax_0\|_2^2,$$

that is, x_0 is a minimum point. □

If the columns of A are linearly independent, that is, if $x \neq 0$ implies $Ax \neq 0$, then the matrix $A^T A$ is nonsingular (and positive definite). If this were not the case, there would exist an $x \neq 0$ satisfying $A^T Ax = 0$, from which

$$0 = x^T A^T A x = \|Ax\|_2^2$$

would yield a contradiction, since $Ax \neq 0$. Therefore the normal equations

$$A^T Ax = A^T y$$

have a unique solution $x = (A^T A)^{-1} A^T y$, which can be computed using the Choleski factorization of $A^T A$.

5.1.2 The use of orthogonalization in solving linear least-squares problems

The problem of determining an $x \in \mathbb{R}^n$ which minimizes

$$\|y - Ax\|_2, \quad (A \in \mathcal{M}_{m \times n}, \;\; m \geq n)$$

can be solved using the orthogonalization techniques (for instance: Gaussian Elimination, Gauss-Jordan algorithm and Choleski Decomposition). Let the matrix $A \equiv: A^{(0)}$ and the vector $y \equiv: y^{(0)}$ be transformed by a sequence of Householder transformations P_i: $A^{(i)} = P_i A^{(i-1)}$, $y^{(i)} = P_i y^{(i-1)}$, where $P_i := I - 2\omega_i \omega_i^\star$, with $\omega_i^\star \omega_i = 1$ and $\omega_i \in \mathbb{C}^n$. The final matrix $A^{(n)}$ has the form

(5.4) $$A^{(n)} = \begin{bmatrix} R \\ 0 \end{bmatrix} \;\text{ with }\; R = \begin{bmatrix} r_{11} & \cdots & r_{1n} \\ & \ddots & \vdots \\ 0 & & r_{nn} \end{bmatrix},$$

since $m \geq n$. Let the vector $h := y^{(n)}$ be the partitioned correspondingly:

$$h = \begin{bmatrix} h_1 \\ h_2 \end{bmatrix}, \quad h_1 \in \mathbb{R}^n, \quad h_2 \in \mathbb{R}^{m-n}. \tag{5.5}$$

The matrix $P = P_n P_{n-1} ... P_1$, being the product of unitary matrices, is unitary itself:

$$P^\star P = P_1^\star ... P_n^\star P_n ... P_1 = I,$$

and satisfies

$$A^{(n)} = PA, \quad h = Py.$$

Unitary transformations leave the Euclidian norm $\|u\|_2$ of a vector u invariant ($\|Pu\|_2^2 = u^\star P^\star P u = u^\star u = \|u\|_2^2$), so

$$\|y - Ax\|_2 = \|P(y - Ax)\|_2 = \|y^{(n)} - A^{(n)}x\|_2.$$

However, from (5.4) and (5.5), the vector $y^{(n)} - A^{(n)}x$ has the structure

$$y^{(n)} - A^{(n)}x = \begin{bmatrix} h_1 - Rx \\ h_2 \end{bmatrix}.$$

Hence $\|y - Ax\|_2$ is minimized if x is chosen so that

$$h_1 = Rx. \tag{5.6}$$

The matrix R has an inverse R^{-1} if and only if the columns $a_1, a_2, ..., a_n$ of A are linearly independent. $Az = 0$ for $z \neq 0$ is equivalent to

$$PAz = 0$$

and therefore to

$$Rz = 0.$$

If we assume that the columns of A are linearly independent, then

$$h_1 = Rx,$$

which is a triangular system, can be solved uniquely for x. This x is, moreover, the unique minimum point for the given least-squares problem. [If the columns of A, and with them of R, are linearly dependent, then, although the value of $\min_x \|y - Ax\|_2$ is uniquely determined, there are many minimum points x].
The size $\|y - Ax\|_2$ of the residual of the minimum point is seen to be

$$\|y - Ax\|_2 = \|h_2\|_2. \tag{5.7}$$

We conclude by mentioning that instead of using unitary transformations, the Gram-Schmidt technique with reorthogonalization can be used to obtain the solution, as should be evident.

5.1.3 The condition of the linear least-squares problem

In this part, we try to show how a minimum point x for the linear least-squares problem

$$\min_x \|y - Ax\|_2 \tag{5.8}$$

changes if the matrix A and the vector y are perturbed. We assume that the columns of A are linearly independent. If the matrix A is replaced by $(A + \Delta A)$, and y is replaced by $y + \Delta y$, then the solution $x = (A^T A)^{-1} A^T y$ of (5.8) changes to

$$x + \Delta x = [(A + \Delta A)^T (A + \Delta A)]^{-1} (A + \Delta A)^T (y + \Delta y).$$

If ΔA is small relative to A, the $[(A+\Delta A)^T(A+\Delta A)]^{-1}$ exists and satisfies, to a first approximation,

$$\begin{aligned}[(A+\Delta A)^T(A+\Delta A)]^{-1} &\doteq (A^TA(I+(A^TA)^{-1}[A^T\Delta A+\Delta A^TA]))^{-1}\\ &\doteq (I-(A^TA)^{-1}[A^T\Delta A+\Delta A^TA])(A^TA)^{-1}.\end{aligned}$$

[to a first approximation, $(I+F)^{-1}\doteq I-F$ if the matrix F is "small" relative to I]. Thus it follows that

$$x+\Delta x \doteq (A^TA)^{-1}A^Ty-(A^TA)^{-1}[A^T\Delta A+\Delta A^TA](A^TA)^{-1}A^Ty \tag{5.9}$$

$$+(A^TA)^{-1}\Delta A^Ty+(A^TA)^{-1}A^T\Delta y.$$

Noting that

$$x=(A^TA)^{-1}A^Ty$$

and introducing the residual

$$r:=y-Ax,$$

it follows immediately from (5.9) that

$$\Delta x \doteq -(A^TA)^{-1}A^T\Delta Ax+(A^TA)^{-1}\Delta A^Tr+(A^TA)^{-1}A^T\Delta y.$$

Therefore, for the Euclidean norm $\|\cdot\|_2$ and the associated matrix norm "lub",

$$\|\Delta x\|_2 \dot{\leq} lub((A^TA)^{-1}A^T)lub(A)\frac{lub(\Delta A)}{lub(A)}\|x\|_2+lub((A^TA)^{-1}A^T)\times \tag{5.10}$$

$$lub(A)\frac{\|y\|_2}{\|Ax\|_2}\frac{\|\Delta y\|_2}{\|y\|_2}\|x\|_2+lub((A^TA)^{-1})lub(A^T)lub(A)\frac{lub(\Delta A^T)}{lub(A^T)}\frac{\|r\|_2}{\|Ax\|_2}\|x\|_2.$$

This approximate bound can be simplified. According to subsection 5.1.2, a unitary matrix P and a upper triangular matrix R can be found such that

$$PA=\begin{bmatrix}R\\0\end{bmatrix},\quad A=P^T\begin{bmatrix}R\\0\end{bmatrix},$$

and it follows that

$$A^TA=R^TR,$$

$$(A^TA)^{-1}=R^{-1}(R^T)^{-1}, \tag{5.11}$$

$$(A^TA)^{-1}A^T=[R^{-1},0]P.$$

If it is observed that

$$lub(C^T)=lub(C)$$

$$lub(PC)=lub(CP)=lub(C),$$

holds for the Euclidean norm, where P is unitary, then (5.10) and (5.11) imply

$$\frac{\|\Delta x\|_2}{\|x\|_2}\dot{\leq} cond(R)\frac{lub(\Delta A)}{lub(A)}+cond(R)^2\frac{\|r\|_2}{\|Ax\|_2}\frac{lub(\Delta A)}{lub(A)}$$

$$+cond(R)\frac{\|y\|_2}{\|Ax\|_2}\frac{\|\Delta y\|_2}{\|y\|_2}.$$

If we define the angle φ by

$$\tan\varphi = \frac{\|r\|_2}{\|Ax\|_2},\ \ 0 \le \varphi < \frac{\pi}{2},$$

then $\|y\|_2/\|Ax\|_2 = (1+\tan^2\varphi)^{1/2}$ because of $y = Ax + r$, $r \perp Ax$. Therefore

$$\frac{\|\Delta x\|_2}{\|x\|_2} \dot{\le} cond(R)\frac{lub(\Delta A)}{lub(A)} + cond(R)^2\tan\varphi\frac{lub(\Delta A)}{lub(A)} + cond(R)\sqrt{1+\tan^2\varphi}\frac{\|\Delta y\|_2}{\|y\|_2}. \tag{5.12}$$

Remark 5.1.1. *According to* (5.12), *it follows that the condition of the least-squares problem depends on cond(R) and the angle φ : If φ is small, say if $cond(R)\tan\varphi \le 1$, then the condition is measured by cond(R). With increasing $\varphi \uparrow \pi/2$ the condition gets worse: it is then measured by $cond(R)^2\tan\varphi$.*

Conclusion. The relation (5.12) shows that the use of the normal equations is not numerically stable if the first term dominates in this relation. Another situation holds if the second term dominates. If $\tan\varphi = \|r\|_2/\|Ax\|_2 \ge 1$, for example, then the use of the normal equations will be numerically stable and will yield results which are comparable to those obtained through the use of orthogonal transformations.

5.1.4 The Moore-Penrose inverse of a matrix

For any arbitrary (complex) $m \times n$ matrix A there is an $n \times m$ matrix A^+, the so-called **Moore-Penrose inverse**, which generalizes in a particular way the classical inverse matrix. It is associated with A in a natural fashion and agrees with the inverse A^{-1} of A in the case $m = n$ and A is nonsingular.

Consider the range space $\mathbf{R(A)}$ and the null space $\mathbf{N(A)}$ of A,

$$\begin{aligned}\mathbf{R(A)} &:= \{Ax \in \mathbb{C}^m | x \in \mathbb{C}^n\},\\ \mathbf{N(A)} &:= \{x \in \mathbb{C}^n | Ax = 0\},\end{aligned}$$

together with their orthogonal complement spaces $\mathbf{R(A)}^\perp \subset \mathbb{C}^m$, $\mathbf{N(A)}^\perp \subset \mathbb{C}^n$. Further, let P be the $n \times n$ matrix which projects $\mathbb{C}^n$ onto $\mathbf{N(A)}^\perp$, and $\bar{P}$ be the $m \times m$ matrix which projects $\mathbb{C}^m$ onto $\mathbf{R(A)}$:

$$\begin{aligned}Px = 0 &\ \Leftrightarrow\ x \in \mathbf{N(A)},\ \ P = P^\star = P^2,\\ \bar{P}y = y &\ \Leftrightarrow\ y \in \mathbf{R(A)},\ \ \bar{P} = \bar{P}^\star = \bar{P}^2.\end{aligned}$$

For each $y \in \mathbf{R(A)}$ there is a uniquely determined $x_1 \in \mathbf{N(A)}^\perp$ satisfying $Ax_1 = y$, i.e., there is a well-defined mapping $f : \mathbf{R(A)} \to \mathbb{C}^n$ with

$$Af(y) = y,\ \ f(y) \in \mathbf{N(A)}^\perp,\ \text{ for all } y \in \mathbf{R(A)}.$$

For given $y \in \mathbf{R(A)}$, there is an x which satisfies $y = Ax$, hence

$$y = A\left(Px + (I-P)x\right) = APx = Ax_1,\ \text{ where } x_1 := Px \in \mathbf{N(A)}^\perp,$$

since $(I-P)x \in \mathbf{N(A)}$. Further, if $x_1, x_2 \in \mathbf{N(A)}^\perp$ with $Ax_1 = Ax_2 = y$, it follows that

$$x_1 - x_2 \in \mathbf{N(A)} \cap \mathbf{N(A)}^\perp = \{0\},$$

which implies that $x_1 = x_2$. f is obviously linear. The composite mapping

$$f \circ \bar{P} :\ y \in \mathbb{C}^m \to f\left(\bar{P}y\right) \in \mathbb{C}^n$$

is well defined and linear, since $\bar{P}y \in \mathbf{R(A)}$, hence it is represented by an $n \times m$ matrix, which is precisely A^+, the Moore-Penrose inverse of A :

$$A^+ y = f\left(\bar{P}y\right) \text{ for all } y \in \mathbb{C}^m.$$

Moreover, one has the following Theorem

Theorem 5.1.2. [149, 100, 112] *Let A be an $m \times n$ matrix. The Moore-Penrose inverse A^+ is an $n \times m$ matrix satisfying*

(1) *$A^+A = P$ is the orthogonal projector $P : \mathbb{C}^n \to \boldsymbol{N(A)}^\perp$ and $AA^+ = \bar{P}$ is the orthogonal projector $\bar{P} : \mathbb{C}^m \to \boldsymbol{R(A)}$.*

(2) *The following formulas holds:*

- **(a)** $A^+A = (A^+A)^\star$,
- **(b)** $AA^+ = (AA^+)^\star$,
- **(c)** $AA^+A = A$,
- **(d)** $A^+AA^+ = A^+$,

Proof. According to the definition of A^+,

$$A^+Ax = f\left(\bar{P}Ax\right) = f(Ax) = Px \text{ for all } x,$$

so that $A^+A = P$. Since $P^\star = P$, the point $2a$) of Theorem 5.1.2 is satisfied. Further, from the definition of f,

$$AA^+y = A\left(f\left(\bar{P}y\right)\right) = \bar{P}y \text{ for all } y \in \mathbb{C}^m,$$

hence $AA^+ = \bar{P}$. Since $\bar{P}^\star = \bar{P}$, the point $2b$) of Theorem 5.1.2 follows too. Finally, for all $x \in \mathbb{C}^n$

$$(AA^+)Ax = \bar{P}Ax = Ax, \text{ according to definition of } \bar{P} \ ,$$

and for all $y \in \mathbb{C}^m$,

$$A^+AA^+y = A^+\bar{P}y = f\left(\bar{P}^2y\right) = f\left(\bar{P}y\right) = A^+y,$$

hence, the points $2c), 2d)$ of Theorem 5.1.2 hold. □

Remark 5.1.2. *The properties $(2 : a - d)$ of Theorem* 5.1.2 *uniquely characterize A^+.*

Furthermore, the following crucial result characterizes the Moore-Penrose inverse of a matrix.

Theorem 5.1.3. [13] *Let A be a matrix of dimension $m \times n$. If Z is a matrix satisfying*

$a^{'})$ $ZA = (ZA)^\star$,

$b^{'})$ $AZ = (AZ)^\star$,

$c^{'})$ $AZA = A$,

$d^{'})$ $ZAZ = Z$,

then the matrix Z is the Moore-Penrose inverse of the matrix A, i.e., $Z = A^+$.

Proof. From $(a)-(d)$ of Theorem 5.1.2 and $(a^{'})-(d^{'})$ we have the following chain of equalities:

$$\begin{aligned} Z = ZAZ &= Z(AA^{+}A)A^{+}(AA^{+}A)Z \qquad \text{from } d^{'}), c) \\ &= (A^{\star}Z^{\star}A^{\star}A^{+\star})A^{+}(A^{+\star}A^{\star}Z^{\star}A^{\star}) \qquad \text{from } a), a^{'}), b), b^{'}) \\ &= (A^{\star}A^{+\star})A^{+}(A^{+\star}A^{\star}) \qquad \text{from } c^{'}) \\ &= (A^{+}A)A^{+}(AA^{+}) \qquad \text{from } a), b) \\ &= A^{+}AA^{+} = A^{+} \qquad \text{from } d)) \end{aligned}$$

□

It follows from Theorem 5.1.3 the following Corollary

Corollary 5.1.1. [13] *For all matrices A,*

$$A^{++} = A, \quad (A^{+})^{\star} = (A^{\star})^{+}.$$

Proof. This holds because $Z := A$ [respectively $Z := (A^{+})^{\star}$] has the properties of $(A^{+})^{+}$ [respectively $(A^{\star})^{+}$] in Theorem 5.1.3.

□

Exploiting these facts, an elegant representation of the solution to the least-squares problem

$$\min_{x} \|Ax - y\|_2$$

can be given with the aid of the Moore-Penrose inverse A^{+}:

Theorem 5.1.4. [96] *The vector $\bar{x} := A^{+}y$ satisfies:*

(a) $\|Ax - y\|_2 \geq \|A\bar{x} - y\|_2$ *for all* $x \in \mathbb{C}^n$.

(b) $\|Ax - y\|_2 = \|A\bar{x} - y\|_2$, *and* $x \neq \bar{x}$ *imply* $\|x\|_2 > \|\bar{x}\|_2$.

In other words, $\bar{x} = A^{+}y$ is the minimum point of the least squares problem which has the smallest Euclidean norm, in the event that the problem does not have a unique minimum point.

Proof. From Theorem 5.1.2, AA^{+} is the orthogonal projector onto $\mathbf{R(A)}$, hence, for all $x \in \mathbb{C}^n$ it follows that

$$Ax - y = u - v$$

$$u := A(x - A^{+}y) \in \mathbf{R(A)}, \quad v := (I - AA^{+})y = y - A\bar{x} \in \mathbf{R(A)}^{\perp}.$$

Consequently, for all $x \in \mathbb{C}^n$

$$\|Ax - y\|_2^2 = \|u\|_2^2 + \|v\|_2^2 \geq \|v\|_2^2 = \|A\bar{x} - y\|_2^2,$$

and $\|Ax - y\|_2 = \|A\bar{x} - y\|_2$ holds precisely if

$$Ax = AA^{+}y.$$

Now, AA^{+} is the projector on $\mathbf{N(A)}^{\perp}$. Therefore, for all x such that $Ax = AA^{+}y$,

$$x = u_1 + v_1, \quad u_1 =: A^{+}Ax = A^{+}AA^{+}y = A^{+}y = \bar{x} \in \mathbf{N(A)}^{\perp},$$

$$v_1 := x - u_1 = x - \bar{x} \in \mathbf{N(A)},$$

from which it follows that $\|x\|_2^2 > \|\bar{x}\|_2^2$ for all $x \in \mathbb{C}^n$ satisfying $x - \bar{x} \neq 0$ and $\|Ax - y\|_2 = \|A\bar{x} - y\|_2$. □

Remark 5.1.3. *(important)*[49]. *If the $m \times n$ matrix A with $m \geq n$ has maximal rank, i.e., $rank(A) = n$, then there is an explicit formula for A^+: it is easily verified that the matrix $Z := (A^\star A)^{-1} A^\star$ has all properties given in Theorem* 5.1.3 *characterizing the Moore-Penrose inverse A^+ so that*

$$A^+ = (A^\star A)^{-1} A^\star.$$

By means of the QR decomposition of A, i.e., $A = QR$ where Q is a unitary matrix and R is a upper triangular matrix, this formula for A^+ is equivalent to

$$A^+ = (R^\star Q^\star Q R)^{-1} R^\star Q^\star = R^{-1} Q^\star.$$

This allows a numerically more stable computation of the Moore-Penrose inverse, $A^+ = R^{-1}Q^\star$.

If $m < n$ and $rank(A) = m$, then because of $(A^+)^\star = (A^\star)^+$, the Moore-Penrose inverse A^+ is given by

$$A^+ = Q(R^\star)^{-1},$$

if the matrix $A^\star$ has the QR decomposition, $A^\star = QR$.

For General $m \times n$ matrices A of arbitrary rank the Moore-Penrose inverse A^+ can be computed by means of the singular value decomposition of A.

5.2 On the convergence of the minimization methods

The purpose of this section is to recall some fundamental results on the minimization problems, which are important in the study of the conjugate gradient methods.

Let $h : \mathbb{R}^n \to \mathbb{R}$ be a function which has a continuous derivative $Dh(x)$ for all $x \in V(\bar{x})$ in a neighborhood $V(\bar{x})$ of $\bar{x}$. We consider the set

$$D(\gamma, x) := \{s \in \mathbb{R}^n |\ \|s\|_2 = 1 \text{ with } Dh(x)s \geq \gamma \|Dh(x)\|_2\} \tag{5.13}$$

of all directions s forming a not-too-large acute angle with the gradient $\nabla h(x)$,

$$\nabla h(x)^T = Dh(x) = \left(\frac{\partial h(x)}{\partial x^1}, ..., \frac{\partial h(x)}{\partial x^n} \right), \text{ where } (x^1, ..., x^n)^T \in \mathbb{R}^n.$$

The following lemma shows, given x, under which conditions a scalar λ and an $s \in \mathbb{R}^n$ exist such that $h(x - \mu s) < h(x)$ for $0 < \mu < \lambda$:

Lemma 5.2.1. [23, 64, 148]. *Let $h : \mathbb{R}^n \to \mathbb{R}$ be a function which has a continuous derivative $Dh(x)$ for all $x \in V(\bar{x})$ in a neighborhood $V(\bar{x})$ of $\bar{x}$. Suppose further that $Dh(\bar{x}) \neq 0$, and let $1 \geq \gamma > 0$. Then there is a neighborhood $U(\bar{x}) \subset V(\bar{x})$ of $\bar{x}$ and a number $\lambda > 0$ such that*

$$h(x - \mu s) \leq h(x) - \frac{\mu\gamma}{4} \|Dh(\bar{x})\|_2$$

for all $x \in U(\bar{x})$, $s \in D(\gamma, x)$, and $0 \leq \mu \leq \lambda$.

Proof. The set

$$U^1(\bar{x}) := \{x \in V(\bar{x}) |\ \|Dh(x) - Dh(\bar{x})\|_2 \leq \frac{\gamma}{4} \|Dh(\bar{x})\|_2\}$$

is nonempty and a neighborhood of $\bar{x}$, since $\|Dh(\bar{x})\|_2 \neq 0$ and $\|Dh(x)\|_2$ is continuous on $V(\bar{x})$. Now, for $x \in U^1(\bar{x})$,

$$\|Dh(x)\|_2 \geq \|Dh(\bar{x})\|_2 - \frac{1}{4}\gamma \|Dh(\bar{x})\|_2 \geq \frac{3}{4} \|Dh(\bar{x})\|_2,$$

so that for $s \in D(\gamma, x)$

$$Dh(x)s \geq \gamma \|Dh(x)\|_2 \geq \frac{3}{4}\gamma \|Dh(\bar{x})\|_2.$$

Choose $\lambda > 0$ with

$$B_{2\lambda} := \{x|\ \|x - \bar{x}\|_2 \leq 2\lambda\} \subset U^1(\bar{x})$$

and define

$$U(\bar{x}) := B_\lambda = \{x|\ \|x - \bar{x}\|_2 \leq \lambda\}.$$

Then for all $x \in U(\bar{x})$, $0 \leq \mu \leq \lambda$, and $s \in D(\gamma, x)$ there is a $0 < \theta < 1$ with

$$\begin{aligned} h(x) - h(x - \mu s) &= \mu Dh(x - \theta\mu s)s \\ &= \mu([Dh(x - \theta\mu s) - Dh(x)]s + Dh(x)s). \end{aligned}$$

Now, $x \in U(\bar{x}) = B_\lambda$ implies $x, x - \mu s, x - \theta\mu s \in B_{2\lambda} \subset U^1$, and therefore

$$[(Dh(x - \theta\mu s) - Dh(\bar{x})) + (Dh(\bar{x}) - Dh(x))]s \leq \frac{1}{2}\gamma \|Dh(\bar{x})\|_2.$$

Using $Dh(x)s \geq \frac{3}{4}\gamma \|Dh(\bar{x})\|_2$, we obtain

$$\begin{aligned} h(x) - h(x - \mu s) &\geq -\frac{\mu\gamma}{2}\|Dh(\bar{x})\|_2 + \frac{3\mu\gamma}{4}\|Dh(\bar{x})\|_2 \\ &= \frac{\mu\gamma}{4}\|Dh(\bar{x})\|_2. \end{aligned}$$

□

Let us consider the following method for minimizing a differentiable function $h : \mathbb{R}^n \to \mathbb{R}$.

(5.14) **Method**

(a) Choose numbers $\gamma_k \leq 1$, $\sigma_k > 0$, $k = 0, 1, ...$, with

$$\inf_k \gamma_k > 0, \ \inf_k \sigma_k > 0,$$

and choose a starting point $x_0 \in \mathbb{R}^n$.

(b) For all $k = 0, 1, ...$, choose an $s_k \in D(\gamma_k, x_k)$ and set

$$x_{k+1} := x_k - \lambda_k s_k$$

where $\lambda_k \in [0, \sigma_k \|Dh(x_k)\|_2]$ is such that

$$h(x_{x+1}) = \min_\mu \{h(x_k - \mu s_k)|\ 0 \leq \mu \leq \sigma_k \|Dh(x_k)\|_2\}.$$

The convergence properties of this method are given by the following theorem:

Theorem 5.2.1. [21, 23, 64]. *Let $h : \mathbb{R}^n \to \mathbb{R}$ be a function, and let $x_0 \in \mathbb{R}^n$ be chosen so that*

(a) *$K := \{x|h(x) \leq h(x_0)\}$ is compact, and*

(b) *h is continuously differentiable in some open sets containing K.*
Then for any sequence $\{x_k\}_k$ defined by a method of the type (5.14)*:*

(1) *$x_k \in K$ for all $k = 0, 1, ...$, $\{x_k\}_k$ has at least one accumulation point $\bar{x}$ in K.*

(2) *Each accumulation point $\bar{x}$ of $\{x_k\}_k$ is a stationary point of h :*

$$Dh(\bar{x}) = 0.$$

Proof. (1) : From the definition of the sequence $\{x_k\}_k$ it follows immediately that the sequence $\{h(x_k)\}_k$ is monotone: $h(x_0) \geq h(x_1) \geq ....$ Hence $x_k \in K$ for all k. K is compact, therefore $\{x_k\}_k$ has at least one accumulate point $\bar{x} \in K$.
(2) : Assume that $\bar{x}$ is an accumulation point of $\{x_k\}_k$, but is not a stationary point of h:

$$Dh(\bar{x}) \neq 0. \tag{5.15}$$

Without loss of generality, let $\lim_{k\to\infty} x_k = \bar{x}$. Let

$$\gamma := \inf_k \gamma_k > 0, \quad \sigma := \inf_k \sigma_k > 0.$$

According to Lemma 5.2.1 there is a neighborhood $U(\bar{x})$ of $\bar{x}$ and a number $\lambda > 0$ satisfying

$$h(x - \mu s) \leq h(x) - \mu \frac{\gamma}{4} \|Dh(\bar{x})\|_2 \tag{5.16}$$

for all $x \in U(\bar{x})$, $s \in D(\gamma, x)$, and $0 \leq \mu \leq \lambda$.
Since $\lim_{k\to\infty} x_k = \bar{x}$, the continuity of $Dh(x)$ together with (5.15) imply the existence of a k_0 such that for all $k \geq k_0$

(a) $x_k \in U(\bar{x})$,

(b) $\|Dh(x_k)\|_2 \geq \frac{1}{2}\|Dh(\bar{x})\|_2$.

Let $\Lambda := \min\{\lambda, \frac{1}{2}\sigma\|Dh(\bar{x})\|_2\}$, $\epsilon := \Lambda\frac{\gamma}{4}\|Dh(\bar{x})\|_2 > 0$. Since $\sigma_k \geq \sigma$, it follows that $[0, \Lambda] \subset [0, \sigma_k\|Dh(x_k)\|_2]$ for all $k \geq k_0$. Therefore, from the definition of x_{k+1},

$$h(x_{k+1}) \leq \min_\mu \{h(x_k - \mu s_k) | 0 \leq \mu \leq \Lambda\}.$$

Since $\Lambda \leq \lambda$, $x_k \in U(\bar{x})$, $s_k \in D(\gamma_k, x_k) \subset D(\gamma, x_k)$, (5.16) implies that

$$h(x_{k+1}) \leq h(x_k) - \frac{\Lambda\gamma}{4}\|Dh(\bar{x})\|_2 = h(x_k) - \epsilon$$

for all $k \geq k_0$. This means that $\lim_{k\to\infty} h(x_k) = -\infty$, which contradicts $h(x_k) \geq h(x_{k+1}) \geq ... \geq h(\bar{x})$. Hence, $\bar{x}$ is a stationary point of h. □

Remark 5.2.1. *Step (b) of Method (5.14) is known as the line research. Even though the method given by (5.14) is quite general, its practical application is limited by the fact that the line research must be exact, i.e, it requires that the exact minimum point of the function*

$$\varphi(\mu) := h(x_k - \mu s_k)$$

be found on the interval $[0, \sigma_k\|Dh(x_k)\|_2]$ in order to determine x_{k+1}. Generally, a great deal of effort is required to obtain even an approximate minimum point.

The following variant of (5.14) has the virtue that in step (b) the exact minimization is replaced by an inexact line research, in particular by a finite search process ("Armijo line search"): Armijo line search DF

(5.17) **Armijo Method**

(a) Choose numbers $\gamma_k \leq 1$, σ_k, $k = 0, 1, ...$, so that

$$\inf_k \gamma_k > 0, \quad \inf_k \sigma_k > 0.$$

Choose a starting point $x_0 \in \mathbb{R}^n$.

(b) For each $k = 0, 1, ...$, obtain x_{k+1} from x_k as follows:

(α) select

$$s_k \in D(\gamma_k, x_k),$$

define

$$\rho_k := \sigma_k \|Dh(x_k)\|_2, \quad h_k(\mu) = h(x_k - \mu s_k),$$

and determine the smallest integer $j \geq 0$ such that

$$h_k(\rho_k 2^{-j}) \leq h_k(0) - \rho_k 2^{-j} \frac{\gamma_k}{4} \|Dh(x_k)\|_2.$$

(β) Determine $\bar{i} \in \{0, 1, ..., j\}$, such that $h_k(\rho_k 2^{-\bar{i}})$ is minimum and let $\lambda_k := \rho_k 2^{-\bar{i}}$, $x_{k+1} := x_k - \lambda_k s_k$.

NB: $h(x_{k+1}) \leq \min_{1 \leq i \leq j} h_k(\rho_k 2^{-i})$.

Theorem 5.2.2. *Under the hypotheses of Theorem* 5.2.1*, each sequence* $\{x_k\}_k$ *produced by a method of the type* (5.17) *satisfies the conclusion of Theorem* 5.2.1.

Proof. We assume as before that $\bar{x}$ is an accumulation point of a sequence $\{x_k\}_k$ defined in (5.17), but not a stationary point, i.e.,

$$Dh(\bar{x}) \neq 0.$$

Again, without loss of generality, let $\lim_{k \to \infty} x_k = \bar{x}$. Also let $\sigma := \inf_k \sigma_k > 0$, $\gamma := \inf_k \gamma_k > 0$. According to Lemma 5.2.1 there is a neighborhood $U(\bar{x})$ and a number $\lambda > 0$ such that

$$h(x - \mu s) \leq h(x) - \mu \frac{\gamma}{4} \|Dh(\bar{x})\|_2 \tag{5.18}$$

for all $x \in U(\bar{x})$, $s \in D(\gamma, x)$, $0 \leq \mu \leq \lambda$. Again, the fact that $\lim_{k \to \infty} x_k = \bar{x}$, that $Dh(x)$ is continuous, and that $Dh(\bar{x}) \neq 0$ imply the existence of a k_0 such that

$$x_k \in U(\bar{x}), \tag{5.19}$$

$$\|Dh(x_k)\|_2 \geq \frac{1}{2} \|Dh(\bar{x})\|_2 \tag{5.20}$$

for all $k \geq k_0$.
We need to show that there exists $\epsilon > 0$ for which

$$h(x_{k+1}) \leq h(x_k) - \epsilon \text{ for all } k \geq k_0.$$

Note first that (5.19) – (5.20) and $\gamma_k \geq \gamma$ imply

$$\gamma_k \|Dh(x_k)\|_2 \geq \frac{\gamma}{2}\|Dh(\bar{x})\|_2 \text{ for all } k \geq k_0.$$

Consequently, according to the definition of x_{k+1} and for $j \geq 0$

$$h(x_{k+1}) \leq h_k(\rho_k 2^{-j}) \leq h(x_k) - \rho_k 2^{-j}\frac{\gamma_k}{4}\|Dh(x_k)\|_2 \leq h(x_k) - \rho_k 2^{-j}\frac{\gamma}{8}\|Dh(\bar{x})\|_2. \tag{5.21}$$

Now, let $\bar{j} \geq 0$ be the smallest integer satisfying

$$h_k(\rho_k 2^{-\bar{j}}) \leq h(x_k) - \rho_k 2^{-\bar{j}}\frac{\gamma}{8}\|Dh(\bar{x})\|_2. \tag{5.22}$$

According to (5.21), $\bar{j} \leq j$, and the definition of x_{k+1}, we have

$$h(x_{k+1}) \leq h_k(\rho_k 2^{-\bar{j}}). \tag{5.23}$$

There are two cases:
Case 1 : $\bar{j} = 0$. Note that $\rho_k = \sigma_k\|Dh(x_k)\|_2 \geq \frac{1}{2}\sigma\|Dh(\bar{x})\|_2$. Then (5.22) and (5.23) imply

$$h(x_{k+1}) \leq h(x_k) - \rho_k\frac{\gamma}{8}\|Dh(\bar{x})\|_2 \leq h(x_k) - \frac{\sigma\gamma}{16}\|Dh(\bar{x})\|_2^2 = h(x_k) - \epsilon_1.$$

with $\epsilon_1 > 0$ independent of x_k.
Case 2 : $\bar{j} > 0$. From the minimality of $\bar{j}$, we have

$$h_k(\rho_k 2^{-(\bar{j}-1)}) > h(x_k) - \rho_k 2^{-(\bar{j}-1)}\frac{\gamma}{8}\|Dh(\bar{x})\|_2 \geq h(x_k) - \rho_k 2^{-(\bar{j}-1)}\frac{\gamma_k}{4}\|Dh(x_k)\|_2.$$

Because $x_k \in U(\bar{x})$ and $s_k \in D(\gamma_k, x_k) \subset D(\gamma, x_k)$, it follows immediately from (5.16) that

$$\rho_k 2^{-(\bar{j}-1)} > \lambda.$$

Combining this with (5.22) and (5.23) yields

$$h(x_{k+1}) \leq h_k(\rho_k 2^{-\bar{j}}) \leq h(x_k) - \frac{\lambda\gamma}{16}\|Dh(\bar{x})\|_2 = h(x_k) - \epsilon_2$$

with $\epsilon_2 > 0$ independent of x_k. Hence, for $\epsilon = \min\{\epsilon_1, \epsilon_2\}$

$$h(x_{k+1}) \leq h(x_k) - \epsilon$$

for all $k \geq k_0$, contradicting the fact that $h(x_k) \geq h(\bar{x})$ for all k. Therefore $\bar{x}$ is a stationary point of h. □

5.3 Techniques of construction of the Krylov spaces: The method of Lanczos

In this section, we present the techniques of construction of Krylov spaces and recall the procedure of reduction of a Hermitian matrix to the Tridiagonal form, all based on the Lanczos method.

5.3.1 Techniques of construction of Krylov spaces

Krylov sequences of vectors $q, Aq, A^2q, ...$ belonging to an $n \times n$ matrix and a starting vector $q \in \mathbb{C}^n$ are used for the derivation of the Frobenius normal form of a general matrix. They also play an important role in the Lanczos method (1950) (see [95]) for reducing a Hermitian matrix to tridiagonal form. Closely related to such a sequence of vectors is a sequence of subspaces of $\mathbb{C}^n$

$$K_i(q, A) := span[q, Aq, A^2q, ..., A^{i-1}q], \;\; i \geq 1, \;\; K_0(q, A) := \{0\},$$

called Krylov spaces: $K_i(q, A)$ is the subspace spanned by the first i vectors of the sequence $\{A^j q\}_{j \geq 0}$. If we denote by m the largest index i for which $q, Aq, A^2q, ..., A^{i-1}q$ are still linearly independent, that is, $dim\, K_i(q, A) = i$. Then $m \leq n$, $A^m q \in K_m(q, A)$, the vectors $q, Aq, A^2q, ..., A^{m-1}q$ form a basis of $K_m(q, A)$, and therefore $AK_m(q, A) \subset K_m(q, A)$: the Krylov space $K_m(q, A)$ is A-invariant and the map $x \mapsto \Phi(x) := Ax$ describes a linear map of $K_m(q, A)$ into itself.

Theorem 5.3.1. [95, 108, 115] *There exists polynomials $p_j \in \overline{\Pi}_j$ where $\overline{\Pi}_j := \{p|\; p(x) = x^j + a_1 x^{j-1} + ... + a_j\}$, $j = 0, 1, 2, ...$, such that*

$$(p_i, p_k) = 0 \;\; for\; i \neq k. \tag{5.24}$$

These polynomials are uniquely defined by the recursions

$$p_0(x) \equiv 1, \tag{5.25}$$

$$p_{i+1}(x) \equiv (x - \delta_{i+1}) p_i(x) - \gamma_{i+1}^2 p_{i-1}(x) \;\; for\; i \geq 0, \tag{5.26}$$

where $p_{-1}(x) := 0$ and

$$\delta_{i+1} := (xp_i, p_i)/(p_i, p_i) \;\; for\; i \geq 0, \tag{5.27}$$

$$\gamma_{i+1}^2 := \begin{cases} 1 & for\; i = 0 \\ (p_i, p_i)/(p_{i-1}, p_{i-1}) & for\; i \geq 1. \end{cases} \tag{5.28}$$

Proof. The polynomials can be constructed recursively by a technique known as Gram-Schmidt orthogonalization. Clearly $p_0(x) \equiv 1$. Suppose then, as an induction hypothesis, that all orthogonal polynomials p_j with the above properties have been constructed for $j \leq i$ and have been shown to be unique. We want to show that there exists a unique polynomial $p_{i+1} \in \overline{\Pi}_{i+1}$ with

$$(p_{i+1}, p_j) = 0 \;\; \text{for } j \leq i, \tag{5.29}$$

and that this polynomial satisfies (5.26). Any polynomial $p_{i+1} \in \overline{\Pi}_{i+1}$ can be written uniquely in the form

$$p_{i+1}(x) \equiv (x - \delta_{i+1}) p_i(x) + c_{i-1} p_{i-1}(x) + c_{i-2} p_{i-2}(x) + ... + c_0 p_0(x),$$

because its leading coefficient and those of the polynomials p_j, $j \leq i$, have value 1. Since $(p_j, p_k) = 0$ for all $j, k \leq i$ with $j \neq k$, (5.29) holds if and only if

$$(p_{i+1}, p_i) = (xp_i, p_i) - \delta_{i+1}(p_i, p_i) = 0, \tag{5.30}$$

$$(p_{i+1}, p_{j-1}) = (xp_{j-1}, p_i) + c_{j-1}(p_{j-1}, p_{j-1}) = 0, \;\; \text{for } j \leq i. \tag{5.31}$$

The condition (5.32) below

(5.32)
$$\begin{aligned} &\text{For polynomials } s(x) \text{ which are nonnegative on } [a,b], \\ &\textstyle\int_a^b \omega(x)s(x)dx = 0 \text{ implies } s(x) \equiv 0 \end{aligned}$$

(where $\omega(x)$ is a given nonnegative weight function on the interval $[a,b]$) with p_i^2 and p_{j-1}^2, respectively, in the role of the nonnegative polynomial s, rules out $(p_i, p_i) = 0$ and $(p_{j-1}, p_{j-1}) = 0$ for $1 \leq j \leq i$. Therefore, the equations (5.30) – (5.31) can be solved uniquely. (5.30) gives (5.27). By the induction hypothesis,

$$p_j(x) \equiv (x - \delta_j)p_{j-1}(x) - \gamma_j^2 p_{j-2}(x)$$

for $j \leq i$. From this, by solving for $xp_{j-1}(x)$, we have $(xp_{j-1}, p_i) = (p_j, p_i)$ for $j \leq i$, so that

$$c_{j-1} = -\frac{(p_j, p_i)}{(p_{j-1}, p_{j-1})} = \begin{cases} -\gamma_{i+1}^2 & \text{for } j = i; \\ 0 & \text{for } j < i. \end{cases}$$

in view of (5.30) – (5.31). Thus (5.26) has been established for $i+1$. □

The idea of the Lanczos method is closely related: Here, the map Φ is described with respect to a special orthonormal basis $q_1, q_2, ..., q_m$ of $K_m(q, A)$, where the q_j are chosen such that for all $i = 1, 2, ..., m$, the vectors $q_1, q_2, ..., q_i$ form an orthonormal basis of $K_i(q, A)$. If $A = A^\star$ is a Hermitian $n \times n$ matrix, then such a basis is easily constructed for a given starting vector q. We assume $q \neq 0$ in order to exclude the trivial case and suppose in addition that $\|q\|_2 = 1$. It follows from this the following result.

Proposition 5.3.1. [95, 108, 115]. *There is a three-term recursion formula for the vectors q_i [similar recursions are known for orthogonal polynomials, cf. Theorem* 5.3.1*].*

(5.33)
$$\begin{aligned} &q_1 := q, \;\; \gamma_1 q_0 := 0, \\ &Aq_i = \gamma_i q_{i-1} + \delta_i q_i + \gamma_{i+1} q_{i+1} \;\; \textit{for } i \geq 1, \end{aligned}$$

where

(5.34)
$$\begin{aligned} &\delta_i := q_i^\star A q_i, \\ &\gamma_{i+1} := \|r_i\|_2 \;\; \textit{with } r_i := Aq_i - \delta_i q_i - \gamma_i q_{i-1}, \\ &q_{i+1} := r_i/\gamma_{i+1}, \;\; \textit{if } \gamma_{i+1} \neq 0. \end{aligned}$$

Here, all coefficients γ_i, δ_i are real. The recursion breaks off with the first index $i := i_0$ with $\gamma_{i+1} = 0$, and then the following holds

$$i_0 = m = \max_i dim\, K_i(q, A).$$

Proof. We show (5.33) – (5.34) by mathematical induction over i. Clearly, since $\|q\|_2 = 1$, the vector $q_1 := q$ provides an orthonormal basis for $K_1(q, A)$. Assume now that for some $j \geq 1$, the vectors $q_1, q_2, ..., q_i$ are given, so that (5.33) – (5.34) and

$$span[q_1, q_2, ..., q_i] = K_i(q, A)$$

hold for all $i \leq j$, and that $r_i \neq 0$ in (5.34) for all $i < j$. We show first that these statements are also true for $j+1$, if $r_j \neq 0$. In fact, if $r_j \neq 0$ then $\gamma_{j+1} \neq 0$, δ_j and q_{j+1} are well defined

by (5.34), and $\|q_{i+1}\|_2 = 1$. The vector q_{j+1} is orthogonal to all q_i with $i \leq j$: This holds for $i = j$, since $\gamma_{j+1} \neq 0$, because

$$Aq_i = \gamma_i q_{i-1} + \delta_i q_i + \gamma_{i+1} q_{i+1}$$

from the definition of δ_j, and using the induction hypothesis

$$\gamma_{j+1} q_j^\star q_{j+1} = q_j^\star A q_j - \delta_j q_j^\star q_j = 0.$$

For $i = j - 1$, the same reasoning and $A = A^\star$ first give

$$\gamma_{j+1} q_{j-1}^\star q_{j+1} = q_{j-1}^\star A q_j - \gamma_j q_{j-1}^\star q_{j-1} = (Aq_{j-1})^\star q_j - \gamma_j.$$

The orthogonality of the q_i for $i \leq j$ and $Aq_{j-1} = \gamma_{j-1} q_{i-2} + \delta_{j-1} q_{j-1} + \gamma_j q_j$ then imply $(Aq_{j-1})^\star q_j = \bar{\gamma_j} = \gamma_j$, and therefore $q_{j-1}^\star q_{j+1} = 0$. For $i < j - 1$ we get the same result with the aid of $Aq_i = \gamma_i q_{i-1} + \delta_i q_i + \gamma_{i+1} q_{i+1}$:

$$\gamma_{j+1} q_i^\star q_{j+1} = q_i^\star A q_j = (Aq_i)^\star q_j = 0.$$

Finally, since $\text{span}[q_1, q_2, ..., q_i] = K_i(q, A) \subset K_j(q, A)$ for $i \leq j$, we also have

$$Aq_j \in K_{j+1}(q, A),$$

which implies by (5.34)

$$q_{j+1} \in span[q_{j-1}, q_j, Aq_j] \subset K_{j+1}(q, A),$$

and therefore $span[q_1, q_2, ..., q_{j+1}] \subset K_{j+1}(q, A)$. Since the orthonormal vectors $q_1, q_2, ..., q_{j+1}$ are linearly independent and $dim\, K_{j+1}(q, A) \leq j + 1$, we obtain

$$K_{j+1}(q, A) = span[q_1, q_2, ..., q_{j+1}].$$

This also shows $j + 1 \leq m = \max_i dim\, K_i(q, A)$, and $i_0 \leq m$ for the break-off index i_0 of (5.33) – (5.34). On the other hand, by the definition of i_0

$$Aq_{i_0} \in span[q_{i_0-1}, q_{i_0}] \subset span[q_1, q_2, ..., q_{i_0}] = K_{i_0}(q, A),$$

so that, because

$$Aq_i \in span[q_1, q_2, ..., q_{i+1}] = K_{i+1}(q, A) \subset K_{i_0}(q, A) \text{ for } i < i_0,$$

we get the A-invariance of $K_{i_0}(q, A)$, $AK_{i_0}(q, A) \subset K_{i_0}(q, A)$. Therefore $i_0 \geq m$, since $K_m(q, A)$ is the first A-invariant subspace among the $K_i(q, A)$. This finally shows $i_0 = m$, and the proof is complete. □

5.3.2 Reduction of a Hermitian matrix to Tridiagonal form

The recursion (5.33) – (5.34) can be written in terms of the matrices

$$Q_i := [q_1, q_2, ..., q_i], \quad J_i := \begin{bmatrix} \delta_1 & \gamma_2 & & 0 \\ \gamma_2 & \delta_2 & \ddots & \\ & \ddots & \ddots & \gamma_i \\ 0 & & \gamma_i & \delta_i \end{bmatrix}, \quad 1 \leq i \leq m,$$

as a matrix equation

(5.35) $$AQ_i = Q_i J_i + [0, ..., 0, \gamma_{i+1} q_{i+1}] = Q_i J_i + \gamma_{i+1} q_{i+1} e_i^T, \quad i = 1, 2, ..., m,$$

where $e_i = [0, ..., 0, 1]^T \in \mathbb{R}^i$ is the i-th axis vector of $\mathbb{R}^i$. This equation is easily verified by comparing the j-th columns, $j = 1, 2, ..., i$, on both sides. Note that the $n \times i$ matrices Q_i have orthonormal columns, $Q_i^\star Q_i = I_i$ and the J_i are real symmetric tridiagonal matrices. Since $i = m$ is the first index with $\gamma_{m+1} = 0$, the matrix J_m is irreducible, and the matrix equation (5.35) reduces to

$$AQ_m = Q_m J_m$$

where $Q_m^\star Q_m = I_m$. Any eigenvalue of J_m is also an eigenvalue of A, since $J_m z = \lambda z$, $z \neq 0$ implies $x := Q_m z \neq 0$ and

$$Ax = AQ_m z = Q_m J_m z = \lambda Q_m z = \lambda x.$$

If $m = n$, i.e., if the method does not terminate prematurely with an $m < n$, then Q_n is a unitary matrix, and the tridiagonal matrix $J_n = Q_n^\star A Q_n$ is similar to A.

Given any vector $q =: q_1$ with $\|q\|_2 = 1$, the method of Lanczos consists of computing the numbers $\gamma_i, \delta_i, i = 1, 2, ...m, \gamma_1 = 0$, and the tridiagonal matrix J_m by means of (5.33)−(5.34). Subsequently, one may compute the eigenvalues and eigenvectors of J_m (and thereby those of A). Concerning the implementation of the method, the following remarks are in order:

Remark 5.3.1. *The number of operations can be reduced by introducing an auxiliary vector defined by*

$$u_i := Aq_i - \gamma_i q_{i-1}.$$

Then $r_i = u_i - \delta_i q_i$, *and the number*

$$\delta_i = q_i^\star A q_i = q_i^\star u_i$$

can also be computed from u_i, *since* $q_i^\star q_{i-1} = 0$.

Remark 5.3.2. *It is not necessary to store the vectors* q_i *if one is not interested in the eigenvectors of A: In order to carry out* (5.33) − (5.34) *only two auxiliary vectors* $u, v \in \mathbb{C}^n$ *are needed, where initially* $v := q$ *is the given starting vector with* $\|q\|_2 = 1$. *Within the following program, which implements the Lanczos algorithm for a given Hermitian* $n \times n$ *matrix* $A = A^\star$, v_k *and* w_k, $k = 1, 2, ..., n$, *denote the components of* v *and* w, *respectively:*

progam

$w := 0$; $\gamma_1 := 1$; $i := 1$;

1: ***if*** $\gamma_i \neq 0$ ***then***

begin if $i \neq 1$ ***then***

for $k := 1$ ***step*** 1 ***until*** n ***do***

begin $t := v_k$; $v_k := w_k/\gamma_i$; $w_k := -\gamma_i t$ ***end*** ;

$w := Av + w$; $\delta_i := v^\star w$; $w = w - \delta_i v$;

m:=i; $i := i + 1$; $\gamma_i := \sqrt{w^\star v}$;

goto *1;*

end;

Each step $i \to i+1$ requires about $5n$ scalar multiplications and one multiplication of the matrix A with a vector. Therefore, the method is inexpensive if A is sparse, so that it is particularly valuable for solving the eigenvalue problem for large sparse matrices $A = A^\star$.

Remark 5.3.3. *In theory, the method is finite: it stops with the first index $i = m \leq n$ with $\gamma_{i+1} = 0$. However, because of the influence of roundoff, one will rarely find a computed $\gamma_{i+1} = 0$ in practice. Yet, it is usually not necessary to perform many steps of the method until one finds a zero or a very small γ_{i+1} : The reason is that, under weak assumptions, the largest and smallest eigenvalues of J_i converge very rapidly with increasing i toward the largest and smallest eigenvalues of A [Kaniel-Paige theory: see Kaniel* (1966, [87]), *Paige* (1975, [107]) *and Saad* (1980, [115, 116])*]. Therefore, if one is only interested in the extreme eigenvalues of A (which is quite frequently the case in applications), only relatively few steps of Lanczos' method are necessary to find a J_i, $i << n$, with extreme eigenvalues that already approximate the extreme eigenvalues of A to machine precision.*

Remark 5.3.4. *The method of Lanczos will generate orthogonal vectors q_i only in theory: In practice, due to roundoff, the vectors $\tilde{q}_i$ actually computed become less and less orthogonal as i increases. This defect could be corrected by reorthogonalizing a newly computed vector $\hat{q}_{i+1}$ with respect to all previous vectors $\tilde{q}_j$, $j \leq i$, that is, by replacing $\hat{q}_{i+1}$ by*

$$\tilde{q}_{i+1} := \hat{q}_{i+1} - \sum_{j=1}^{i} (\tilde{q}_j^\star, \hat{q}_{i+1}) \tilde{q}_j.$$

However, reorthogonalization is quite expensive: The vectors $\tilde{q}_i$ have to be stored, and step i of the Lanczos method now requires $O(i \cdot n)$ operations instead of $O(n)$ operations as before. But it is possible to avoid a full reorthogonalization to some extent and still obtain very good approximations for the eigenvalues of A in spite of the difficulties mentioned. Details can be found in the following literature, which also contains a systematic investigation of the interesting numerical properties of the Lanczos method: Paige (1971, [107]), *Parlett and Scott* (1979, [108]) *and Cullum and Willoughby* (1985, [47]), *where one can also find programs.*

Conclusion

We have presented in this chapter some fundamental results in view of studying the Krylov Space Methods. Section 5.2 is particularly exploited in the part on the conjugate gradient (CG) method while the other sections find their use in the parts: generalized minimal residual (GMRES) method, quasi-minimal residual (QMR) method, Bi-conjugate gradient (BCG) method and Bi-conjugate gradient STAB (Bi-GCSTAB) algorithm. Finally, subsection 5.1.1 also finds its utility in chapters 6 and 9.

Chapter Six

Krylov Space Methods and General Idea on Multigrid Methods

6.1 Introduction

We present in this chapter powerful Iterative methods for solving of large systems of linear equations $Ax = b$ in which A is an $n \times n$ (real) nonsingular matrix. In general, one obtains such systems by using difference methods or finite element methods for solving boundary value problems in partial differential equations. The Krylov space methods start with an initial vector $x^{(0)}$ and subsequently produce a sequence of vectors

$$x^{(0)} \to x^{(1)} \to x^{(2)} \to ... \to x^{(m)} \to ...$$

which converge toward the desired solution $x = A^{-1}b$. The general characteristic of these methods is that the methods, in exact arithmetic, terminate with the exact solution x_m after at most n steps, $m \leq n$. The Krylov space methods generate iterates x_k that approximate the solution of linear equations $Ax = b$ best among all vectors x_j such that $x_j - x_0$ belong to $K_k(r_0, A)$, where $K_k(r_0, A)$ is the Krylov space belonging to the matrix A and the starting vector $r_0 := b - Ax_0$ given by the residual of x_0. Because of roundoff errors, these methods do not terminate with the desired solution after finitely many steps. As in true iterative methods, an infinite number of steps needs to be carried out to speed of convergence of the iterates x_k. The amount of work per step $x^{(k)} \to x^{(k+1)}$ roughly equals to that of multiplying the matrix A by a vector. For this reason, these methods are advantageous for sparse unstructured matrices A as they occur, e.g: in network calculations, but are not recommended for dense matrices or band matrices. Throughout this chapter, we restrict our study on the following methods.

i. The conjugate gradient (CG) method proposed by Hestenes and Stiefel (1952, [80]) for systems with a symmetric positive definite matrix.

ii. The generalized minimal residual (GMRES) method of Saad and Schultz (1986, [116]) (more expensive) but is defined for general linear systems with a nonsymmetric nonsingular matrix.

iii. The quasi-minimal residual method (QMR method) of Freud and Nachtigal (1991, [65]), for solving arbitrary sparse linear systems of equations. This methods is based on the more efficient (but numerically more sensitive) biorthogonalization algorithm of Lanczos (1950, [95]), provides non-orthogonal bases $v_1, v_2, ..., v_k$ for the Krylov spaces $K_k(r_0, A)$ of dimension k. Using these bases, one can compute iterates $x_k \in x_0 + K_k(r_0, A)$ with an approximately minimal residual.

iv. The biconjugate gradient (Bi-CG) algorithm due to Lanczos (1950, [95]) and thoroughly studied by Fletcher (1976, [63]) is also a method for solving linear systems of equations with an arbitrary matrix A. It is an inexpensive, natural generalization of the cg-algorithm, and also generates iterates $x_k \in x_0 + K_k(r_0, A)$.

With regard to the applicability of Krylov space methods, the same remarks apply as for the true iterative methods [see Varga (2000), Young and Axelsson (1994), and Saad (1996, [114])].

However, there exists other methods for solving certains special systems of linear equations arising, for instance, with solution of so-called model problem (the poisson problem on a rectangle) which give the solution after finitely many steps and are superior to most iterative methods. Such methods are the algorithm of Buneman (1968), and the Fourier methods that use the FFT-algorithm of trigonometric interpolation. For a detailed treatment, see cf. J.Stoer and R.Bulirsch [149].

Nowadays, the very large systems of equations are related with the solution of boundary-value problems of partial differential equations by finite element techniques and are mainly solved by Multigrid methods. We describe in section 6.6 only concepts of these important iterative methods, using the context of a boundary value problem for ordinary differential equations. For thorough treatments of multigrid methods, which are closely connected to the numerics of partial differential equations, one can read the special literature of this subject, e.g, Hackbusch (1985, [78]), Braess (1997, [18]), Bramble (1993, [19]), Quarteroni and Valli (1997).

6.2 Conjugate gradient method (cg-method)

In this section, we only work with the matrices A, which are symmetric positive definite. These matrices define a vector norm $\|x\|_A := (x^T Ax)^{1/2}$, and the cg-method generates a vector sequence $x_k \in x_0 + K_k(r_0, A)$ with the minimality property

$$\|x_k - \bar{x}\|_A = \min_{x \in x_0 + K_k(r_0, A)} \|x - \bar{x}\|_A.$$

An important role in this method, is played by A-conjugate vectors $p_k \in \mathbb{R}^n$, $k = 0, 1, 2, ...$

$$p_i^T A p_k = 0 \quad \text{for} \quad i \neq k,$$

that span the Krylov spaces $K_k(r_0, A)$, i.e.,

$$span[p_0, p_1, ..., p_{k-1}] = K_k(r_0, A), \quad k = 1, 2, ...$$

Now, let us consider a system of linear equations

$$Ax = b \tag{6.1}$$

where $b \in \mathbb{R}^n$ and A is a real symmetric positive definite $n \times n$ matrix. Furthermore, defining the quadratic functional $h : \mathbb{R}^n \to \mathbb{R}$ by:

$$h(x) = \frac{1}{2}(b - Ax)^T A^{-1}(b - Ax)$$

and developing the quantity $\frac{1}{2}(b - Ax)^T A^{-1}(b - Ax)$, one obtains $h(x) = \frac{1}{2}\|x - \bar{x}\|_A^2$, where $\bar{x} = A^{-1}b$. So, h is minimized by $\bar{x}$ over $\mathbb{R}^n$, i.e.,

$$0 = h(\bar{x}) = \min_{x \in \mathbb{R}^n} h(x).$$

This part suggests using the methods 5.14, 5.17 (studied in section 5.2), in which the sequence $x_0 \to x_1 \to x_2 \to ... \to x_k \to ...$ is found by one-dimensional minimization of h in the direction of the gradient, i.e., find x_{k+1} such that

$$h(x_{k+1}) = \min_u h(x_k + u r_k)$$

$$\text{with } r_k := -\nabla h(x_k) = -Ax_k + b.$$

At the step $x_k \to x_{k+1}$, a $(k+1)$-dimensional minimization is carried out:

$$(6.2) \qquad \begin{cases} x_{k+1}: & h(x_{k+1}) = \min\limits_{u_0,u_1,\ldots,u_k} h\left(x_k + \sum\limits_{j=0}^{k} u_j r_j\right) \\ & r_j := b - Ax_j \text{ for } j \le k. \end{cases}$$

From (6.2), it follows that the vector x_{k+1} can be computed easily. The r_j obtained are orthogonal, as long as $r_k \neq 0$. Since $\mathbb{R}^n$ is of dimension n, one deduces in exact arithmetic that there exists a first nonnegative integer $m \le n$ such that $r_m = 0$. So, the corresponding x_m is the desired solution of (6.2), (see section 5.2).

(6.3) **Conjugate-gradient algorithm**

Initialization: Choose $x_0 \in \mathbb{R}^n$, and put

$$p_0 := r_0 := b - Ax_0$$

For $k = 0, 1, 2, ...$

i) If $p_k = 0$, set $m := k$ and stop: x_k is the solution of $Ax = b$. Otherwise,

ii) compute:

$$a_k := \frac{r_k^T r_k}{p_k^T A p_k}, \quad x_{k+1} := x_k + a_k p_k,$$

$$r_{k+1} := r_k - a_k A p_k, \quad b_k := \frac{r_{k+1}^T r_{k+1}}{r_k^T r_k},$$

$$p_{k+1} := r_{k+1} + b_k p_k.$$

Remark 6.2.1. *In this method, only four vectors: x_k, r_k, p_k, and Ap_k need to be stored. At each iteration step, only one matrix multiplication: Ap_k is required, the remaining works amounts to the calculation of six inner products in $\mathbb{R}^n$. So, the total computation effort, for sparse matrices is modest.*

Theorem 6.2.1. *Let A be a symmetric positive definite real $n \times n$ matrix and let $b \in \mathbb{R}^n$. Then for each initial vector $x_0 \in \mathbb{R}^n$, there is a smallest nonnegative integer m, $m \le n$, such that $p_m = 0$. The vectors x_k, p_k, r_k, $k \le m$ generated by the cg-method (6.3) have the following properties:*

(a) *$Ax_m = b$ (The method produces the exact solution of the equation: $Ax = b$ after at most n steps).*

(b) *$r_j^T p_i = 0$ for $0 \le i < j \le m$.*

(c) *$r_i^T p_i = r_i^T r_i$ for $i = 0, 1, 2, ..., m$.*

(d) *$p_i^T A p_j = 0$ for $0 \le i < j \le m$, and $p_j^T A p_j > 0$ for $j = 0, 1, ..., m-1$.*

(e) *$r_i^T r_j = 0$ for $0 \le i < j \le m$, and $r_j^T r_j > 0$ for $j = 0, 1, ..., m-1$.*

(f) *$r_i = b - Ax_i$ for $i = 0, 1, 2, ..., m$.*

Proof. By mathematical induction on k, we show that the following statement (A_k) holds for all $K = 0, 1, ...m$, where m is the first index with $p_m = 0$.

$$(A_k)\quad \begin{array}{ll} (N_1) & r_j^T p_i = 0 \text{ for } 0 \le i < j \le k \\ (N_2) & r_i^T r_i > 0 \text{ for } 0 \le i < k,\ r_i^T p_i = r_i^T r_i \text{ for } 0 \le i \le k \\ (N_3) & p_i^T A p_j = 0 \text{ for } 0 \le i < j \le k,\ p_i^T A p_i > 0 \text{ for } 0 \le i < k \\ (N_4) & r_i^T r_j = 0 \text{ for } 0 \le i < j \le k \\ (N_5) & r_i = b - A x_i \text{ for } i = 0, 1, 2, ..., k \end{array}$$

(A_0) is trivially true. We assume inductively that, (A_k) holds for some $0 \le k < m$ and we show $(A_{k+1})(N_1)$. Since $p_k \neq 0$ and A is symmetric positive definite, one has $p_k^T A p_k > 0$, from (6.3) it follows that:

$$r_{k+1}^T p_k = (r_k - a_k A p_k)^T p_k = r_k^T p_k - a_k p_k^T A p_k = r_k^T p_k - \frac{r_k^T r_k}{p_k^T A p_k} p_k^T A p_k \tag{6.4}$$

$$= r_k^T p_k - r_k^T r_k = 0$$

according to $(A_k)(N_2)$.

For $j < k$, $r_{k+1}^T p_j = r_k^T p_j - a_k p_k^T A p_j = 0$ because of $(A_k)(N_1)(N_3)$. So, $(A_{k+1})(N_1)$ is proved.

(N_2): We have $r_k^T r_k > 0$, since otherwise $r_k = 0$, and thus, in view of (6.3)

$$p_k = \begin{cases} r_0 & \text{if } k = 0; \\ b_{k-1} p_{k-1} & \text{if } k \neq 0. \end{cases} \tag{6.5}$$

Since $k < m$, we must have $k > 0$, otherwise $p_0 = r_0 = 0$ and $m = 0$. For $k > 0$, in view of $p_k \neq 0$ it follows from (6.5) and $(A_k)(N_3)$ the contradiction: $0 < p_k^T A p_k = b_{k-1} p_k^T A p_{k-1} = 0$. Then $r_k^T r_k > 0$, so that b_k and p_{k+1} are defined through (6.3). From (6.3) and (6.4), one deduces that:

$$r_{k+1}^T p_{k+1} = r_{k+1}^T (r_{k+1} + b_k p_k) = r_{k+1}^T r_{k+1}.$$

whence $(A_{k+1})(N_2)$.

(N_3) : From what was just proved, $r_k \neq 0$, so that a_j^{-1} is well defined for $j \le k$. From (6.3), we have for $j \le k$

$$\begin{aligned} p_{k+1}^T A p_j &= r_{k+1}^T A p_j + b_k p_k^T A p_j \\ &= a_j^{-1} r_{k+1}^T (r_j - r_{j+1}) + b_k p_k^T A p_j \\ &= a_j^{-1} r_{k+1}^T (p_j - b_{j-1} p_{j-1} - p_{j+1} + b_j p_j) + b_k p_k^T A p_j \\ &= -a_j^{-1} r_{k+1}^T p_{j+1} + b_k p_k^T A p_j \\ &= \begin{cases} 0 & \text{if } j < k \text{ because of } (A_k)(N_3) \text{ and } (A_k)(N_1); \\ 0 & \text{for } j = k \text{ by definition of } a_k,\ b_k, \text{ and } (A_k)(N_1)(N_2). \end{cases} \end{aligned}$$

NB: When $j = 0$, the vector p_{-1} has to be interpreted as the zero vector, i.e, $p_{-1} = 0$. This proves $(A_{k+1})(N_3)$.

(N_4) : From (6.3) and $(A_{k+1})(N_1)$, one has for $i \le k$ and $(p_{-1} = 0)$

$$r_i^T r_{k+1} = (p_i - b_{i-1} p_{i-1})^T r_{k+1} = p_i^T r_{k+1} - b_{i-1} p_{i-1}^T r_{k+1} = 0.$$

(N_5) : It follows from (6.3) and $(A_k)(N_5)$ that

$$b - A x_{k+1} = b - A(x_k + a_k p_k) = r_k - a_k A p_k = r_{k+1},$$

whence, (A_{k+1}) is checked in its entirely, and consequently, by induction, (A_m) holds true.

Because of $(A_m)(N_2)(N_4)$, we have $r_i \neq 0$ for all $i < m$, and these vectors form an orthogonal system in $\mathbb{R}^n$. Consequently, $m \leq n$. From $p_m = 0$, it finally follows, by virtue of $(A_m)(N_2)$, that $r_m^T r_m = r_m^T p_m = 0$, and thus $r_m = 0$, so that x_m is a solution of $Ax = b$. The proof of Theorem 6.2.1 is complete. □

Remark 6.2.2. *It follows from Theorem* 6.2.1 *that the cg-method is well defined since* $r_k^T r_k > 0$, $p_k^T A p_k > 0$ *for* $p_k \neq 0$. *Furthermore, property* (d) *states that the vectors* p_k *are* A*-conjugate.*

Proof. **of relation** (6.2).
According to (6.3) and for $k < m$, the vectors r_i and p_i, $i \leq k$ span the same subspace of $\mathbb{R}^n$, i.e,

$$S_k := \left\{ \sum_{j=0}^{k} u_j r_j,\ u_j \in \mathbb{R} \right\} = \left\{ \sum_{j=0}^{k} \rho_j p_j,\ \rho_j \in \mathbb{R} \right\}.$$

For the function $\Phi(\rho_0, \rho_1, ..., \rho_k) = h(x_k + \sum_{j=0}^{k} \rho_j p_j)$, we have

$$\frac{\partial \Phi}{\partial \rho_j}(\rho_0, \rho_1, ..., \rho_k) = -r^T p_j,$$

where $r = b - Ax$, $x := x_k + \sum_{j=0}^{k} \rho_j p_j$. Choosing

$$\rho_j := \begin{cases} a_k & \text{for } j = k; \\ 0 & \text{for } j < k. \end{cases}$$

We thus obtain, by (6.3), $x = x_{k+1}$, $r = r_{k+1}$ and by Theorem 6.2.1 (part (b)), $-r_{k+1}^T p_j = 0$, so that indeed

$$\min_{\rho_0, \rho_1, ..., \rho_k} \Phi(\rho_0, \rho_1, ..., \rho_k) = \min_{\rho_0, \rho_1, ..., \rho_k} h\left(x_k + \sum_{j=0}^{k} \rho_j p_j \right) = h(x_{k+1}).$$

Using the recursion (6.3) for the vectors r_k and p_k, it is readily verified that:

$$p_k \in span[r_0, Ar_0, ..., A^k r_0],$$

so that

$$S_k = span[p_0, p_1, ..., p_k] = span[r_0, Ar_0, ..., A^k r_0] = K_{k+1}(r_0, A)$$

is the $(k+1)$-st Krylov space of A belonging to the vector r_0. If one replaces $k+1$ by k and uses (6.2) and $h(x) = \frac{1}{2}\|x - \bar{x}\|_A^2$, one obtains

$$S_{k-1} = K_k(r_0, A),\ \ x_k - x_0 \in K_k(r_0, A) \text{ and}$$

$$\|x - \bar{x}\|_A = \min\{\|u - \bar{x}\|_A : u \in x_0 + K_k(r_0, A)\}. \tag{6.6}$$

□

Remark 6.2.3. *In exact arithmetic, we would have, at the latest* $r_n = 0$, *and thus in* x_n *the desired solution of* (6.1). *Because of the effects of rounding errors, the value computed for* r_n *is, as a rule, different from zero. In actual computation, the method is then simply continued beyond the value* $k = n$ *until an* r_k *(or* p_k*) is found which is sufficiently small. An algo program for a variant of this algorithm can be found in Wilkinson, Reinsch* (1971, [174]), *an extensive account of numerical experiments, in Reid* (1971, [111]) *and further results in Axelsson* (1976, [4]).

6.2.1 Estimation of the speed of the cg-method

When introducing the error $e_j = x_j - \bar{x}$ of x_j and by $r_0 = -Ae_0$, any $u \in x_0 + K_k(r_0, A)$ satisfies

$$u - \bar{x} \in e_0 + span[Ae_0, A^2e_0, ..., A^ke_0],$$

that is, there exists a polynomial $p(t) = 1 + \sum_{j=1}^{k} \alpha_j t^j$ with $u - \bar{x} = p(A)e_0$. Then

$$\|e_k\|_A = \min\{\|p(A)e_0\|_A : \ \ p \in \overline{\Pi}_k\},$$

where $\overline{\Pi}_k$ denotes the set of all real polynomials of degree$\leq k$ with $p(0) = 1$. Or, the positive definite matrix A has n eigenvalues: $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_n > 0$ and associated orthonormal eigenvectors z_i. Let us write e_0 in the form $e_0 = \sum_{j=1}^{n} \rho_j z_j$, then

$$\|e_0\|_A^2 = e_0^T A e_0 = \sum_{j=1}^{n} \lambda_j \rho_j^2,$$

$$\|p(A)e_0\|_A^2 = \sum_{j=1}^{n} \lambda_j p(\lambda_j)^2 \rho_j^2 \leq (\max_j p(\lambda_j)^2).\|e_0\|_A^2$$

and therefore

$$\frac{\|e_k\|_A}{\|e_0\|_A} \leq \min_{p \in \overline{\Pi}_k} \max_j |p(\lambda_j)| \leq \min_{p \in \overline{\Pi}_k} \max_{\lambda \in [\lambda_n, \lambda_1]} |p(\lambda)|. \tag{6.7}$$

Given infinite floating point precision, the number of iterations required to compute an exact solution is at most the number of distinct eigenvalues. (There is one other possibility for early termination: x_0 may already be A-orthogonal to some of the eigenvectors of A. If eigenvectors are missing from the expansion of x_0, their eigenvalues may be omitted from consideration in Relation (6.7). Be forewarned, however, that these eigenvectors may be reintroduced by floating point roundoff error).

If we know something about the characteristics of the eigenvalues of A, it is sometimes possible to suggest a polynomial that leads to a proof of a fast convergence. For the remainder of this analysis, however, we shall assume the most general case: the eigenvalues are evenly distributed between $\lambda_{\min}$ and $\lambda_{\max}$, the number of distinct eigenvalues is large, and floating point roundoff occurs.

6.2.2 Application of (6.7) to Chebyshev polynomials

A useful approach is to minimize equation (6.7) over the range $[\lambda_n, \lambda_1]$ rather than a finite number of points. The polynomials that accomplish this are based on Chebyshev polynomials. The Chebyshev polynomial of degree k is

$$T_k(x) = \cos(k \arccos x) := \cos(k\theta) \ \text{ if } x = \cos\theta, \ \ k = 0, 1, 2, ...$$

Remark 6.2.4. *The Chebyshev polynomials have the property that $|T_k(x)| \leq 1$ (in fact, they oscillate between 1 and -1) on the domain $x \in [-1, 1]$, and $|T_k(x)|$ is maximum on the domain $x \notin [-1, 1]$ among all such polynomial. Loosely speaking, $|T_k(x)|$ increases as quickly as possible outside the boxes in the illustration.*

We can construct a polynomial of $\overline{\Pi}_k := \{p : \text{polynomial of degree at most } k \,|\, p(0) = 1\}$ with small $\max\{|p(\lambda)| : \lambda \in [\lambda_n, \lambda_1]\}$ in the following way (in fact, we so obtain the optimal polynomial): consider the mapping

$$\begin{aligned} x(\cdot) : \ [\lambda_n, \lambda_1] &\to [-1, 1] \\ \lambda &\mapsto x(\lambda) = \frac{-2\lambda + (\lambda_1 + \lambda_n)}{\lambda_1 - \lambda_n}. \end{aligned}$$

According to Remark 6.2.4, the Chebyshev polynomial $\lambda \mapsto T_k(x(\lambda))$ is maximum outside on the domain $[\lambda_n, \lambda_1]$. In particular, it is maximum at $\lambda_0 = 0$. So the polynomial

$$p_k(\lambda) := \frac{T_k(x(\lambda))}{T_k(x(0))}$$

belongs to $\overline{\Pi}_k$ and satisfies

$$\max_{\lambda\in[\lambda_n,\lambda_1]} |p_k(\lambda)| = |T_k(x(0))|^{-1} \max_{\lambda\in[\lambda_n,\lambda_1]} |T_k(x(\lambda))| = |T_k(x(0))|^{-1} = \left| T_k\left(\frac{c+1}{c-1}\right)\right|^{-1}, \tag{6.8}$$

where $c = \lambda_1/\lambda_n$ is just the condition number of the matrix A with respect to the $lub_2(\cdot)$-norm. On the other side

$$\begin{aligned} T_k(x) = \cos(k \arccos x) &= \frac{1}{2}\left(e^{\hat{i}k \arccos x} + e^{-\hat{i}k \arccos x}\right) \\ &= \frac{1}{2}\left((e^{\hat{i} \arccos x})^k + (e^{\hat{i} \arccos x})^{-k}\right) \\ &= \frac{1}{2}\left((x + \hat{i}\sqrt{1-x^2})^k + (x + \hat{i}\sqrt{1-x^2})^{-k}\right) \end{aligned}$$

and for $x = x(0) = \frac{c+1}{c-1}$, one has

$$\begin{aligned} x + \hat{i}\sqrt{1-x^2} &= \frac{c+1}{c-1} + \hat{i}\sqrt{1 - \left(\frac{c+1}{c-1}\right)^2} \\ &= \frac{c+1}{c-1} + \hat{i}\sqrt{\frac{(c-1)^2 - (c+1)^2}{(c-1)^2}} \\ &= \frac{c+1}{c-1} + \hat{i}\sqrt{\frac{4\hat{i}^2 c}{(c-1)^2}} \\ &= \frac{c+1}{c-1} + 2\hat{i}^2\sqrt{\frac{c}{(c-1)^2}} \\ &= \frac{c - 2\sqrt{c} + 1}{c-1} = \frac{(\sqrt{c}-1)^2}{(\sqrt{c}-1)(\sqrt{c}+1)} = \frac{\sqrt{c}-1}{\sqrt{c}+1}. \end{aligned}$$

So, (6.8) becomes

$$\left| T_k\left(\frac{c+1}{c-1}\right)\right|^{-1} = 2\left(\left(\frac{\sqrt{c}-1}{\sqrt{c}+1}\right)^k + \left(\frac{\sqrt{c}-1}{\sqrt{c}+1}\right)^{-k}\right)^{-1} \le 2\left(\frac{\sqrt{c}-1}{\sqrt{c}+1}\right)^k,$$

we finally obtain the estimate

$$\frac{\|e_k\|_A}{\|e_0\|_A} \leq \left(T_k\left(\frac{c+1}{c-1}\right)\right)^{-1} \leq 2\left(\frac{\sqrt{c}-1}{\sqrt{c}+1}\right)^k. \tag{6.9}$$

In practice, CG usually converges faster than Equation (6.9) would suggest, because of good eigenvalue distributions or good starting points. However, it is not necessarily true that every iteration of CG enjoys faster convergence. The factor of 2 in Equation (6.9) allows CG a little slack for these poor iterations.

Thus, the speed of convergence of the conjugate gradient method is determined by $\sqrt{c} = \sqrt{\frac{\lambda_1}{\lambda_n}}$ and increases if the condition number c of A decreases. This behavior is exploited by the so-called preconditioning techniques in order to accelerate the conjugate gradient method. One tries to approximate as well as possible the positive definite matrix A by another positive definite matrix B, the preconditioner, so that $B^{-1}A$ is a good approximation of the unity matrix. The positive definite matrix

$$A' = B^{1/2}(B^{-1}A)B^{-1/2} = B^{-1/2}AB^{-1/2},$$

which is similar to $B^{-1}A$, has a much smaller condition number than A, $c' = cond(A') << c = cond(A)$ (since for any positive definite matrix B, there exists a positive definite matrix $C := B^{1/2}$ with $C^2 = B$). Moreover, the matrix B should be chosen such that linear system $Bq = r$ is easily solvable, which is the case, e.g., if B has a Choleski factor L. After having chosen B, the vector $\bar{x}' := B^{1/2}\bar{x}$ solves the system

$$A'x' = b';\ \ b' := B^{-1/2}b,$$

which is equivalent to $Ax = b$. We apply the cg-method (6.3) to solve the system $A'x' = b'$, using $x'_0 := B^{1/2}x_0$ as starting vector. Because of (6.13) and $c' << c$, the sequence $\{x'_k\}$ generated by the cg-method will converge very rapidly toward $\bar{x}'$. But instead of computing the matrix A' and the vectors x'_k explicitly, we generate the sequence $x_k := B^{-1/2}x'_k$ associated with directly as follows: using the transformation rules

$$A' = B^{-1/2}AB^{-1/2},\ \ b' = B^{-1/2}b$$

$$x'_k = B^{1/2}x_k,\ \ r'_k = b' - A'x'_k = B^{-1/2}r_k,\ \ p'_k = B^{1/2}p_k,$$

we obtain from the recursions of (6.3) for the system $A'x' = b'$ immediately the recursions of the following method.

(6.10) **Preconditioned conjugate gradient method**

Initialization: Choose $x_0 \in \mathbb{R}^n$, compute $r_0 := b - Ax_0$, $q_0 := B^{-1}r_0$ and put $p_0 := q_0$.

For $k = 0, 1, 2, ...$

(1) If $p_k = 0$, set $m := k$ and stop: x_k is the solution of $Ax = b$. Otherwise,

(2) compute:

$$a_k := \frac{r_k^T q_k}{p_k^T A p_k},\quad x_{k+1} := x_k + a_k p_k,$$

$$r_{k+1} := r_k - a_k A p_k,\quad q_{k+1} := B^{-1}r_{k+1},$$

$$b_k := \frac{r_{k+1}^T q_{k+1}}{r_k^T q_k},\quad p_{k+1} := q_{k+1} + b_k p_k.$$

Remark 6.2.5. *The only difference, compared to* (6.3)*, is that we have to solve at each step an additional linear system $Bq = r$ with the matrix B.*

Now, the problem arises of finding an appropriate preconditioning matrix B, a problem similar to the problem of finding a suitable iterative method. When solving the linear equations $Ax = b$ arising from the discretization of boundary value problems for elliptic equations, that is, the following model problem:

$$\begin{cases} -u_{xx} - u_{yy} = f & 0 < x, y < 1, \\ u(x,y) = 0 & \text{for } (x,y) \in \partial\Omega. \end{cases}$$

(for the unit square $\Omega := \{(x,y) | 0 < x, y < 1\} \subset \mathbb{R}^2$ with boundary $\partial\Omega$),
it turned out to be useful to choose B as the $SSOR$ matrix defined by

$$B := \frac{1}{2-w}\left(\frac{1}{w}D - E\right)\left(\frac{1}{w}D\right)^{-1}\left(\frac{1}{w}D - E^T\right) \tag{6.11}$$

with a suitable $w \in (0,2)$ [see Axelsson (1977, [4])]. Here, D and E are defined as in the standard decomposition of A, i.e., $A = D - E - E^T$.

Note that the factor $L = \frac{1}{w}D - E$ of B is a lower triangular matrix that is as sparse as the matrix A. Indeed, below the diagonal, it is nonzero at the same positions as A.

Another more general proposal is due to Meijerink and Van der Vorst (1977, [99]): They proposed to determine the preconditioner B and its choleski decomposition by the so-called `incomplete Choleski factorization` of A. Here, we consider the Choleski decompositions of the form $B = LDL^T$, where L is a lower triangular matrix with $l_{ii} = 1$ and D is a positive definite diagonal matrix. With the `incomplete Choleski decomposition` it is even possible to prescribe the sparsity structure of L: Given an arbitrary set $G \subset \{(i,j) | j \leq i \leq n\}$ of pairs of indices with $(i,i) \in G$ for all i, it is possible to find an L with

$$l_{i,j} \neq 0 \Rightarrow (i,j) \in G.$$

However, `incomplete Choleski factorization` gives a decent B approximation to A only for positive definite matrices A, which are also M-matrices, that is, matrices A with $a_{ij} \leq 0$ for $i \neq j$ and $A^{-1} \geq 0$ [see Meijerink and Van der Vorst [99]].

Fortunately, M-matrices occur very frequently in applications and there are simple sufficient criteria for A to be an M-matrix. For instance, any matrix $A = A^T$ with $a_{ii} > 0$, $a_{ij} \leq 0$ for $i \neq j$ that also satisfies the hypotheses of **Weak row sum criterion** (i.e., A is irreducible and

$$|a_{ii}| \geq \sum_{k \neq i} |a_{ik}| \quad \text{for all } i = 1, 2, ..., n$$

further, $|a_{i_0 i_0}| > \sum_{k \neq i_0} |a_{i_0 k}|$ for at least one i_0) is an M-matrix (e.g: the matrix A of the model problem). This is shown, by establishing the convergence of the Neumann series

$$A^{-1} = (I + J + J^2 + ...)D^{-1} \geq 0$$

for $A = D(I - J)$, with $\|J\|_2 < 1$.

Given an index set G as earlier, the incomplete Choleski factorization of a positive definite M-matrix A produces the factors D and L of a positive definite matrix $B = LDL^T$ approximating A according to the following rules:

(6.12) **Incomplete choleski factorization**

For $i = 1, 2, ..., n$:

$$d_i := a_{ii} - \sum_{k=1}^{i-1} d_k l_{ik}^2$$

For $j = i+1, ..., n$:

$$d_i l_{ji} := \begin{cases} a_{ji} - \sum_{k=1}^{i-1} d_k l_{jk} l_{ik} & \text{if } (i,j) \in G \\ 0 & \text{otherwise.} \end{cases}$$

That is, the only difference, compared to the ordinary Choleski algorithm, is that $l_{ij} = 0$ is set equal to zero at the "forbidden" places $(i,j) \in G$.

6.2.3 Application to the least-squares problems

The cg-method can also be used to solve the least-squares problems for over determined systems

$$\min_x \|Bx - c\|_2, \tag{6.13}$$

where B is a sparse $m \times n$ matrix with $m \geq n$ and $rank(B) = n$. According to (5.1), the optimal solution $\bar{x}$ of (6.13) is also solution of the normal equations

$$Ax = b, \;\; A = B^T B, \;\; b := B^T c$$

where A is a positive definite matrix. Even if B is sparse, the matrix $A = B^T B$ can be dense. This suggests the following variant of the conjugate-gradient method (6.3) for solution of (6.13), and has proved useful in practice

(6.14) **Algorithm**

Initialization: Choose $x_0 \in \mathbb{R}^n$ and compute $s_0 := c - Bx_0$, $p_0 := r_0 := B^T s_0$.

For $k = 0, 1, 2, ...$

(1) If $p_k = 0$, stop: x_k is the optimal solution of (6.13). Otherwise,

(2) compute:

$$q_k := Bp_k, \;\; a_k := \frac{r_k^T r_k}{q_k^T q_k}, \;\; x_{k+1} := x_k + a_k p_k,$$

$$s_{k+1} := s_k - a_k q_k, \;\; r_{k+1} := B^T s_{k+1},$$

$$b_k := \frac{r_{k+1}^T r_{k+1}}{r_k^T r_k}, \;\; p_{k+1} := r_{k+1} + b_k p_k.$$

Clearly, the iterates $\{x_k\}_k$ generated by this method satisfy

$$x_k - x_0 \in K_k(r_0, B^T B),$$

$$\|x_k - \bar{x}\|_{B^T B} = \min\{\|u - \bar{x}\|_{B^T B} : u \in x_0 + K_k(r_0, B^T B)\}$$

Remark 6.2.6. *Since the $cond_2(B^T B) = cond_2(B)^2$, the convergence speed of this variant of the Conjugate gradient method is slow if the condition number $cond_2(B) >> 1$ of B is large.*

Conclusion: For a quadratic matrix, this algorithm could also be used to solve the linear equation $Bx = c$ even for a nonsymmetric matrix B. However, it is usually better to apply one of the algorithms (GMRES, QMR, Bi-CGSTAB) which will be described in the following for solving such linear equations.

6.3 Generalized minimum residual (GMRES) algorithm

The generalized minimum residual (GMRES) method [Saad, Schulz (1986, [116]), Saad (1996, [114])] is more expensive but it is defined also for general linear systems with a nonsymmetric nonsingular matrix A. It generates vectors $x_k \in x_0 + K_k(r_0, A)$ with a minimal residual $b - Ax_k$,

$$\|b - Ax_k\|_2 = \min_{x \in x_0 + K_k(r_0, A)} \|b - Ax\|_2,$$

and uses, as main tool, orthonormal vectors $v_1, v_2, ...$, that provide an orthonormal basis for the Krylov spaces $K_k(r_0, A)$ of dimension k,

$$span[v_1, v_2, ..., v_k] = K_k(r_0, A).$$

These vectors are given by the method of Arnoldi (6.17), (1951).

Let us consider the system of linear equations

$$Ax = b$$

with a general real nonsingular $n \times n$ matrix A which may be nonsymmetric, and solution $\bar{x} := A^{-1}b$. There were many efforts to develop conjugate-gradient type algorithms for solving such systems [see Saad (1996, [114])] for a comprehensive representation] that, among others, lead to the generalized minimum residual (GMRES) method of Saad and Schult (1986, [116]). It is a Krylov space method: starting with any approximate solution $x_0 \neq \bar{x}$ with residual $r_0 := b - Ax_0 \neq 0$, it generates subsequence approximations x_k to x_0 with the following properties:

$$\text{(6.15)} \qquad \begin{cases} x_k \in x_0 + K_k(r_0, A) \\ \|x_k - \bar{x}\|_2 = \min\{\|u - \bar{x}\|_2 : \ u \in x_0 + K_k(r_0, A)\}. \end{cases}$$

As a tool, we use orthonormal bases of the Krylov spaces $K_k(r_0, A)$, $k \geq 1$. In view of the definition

$$K_k(r_0, A) = span[r_0, Ar_0, A^2 r_0, ..., A^{k-1} r_0],$$

and since $r_0 \neq 0$, then

$$1 \leq dim\, K_k(r_0, A) \leq k, \quad k \geq 1.$$

Thus, there is a largest index m with $1 \leq m \leq n$ so that

$$dim\, K_k(r_0, A) = k, \ \text{ for } 1 \leq k \leq m.$$

The positive integer number m is the smallest integer for which the Krylov space $K_m(r_0, A)$ is A-invariant, that is,

$$\text{(6.16)} \qquad AK_m(r_0, A) := \{Ax : \ x \in K_m(r_0, A)\} \subset K_m(r_0, A).$$

Indeed: The A-invariance of $K_m(r_0, A)$ is equivalent to

$$A^m r_0 \in K_m(r_0, A),$$

so that

$$dim\, K_{m+1}(r_0, A) = dim\, K_m(r_0, A) < m + 1.$$

Remark 6.3.1. *A main feature of the GMRES method is that it uses orthonormal vectors $v_i \in \mathbb{R}^n$ to span all the Krylov spaces $K_k(r_0, A)$, $k \leq m$.*

$$span[v_1, v_2, ..., v_k] = K_k(r_0, A), \ \textit{for } 1 \leq k \leq m.$$

The vector v_1 is determined, up to the sign:

$$v_1 := \frac{r_0}{\beta}, \quad \beta := \|r_0\|_2$$

and the remaining vectors v_j $(j = 2, 3, ..., k)$ is computed by exploiting the Arnoldi algorithm (1951, [3]*); it generates the algorithm of Lanczos to nonsymmetric matrices A.*

(6.17) **Arnoldi's orthonormalization method**

Initialization: Given $r_0 \neq 0$, put $\beta := \|r_0\|_2$, $v_1 := r_0/\beta$.

For $k = 0, 1, 2, ...$

(1) Compute $u := Av_k$.

(2) For $i = 1, 2, ..., k$
Compute $h_{ik} := v_i^T u$.

(3) Compute $w_k := u - \sum_{i=1}^{k} h_{ik} v_i$ and $h_{k+1,k} := \|w_k\|_2$.

(4) If $h_{k+1,k} = 0$, set $m := k$ and stop. Otherwise,

(5) Compute $v_{k+1} := w_k / h_{k+1,k}$.

In the algorithm (6.17), the h_{ik} are determined in step (2) such that

$$w_k \perp v_i, \quad \text{for} \quad i = 1, 2, ..., k.$$

Therefore, if $\|w_k\|_2 \neq 0$, step (5) determines a new vector v_{k+1} such that $v_1, v_2, ..., v_{k+1}$ form an enlarged orthonormal system of $k+1$ vectors. It follows by induction that Arnoldi's method generates vectors $v_1, v_2, ..., v_m$ such that

$$span[v_1, v_2, ..., v_k] = K_k(r_0, A) = span[r_0, Ar_0, ..., A^{k-1} r_0] \tag{6.18}$$

for $k \leq m$. It follows from (6.18) the following Lemma

Lemma 6.3.1. *Each v_k, $1 \leq k \leq m$, generated by the method* (6.17) *can be represented in the form*

$$v_k = \sum_{j=1}^{k} \gamma_j A^{j-1} r_0 \quad \textit{with} \quad \gamma_k \neq 0. \tag{6.19}$$

Proof. This is true for $k = 1$ by definition of $v_1 = r_0/\beta$. If (6.19) holds for some $k \geq 1$, then by the induction hypothesis, $v_i \in K_k(r_0, A)$ for $i \leq k$ and (6.19) implies

$$Av_k = \sum_{j=1}^{k} \gamma_j A^j r_0, \quad \gamma_k \neq 0$$

and

$$w_k = \sum_{j=1}^{k} \gamma_j A^j r_0 - \sum_{j=1}^{k} h_{jk} v_j.$$

By the induction hypothesis, each v_j, $1 \leq j \leq k$, has a representation

$$v_j = \sum_{i=1}^{j} \delta_i A^{i-1} r_0.$$

Thus, w_k has the form

$$w_k = \sum_{j=1}^{k} \epsilon_j A^j r_0 \quad \text{with} \quad \epsilon_k = \gamma_k \neq 0.$$

Hence, if $h_{k+1,k} = \|w_k\|_2 \neq 0$, assertion (6.19) for $k+1$ follows, because of $v_{k+1} = w_k/h_{k+1,k}$.

If $w_k = 0$ then $A^k r_0$ is a linear combination of the vectors $A^j r_0$, $j \leq k-1$, that is the break-off index $m = k$ of the Arnoldi's method is the same index introduced earlier:

$$m = \max\{k \geq 1 : \ dim\, K_k(r_0, A) = k\}.$$

□

Setting $V_k = [v_1, v_2, ..., v_k]$, $k \leq m$, it follows from (6.18) that

$$K_k(r_0, A) = \{V_k y : \ y \in \mathbb{R}^k\}. \tag{6.20}$$

So, for all $x \in x_0 + K_k(r_0, A)$, there is a unique $y \in \mathbb{R}^k$ such that

$$x = x_0 + V_k y. \tag{6.21}$$

The Arnoldi's recursions may be formulated compactly in terms of the $(k+1) \times k$ Hessenberg matrices

$$\bar{H}_k := \begin{bmatrix} h_{11} & h_{12} & \dots & h_{1k} \\ h_{21} & h_{22} & & h_{2k} \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & h_{kk} \\ 0 & \dots & 0 & h_{k+1,k} \end{bmatrix}, \quad 1 \leq k \leq m,$$

and the $k \times k$ submatrices H_k obtained by deleting the last row of $\bar{H}$. The formulae of steps (3) and (5) of algorithm (6.17) imply for all $1 \leq k < m$

$$Av_k = \sum_{j=1}^{k+1} h_{jk} v_j = \sum_{j=1}^{k} h_{jk} v_j + w_k \tag{6.22}$$

and

$$Av_m = \sum_{j=1}^{m} h_{jm} v_j, \quad \text{since } w_m = 0. \tag{6.23}$$

The relations (6.22) − (6.23) are equivalent to

$$AV_m = V_m H_m \tag{6.24}$$

and for $1 \leq k < m$, to

$$AV_k = V_k H_k + w_k e_k^T = V_{k+1} \bar{H}_k, \quad e_k := (0, ..., 0, 1)^T \in \mathbb{R}^k. \tag{6.25}$$

For $1 \leq k \leq m$ and because $V_k^T w_k = 0$, relations (6.24) − (6.25) imply:

$$H_k = V_k^T A V_k. \tag{6.26}$$

Remark 6.3.2. *The matrix H_m is nonsingular, and rank $\bar{H}_k = k$ for $k < m$.*

Indeed: Looking the contradiction. If Remark 6.3.2 does not hold, there exists a vector $y \in \mathbb{R}^m$, $y \neq 0$, with $H_m y = 0$. Or $z := V_m y \neq 0$, and relation (6.24) implies

$$Az = AV_m y = V_m H_m y = 0$$

contradiction with the fact that A is nonsingular. Also, the subdiagonal elements $h_{j+1,j}$ $j = 1, 2, ..., k$ of $\bar{H}_k$, $k < m$, are non zero: this establishes $rank\, \bar{H}_k = k$.

On the other side, the matrices $\bar{H}_k$, H_k and V_k allow a straightforward determination of the solution x_k of (6.15). Since $r_0 = \beta v_1 = \beta V_{k+1} \bar{e}_1$, $\bar{e}_1 = (1, 0, ..., 0)^T \in \mathbb{R}^{k+1}$, $V_{k+1}^T V_{k+1} = I$, and (6.21)-(6.25) for $k < m$, it follows for all $x \in x_0 + K_k(r_0, A)$ that

$$\begin{aligned} \|b - Ax\|_2 &= \|b - Ax_0 - AV_k y\|_2 \\ &= \|r_0 - V_{k+1} \bar{H}_k y\|_2 \\ &= \|V_{k+1}(\beta \bar{e}_1 - \bar{H}_k y)\|_2 \\ &= \|\beta \bar{e}_1 - \bar{H}_k y\|_2. \end{aligned}$$

So, the solution y_k of the least-squares problem

$$\min_{y \in \mathbb{R}^k} \|\beta \bar{e}_1 - \bar{H}_k y\|_2 \tag{6.27}$$

thus provides the solution x_k of (6.15),

$$x_k = x_0 + V_k y_k.$$

In the case $k = m$, Relation (6.24) implies for $x \in x_0 + K_m(r_0, A)$ that

$$\|b - Ax\|_2 = \|r_0 - AV_m y\|_2 = \|V_m(\beta e_1 - H_m y)\|_2 = \|\beta e_1 - H_m y\|_2. \tag{6.28}$$

where the vector $e_1 = (1, 0, ..., 0)^T \in \mathbb{R}^m$.

Lemma 6.3.2. *(Characterization of the break-off index m of (6.17) in terms of x_k : solution of (6.15))*
The vector x_m solves the linear equations $Ax = b$, that is, $x_m = A^{-1}b$, and $x_k \neq A^{-1}b$ for all $k < m$. The break-off index m is the first index for which x_k solves the linear system $Ax = b$.

Proof. From Remark 6.3.2, the matrix $H_m \in \mathbb{R}^{m \times m}$ is nonsingular, then there is a unique $y \in \mathbb{R}^m$ with $H_m y_m = \beta e_1$. According to (6.28), the corresponding $x_m := x_0 + V_m y_m$ solves the linear system $Ax = b$.

For $k < m$, all the subdiagonal elements $h_{j+1,j}$, $j = 1, 2, ..., k$, of the Hessenberg matrix $\bar{H}_k$ are nonzero. The linear equations

$$\bar{H}_k y = \begin{bmatrix} h_{11} & h_{12} & \dots & h_{1k} \\ h_{21} & h_{22} & & h_{2k} \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & h_{kk} \\ 0 & \dots & 0 & h_{k+1,k} \end{bmatrix} y = \begin{bmatrix} \beta \\ 0 \\ \vdots \\ \vdots \\ 0 \end{bmatrix} = \beta \bar{e}_1$$

are not solvable. Indeed, the unique solution of k last equations is $y = 0$, since $h_{j+1,j} \neq 0$, but $y = 0$ does not solve the first equation as $\beta \neq 0$. □

Determination of an approximate solution of (6.15)

The least-squares problems (6.27) may be solved by the orthogonalization methods described in section 5.1.2, taking advantage of the Hessenberg structure of matrices $\bar{H}_k$. The idea is to consider the $(k+1)\times(k+1)$ given rotations $\Omega_j = \Omega_{j,j+1}$ for $j = 1, 2, ..., k$ of the type:

$$\Omega_{j,j+1} = \begin{bmatrix} 1 & & & & & & & \\ & \ddots & & & & & & \\ & & 1 & & & & & \\ & & & c_j & -s_j & & & \\ & & & s_j & c_j & & & \\ & & & & & 1 & & \\ & & & & & & \ddots & \\ & & & & & & & 1 \end{bmatrix}, \quad c_j^2 + s_j^2 = 1.$$

The parameters c_j, s_j are chosen such that for each matrix in the sequence:

$$\bar{H}_k \to \Omega_1 \bar{H}_k \to \Omega_2(\Omega_1 \bar{H}_k) \to \cdots \to \Omega_k(\Omega_{k-1}\Omega_{k-2}\cdots\Omega_1 \bar{H}_k) =: \bar{R}_k,$$

the first nonzero subdiagonal element is annihilated so that the sequence terminates with a "upper triangular" $(k+1)\times k$ matrix

$$\bar{R}_k = \begin{bmatrix} R_k \\ 0 \end{bmatrix}, \quad R_k = \begin{bmatrix} x & x & \dots & x \\ 0 & x & & \vdots \\ \vdots & \ddots & \ddots & x \\ 0 & \dots & 0 & x \end{bmatrix}.$$

The concurrent transformations of the vector $\bar{g}_0 := \beta \bar{e}_1 \in \mathbb{R}^{k+1}$,

$$\bar{g}_0 \to \Omega_1 \bar{g}_0 \to \Omega_2(\Omega_1 \bar{g}_0) \to \cdots \to \Omega_k(\Omega_{k-1}\Omega_{k-2}\cdots\Omega_1 \bar{g}_0) =: \bar{g}_k,$$

lead to the vector

$$\bar{g}_k = \begin{bmatrix} g_k \\ \bar{\gamma}_{k+1} \end{bmatrix}, \quad g_k = \begin{bmatrix} \gamma_1 \\ \gamma_2 \\ \vdots \\ \gamma_k \end{bmatrix} \in \mathbb{R}^k.$$

Remark 6.3.3. *The notation of its components is to indicate, that of the first k components $\gamma_1, \gamma_2, ..., \gamma_k$ of $\bar{g}_k$ will no longer change in the subsequent steps $k \to k+1 \to \cdots$, of the algorithm.*

Example 6.3.1. *Case where $k = 4$.*
$B \overset{\Omega}{\to} C$ stands for a left multiplication with the matrix Ω, that is, $C := \Omega B$. " $\star$ " denotes elements that have changed during the preceding transformation:

$$[\bar{H}_4|\bar{g}_0] = \begin{bmatrix} x & x & x & x & : x \\ x & x & x & x & : 0 \\ 0 & x & x & x & : 0 \\ 0 & 0 & x & x & : 0 \\ 0 & 0 & 0 & x & : 0 \end{bmatrix} \overset{\Omega_1}{\to} \begin{bmatrix} \star & \star & \star & \star & : \star \\ 0 & \star & \star & \star & : \star \\ 0 & x & x & x & : 0 \\ 0 & 0 & x & x & : 0 \\ 0 & 0 & 0 & x & : 0 \end{bmatrix} \overset{\Omega_2}{\to} \begin{bmatrix} x & x & x & x & : x \\ 0 & \star & \star & \star & : \star \\ 0 & 0 & \star & \star & : \star \\ 0 & 0 & x & x & : 0 \\ 0 & 0 & 0 & x & : 0 \end{bmatrix}$$

$$\overset{\Omega_3}{\to} \begin{bmatrix} x & x & x & x & : x \\ 0 & x & x & x & : x \\ 0 & 0 & \star & \star & : \star \\ 0 & 0 & 0 & \star & : \star \\ 0 & 0 & 0 & x & : 0 \end{bmatrix} \overset{\Omega_4}{\to} \begin{bmatrix} x & x & x & x & : x \\ 0 & x & x & x & : x \\ 0 & 0 & x & x & : x \\ 0 & 0 & 0 & \star & : \star \\ 0 & 0 & 0 & 0 & : \star \end{bmatrix} = \begin{bmatrix} R_k & g_4 \\ 0 & \bar{\gamma}_5 \end{bmatrix} = [\bar{R}_4|\bar{g}_4].$$

When setting $Q_k := \Omega_k \Omega_{k-1} \cdots \Omega_1$, then Q_k is a unitary matrix, so

$$\|\beta \bar{e}_1 - \bar{H}_k y\|_2 = \|Q_k(\beta \bar{e}_1 - \bar{H}_k y)\|_2 = \|\bar{g}_k - \bar{R}_k y\|_2.$$

Then the solution y_k of the least squares problem (6.27) is obtained as the solution of

$$\min_{y\in\mathbb{R}^k}\|\bar{g}_k-\bar{R}_k y\|_2=\min_{y\in\mathbb{R}^k}\left\|\begin{pmatrix}g_k\\ \bar{\gamma}_{k+1}\end{pmatrix}-\begin{bmatrix}R_k\\ 0\end{bmatrix}y\right\|_2,$$

that is, as the solution $y_k := R_k^{-1}g_k$ of the linear equations

$$g_k = R_k y. \tag{6.29}$$

(R_k is nonsingular, since for $k < m$, $rank\,\bar{H}_k = k$ implies $rank\,\bar{R}_k = rank(Q_k\bar{H}_k) = k$). Then

$$x_k = x_0 + V_k y_k := x_0 + V_k R_k^{-1} g_k$$

is the solution of (6.15).

NB: The size of the residual "$b - Ax_k$" is given by

$$\|b-Ax_k\|_2=\|\beta\bar{e}_1-\bar{H}_k y_k\|_2=\|\bar{g}_k-\bar{R}_k y_k\|_2=|\bar{\gamma}_{k+1}|. \tag{6.30}$$

Now, we can save a major portion of the previous computations as we step from $k-1$ to k. The reason is that the $(k+1)\times k$ Hessenberg matrix $\bar{H}_k$ differs from the $k\times(k-1)$ matrix $\bar{H}_{k-1}$ essentially only by a computational column,

$$\bar{H}_k=\begin{bmatrix}h_{11} & \dots & h_{1,k-1} & h_{1k}\\ h_{21} & \ddots & \vdots & \vdots\\ 0 & \ddots & h_{k-1,k-1} & \vdots\\ \vdots & \ddots & h_{k,k-1} & h_{k,k}\\ 0 & \dots & 0 & h_{k+1,k}\end{bmatrix}=\begin{bmatrix}\bar{H}_{k-1} & h_k\\ 0 & h_{k+1,k}\end{bmatrix}$$

namely by the last column

$$\bar{h}_k := \begin{bmatrix}h_{1k}\\ \vdots\\ h_{kk}\\ h_{k+1,k}\end{bmatrix}=\begin{bmatrix}h_k\\ h_{k+1,k}\end{bmatrix},$$

the components which are computed in steps (2) and (3) of algorithm (6.17). This can be used for the matrix $Q_{k-1}\bar{H}_k$, (where $Q_{k-1}=\Omega_{k-1}\Omega_{k-2}\cdots\Omega_1$) which has the form

$$Q_{k-1}\bar{H}_k=\begin{bmatrix}R_{k-1} & r_k\\ 0 & \rho\\ 0 & \sigma\end{bmatrix}:=\tilde{R}_k \quad\text{with}\quad \begin{bmatrix}r_k\\ \rho\\ \sigma\end{bmatrix}:=Q_{k-1}\bar{h}_k,\quad r_k\in\mathbb{R}^{k-1}.$$

Therefore, we have to compute only the last column $\tilde{r}_k$ of $\tilde{R}_k$, which amounts to forming the product

$$\tilde{r}_k := \begin{bmatrix}r_k\\ \rho\\ \sigma\end{bmatrix}=Q_{k-1}\bar{h}_k=\Omega_{k-1}\Omega_{k-2}\cdots\Omega_1\bar{h}_k. \tag{6.31}$$

Whence, $\tilde{R}_k$ is transformed by a given rotation Ω_k of the type $\Omega_{k,k+1}$ (with parameters c_k, s_k) to upper triangular form:

$$\tilde{R}_k=\begin{bmatrix}R_{k-1} & r_k\\ 0 & \rho\\ 0 & \sigma\end{bmatrix}\to\Omega_k\tilde{R}_k=:\bar{R}_k=\begin{bmatrix}R_k\\ 0\end{bmatrix}=\begin{bmatrix}R_{k-1} & r_k\\ 0 & r_{k,k}\\ 0 & 0\end{bmatrix},$$

where the quantities c_k and s_k are given by

$$c_k := \frac{\rho}{\sqrt{\rho^2+\sigma^2}}\quad\text{and}\quad s_k=-\frac{\sigma}{\sqrt{\sigma^2+\sigma^2}}. \tag{6.32}$$

The last column $\bar{r}_k$ of $\bar{R}_k$, finally is obtained as follows:

$$\bar{r}_k = \begin{bmatrix} r_k \\ r_{kk} \\ 0 \end{bmatrix} = \begin{bmatrix} r_{1k} \\ \vdots \\ r_{kk} \\ 0 \end{bmatrix} = \Omega_k \tilde{r}_k, \;\; r_{kk} = \sqrt{\rho^2 + \sigma^2}. \tag{6.33}$$

Example 6.3.2. *Case* $k = 5$.

$$\tilde{R}_4 = \begin{bmatrix} x & x & x & x & x \\ & x & x & x & x \\ & & x & x & x \\ & & & x & x \\ & & & & \rho \\ & & & & \sigma \end{bmatrix} \overset{\Omega_5}{\to} \begin{bmatrix} x & x & x & x & x \\ & x & x & x & x \\ & & x & x & x \\ & & & x & x \\ & & & & \star \\ & & & & 0 \end{bmatrix} := \bar{R}_5.$$

Since $\bar{e}_1 \in \mathbb{R}^{k+1}$ is the first axis vector in $\mathbb{R}^{k+1}$, setting $\bar{g}_0 := \beta \bar{e}_1$ one has:

$$Q_{k-1}\bar{g}_0 = \begin{bmatrix} \bar{g}_{k-1} \\ 0 \end{bmatrix}, \quad \text{with } \; \bar{g}_{k-1} = \begin{bmatrix} \gamma_1 \\ \vdots \\ \gamma_{k-1} \\ \bar{\gamma}_k \end{bmatrix} = \begin{bmatrix} g_{k-1} \\ \bar{\gamma}_k \end{bmatrix},$$

as $\bar{g}_{k-1} \in \mathbb{R}^k$. Hence $\bar{g}_k := \Omega_k \Omega_{k-1} ... \Omega_1 \bar{g}_0$ satisfies

$$\bar{g}_k = \begin{bmatrix} \gamma_1 \\ \vdots \\ \gamma_k \\ \bar{\gamma}_{k+1} \end{bmatrix} := \Omega_k \begin{bmatrix} \bar{g}_{k-1} \\ 0 \end{bmatrix} = \Omega_k \begin{bmatrix} \gamma_1 \\ \vdots \\ \gamma_{k-1} \\ \bar{\gamma}_k \\ 0 \end{bmatrix},$$

that is,

$$\gamma_k = c_k \bar{\gamma}_k, \quad \bar{\gamma}_{k+1} = s_k \bar{\gamma}_k. \tag{6.34}$$

Thus, because of (6.30) – (6.34), the size of the residual $b - Ax_k$ can be computed recursively

$$\|b - Ax_k\|_2 = |\bar{\gamma}_{k+1}| = |s_k \bar{\gamma}_k|,$$

so that

$$\|b - Ax_k\|_2 = |\bar{\gamma}_{k+1}| = |s_k s_{k-1} ... s_1| \beta.$$

Remark 6.3.4. *In general, we need not solve the linear equations* $R_k y = g_k$ *for* y_k *and compute* $x_k = x_0 + V_k y_k$ *in order to find* $\|b - Ax_k\|_2$. *This can be used in the GMRES algorithm, if for a desired accuracy* $\epsilon > 0$, *the solution* x_k *of* (6.15) *is computed only when* $|\bar{\gamma}_{k+1}| = |\prod_{j=1}^{k} s_j| \beta \leq \epsilon$ *and not before.*

Lemma 6.3.3. *Let* $\epsilon > 0$ *be a given positive number such that* $|\bar{\gamma}_{k+1}| = |\prod_{j=1}^{k} s_j| \beta \leq \epsilon$. *Then the vectors* x_k *can be computed recursively.*

Proof. Introducing the matrices

$$P_k := V_k R_k^{-1} = [p_1, p_2, ..., p_k] \quad \text{with column } \; p_i.$$

Then

$$x_k = x_0 + V_k y_k = x_0 + V_k R_k^{-1} g_k =: x_0 + P_k g_k.$$

The equality

$$R_k = \begin{bmatrix} R_{k-1} & r_k \\ 0 & r_{kk} \end{bmatrix},$$

shows that P_k satisfies

$$V_k = [v_1, v_2, ..., v_k] = P_k R_k = [P_{k-1}, p_k] \begin{bmatrix} R_{k-1} & r_k \\ 0 & r_{kk} \end{bmatrix},$$

if and only if $v_k = P_{k-1}r_k + r_{kk}p_k$. Thus, the vectors p_k can be computed recursively by:

$$p_k = \frac{1}{r_{kk}} \left(v_k - \sum_{j=1}^{k-1} r_{jk} p_j \right). \tag{6.35}$$

This gives, because of

$$P_k g_k = [P_{k-1}, p_k] \begin{bmatrix} g_{k-1} \\ \gamma_k \end{bmatrix} = P_{k-1} g_{k-1} + \gamma_k p_k$$

the following recursion for the vectors x_k :

$$x_k = x_0 + P_k g_k = x_{k-1} + \gamma_k p_k. \tag{6.36}$$

where

$$\gamma_k = \beta c_k \prod_{j=1}^{k-1} s_j$$

□

Remark 6.3.5. *According to* (6.36)*, a storage of the matrices R_k is not necessary. However, this saving is more than offset since one has to compute the vectors p_k by* (6.35)*, which becomes, with increasing k, progressively expensive. So, in general, the use of relation* (6.36) *is not recommendable. On the other side, this balance changes if the matrices $\bar{H}_k$ are band matrices of small bandwidth l, $l << n$.*

Reorthogonalization technique

A weakness of the Arnoldi's method (6.17) is that, due to roundoff errors, the computed vectors v_i, $i \leq k$ become less and less orthogonal as k increases [this is a known defect of Gram-Schmidt orthogonalization]. An expensive remedy is to use reorthogonalization technique: A new computed vector $\tilde{v}_{k+1}$ is orthogonalized against all already accepted vectors $v_1, v_2, ..., v_k$ as follows

$$\tilde{v}_{k+1} \to \hat{v}_{k+1} := \tilde{v}_{k+1} - \sum_{j=1}^{k} (v_j^T \tilde{v}_{k+1}) v_j,$$

before accepting $v_{k+1} := \hat{v}_{k+1} / \|\hat{v}_{k+1}\|_2$ as next vector. Here, the computing effort is double , but an improvement, at no extra expenses, of the orthogonality of the computed v_i is already obtained if one replaces steps (1) to (3) of the Arnoldi's method by

($1'$) Compute $w := Av_k$.

($2'$) For $i := 1, 2, .., k$:
Compute $h_{ik} := v_i^T w$, $w := w - h_{ik} v_i$.

($3'$) Compute $h_{k+1,k} := \|w\|_2$ and set $w_k := w$.

Remark 6.3.6. *In contrast to the Conjugate-gradient algorithm* (6.3)*, a more serious disadvantage of the GMRES method is that the computational expenses of step $k-1 \to k$ increase proportionally to k, since each vector Av_k has to be orthogonalized with respect to all previous vectors $v_1, v_2, ..., v_k$ in order to find v_{k+1}.*

A drastic remedy is to restart the GMRES method periodically, say every N-th step , where $1 < N << n$ (e.g., $N = 10$), according to the following scheme, which is denoted by GMRES(N):

Scheme (6.37)

(0) Given x_0, compute $r_0 := b - Ax_0$.

(1) Compute x_N by means of the $GMRES$ method (more precisely, by relation (6.36)).

(2) Set $x_0 := x_N$ and go to step (0).

NB: A drawback of this approach is that, after each restart, one loses all information contained in the vectors $v_1, v_2, ..., v_N$.

Incomplete Quasi-minimal GMRES (QGMRES) method.

Instead of using restarts, one could artificially limit the number of orthogonalizations in step (2) of the Arnoldi's method (6.17) : one could fix an integer l with $1 \leq l << n$ and orthogonalize the vector Av_k only against the last l vectors $v_k, v_{k-1}, ..., v_{k-l+1}$. Then, only the information contained in the old vectors v_{k-i} with $i \geq l$ is not used anymore. This leads to an incomplete GMRES method , where one replaces steps (1) to (3) in (6.17) by

$(1')$ Compute $w := Av_k$.

$(2')$ For $i := \max\{1, k-l+1\}, .., k$:
Compute $h_{ik} := v_i^T w$, $w := w - h_{ik} v_i$.

$(3')$ Compute $h_{k+1,k} := \|w\|_2$ and set $w_k := w$.

This method generates $(k+1) \times k$ Hessenberg matrices $\bar{H}_k$ that are band matrices of bandwidth l, and $k \times k$ upper triangular band matrices R_k of bandwidth $l+1$.

Example 6.3.3. *$k = 8$ and $l = 3$.*

$$\bar{H}_8 = \begin{bmatrix} x & x & x & 0 & 0 & 0 & 0 & 0 \\ x & x & x & x & 0 & 0 & 0 & 0 \\ 0 & x & x & x & x & 0 & 0 & 0 \\ 0 & 0 & x & x & x & x & 0 & 0 \\ 0 & 0 & 0 & x & x & x & x & 0 \\ 0 & 0 & 0 & 0 & x & x & x & x \\ 0 & 0 & 0 & 0 & 0 & x & x & x \\ 0 & 0 & 0 & 0 & 0 & 0 & x & x \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & x \end{bmatrix}, \quad R_8 = \begin{bmatrix} x & x & x & x & 0 & 0 & 0 & 0 \\ 0 & x & x & x & x & 0 & 0 & 0 \\ 0 & 0 & x & x & x & x & 0 & 0 \\ 0 & 0 & 0 & x & x & x & x & 0 \\ 0 & 0 & 0 & 0 & x & x & x & x \\ 0 & 0 & 0 & 0 & 0 & x & x & x \\ 0 & 0 & 0 & 0 & 0 & 0 & x & x \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & x \end{bmatrix}$$

The relations (6.25) are still valid, but the columns of the matrices $V_k = [v_1, v_2, ..., v_k]$ are no longer orthogonal. Yet, the vectors $v_1, v_2, ..., v_k$ remain linearly independent for $k \leq m$, and they form a basis of $K_k(r_0, A)$ [cf. (6.18), the induction proof given there remains valid] so that each $x \in x_0 + K_k(r_0, A)$ has the form $x = x_0 + V_k y$ with a unique y. But because of the lacking orthogonality of the v_i, we have for $k < m$

$$\begin{aligned} \|b - Ax\|_2 &= \|b - Ax_0 - AV_k y\|_2 \\ &= \|r_0 - V_{k+1}\bar{H}_k y\|_2 \\ &= \|V_{k+1}(\beta\bar{e}_1 - \bar{H}_k y)\|_2 \neq \|\beta\bar{e}_1 - \bar{H}_k y\|_2 \end{aligned}$$

(since the columns of V_{k+1} are no longer orthogonal), so that the minimization of $\|\beta\bar{e}_1 - \bar{H}_k y\|_2$ is no longer equivalent with the minimization of $\|b - Ax\|_2$. Since, in general, the vectors v_i are approximately orthogonal, it is still meaningful to compute the optimal solution y_k of

$$\min_{y \in \mathbb{R}^k} \|\beta\bar{e}_1 - \bar{H}_k y\|_2$$

and the associated vector $x_k := x_0 + V_k y_k$: the vector x_k will minimize $\|b - Ax\|_2$ on $x_0 + K_k(r_0, A)$ not exactly, but to a good approximation.

Since $\bar{H}_k$ and the triangular matrix R_k now are band matrices of bandwidth l and $l+1$, respectively, the use of the recursions (6.35) and (6.36) is advantageous. One then has to

store only the l vectors $v_k, v_{k-1}, ..., v_{k-l+1}$ and l additional vectors $p_{k-1}, ..., p_{k-l}$. A storage of the full matrix R_k is no longer necessary, only the last column of R_k is needed. Formulae (6.31) and (6.35) simplify, because of $h_{ik} = 0$ for $i \leq k-l$, $r_{ik} = 0$ for $i \leq k-(l+1)$, and they read

$$\begin{aligned} \tilde{r}_k &= \Omega_{k-1}\Omega_{k-2}...\Omega_1 \bar{h}_k \\ p_k &= \frac{1}{r_{kk}} \left(v_k - \sum_{i=\max\{1,k-l\}}^{k-1} r_{ik} p_i \right). \end{aligned}$$

In sum, one obtains the following incomplete quasi-minimal GMRES method (QGMRES), also denoted by QGMRES(l):

(6.38) **QGMRES algorithm**

Given $\epsilon > 0$, l an integer with $2 \leq l << n$, and x_0 with $r_0 := b - Ax_0 \neq 0$.

(0) Put $\beta := \bar{\gamma}_0 := \|r_0\|_2$, $v_1 := r_0/\beta$, $k := 1$.

(1) Compute $w := Av_k$.

(2) For $i = 1, 2, ..., k$, compute

$$\begin{aligned} h_{ik} &:= \begin{cases} 0 & \text{if } i \leq k-l; \\ v_i^T w & \text{otherwise.} \end{cases} \\ w &:= w - h_{ik} v_i. \end{aligned}$$

(3) Compute $h_{k+1,k} := \|w\|_2$, and thus the vector $\bar{h}_k = [h_{1k}, h_{2k}, ..., h_{k+1,k}]^T$.

(4) Compute $\tilde{r}_k := \Omega_{k-1}\Omega_{k-2}...\Omega_1 \bar{h}_k$, the rotation parameters c_k, s_k by (6.32), $\gamma_k, \bar{\gamma}_{k+1}$ by (6.34) and the vector $\bar{r}_k$ by (6.33) i.e.,

$$\bar{r}_k = \begin{bmatrix} r_{1k} \\ r_{2k} \\ \vdots \\ r_{kk} \\ 0 \end{bmatrix} := \Omega_k \tilde{r}_k.$$

(5) Compute

$$p_k := \frac{1}{r_{kk}} \left(v_k - \sum_{i=\max\{1,k-l\}}^{k-1} r_{ik} p_i \right).$$

(6) Compute $x_k := x_{k-1} + \gamma_k p_k$.

(7) If $|\bar{\gamma}_{k+1}| \leq \epsilon$, stop. Otherwise, set $v_{k+1} := w/h_{k+1,k}$, $k := k+1$ and go to (1).

Remark 6.3.7. *For symmetric, but indefinite matrices $A = A^T$, the method of Arnoldi is identical with the Lanczos method (see Remark (5.3.2)). As in section 5.3, one can show that all scalar products $h_{ik} = v_i^T A v_k = 0$, $1 \leq i \leq k-2$, vanish in this case and*

$$h_{k,k+1} = h_{k+1,k}, \quad k = 1, 2, ..., n.$$

Then also the matrices

$$H_k := \begin{bmatrix} h_{11} & h_{12} & 0 & \dots & 0 \\ h_{21} & h_{22} & h_{23} & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & h_{k-1,k-1} & h_{k-1,k} \\ 0 & \dots & 0 & h_{k,k-1} & h_{kk} \end{bmatrix}$$

are symmetric tridiagonal matrices. Therefore, we have before us, without artificial truncation, the case $l = 2$ of the incomplete GMRES method, and the method reduces to the ***SYMMLQ*** *method of Paige and Saunders* (1975, [107]*).*

Preconditioned GMRES method.

Similarly as in the Conjugate-gradient algorithm [cf. (6.10)], it is possible to accelerate also the convergence of the GMRES-method by preconditioning techniques. These techniques are based on the choice of a preconditioning matrix B with the following properties:

(1) B is a good approximation of A so that $B^{-1}A$ respectively AB^{-1} are approximations of the unity matrix.

(2) Equations of the form $Bu = v$ are easy to solve, that is, it is simple to compute $B^{-1}v$.

NB: Property (2) is satisfied if one knows the LR decomposition of $B = LR$, where L and R are sparse triangular matrices.

Having a preconditioner B, one has the choice between **left preconditioning** and **right preconditioning:** with left preconditioning, the GMRES method is applied to the system

$$B^{-1}Ax = B^{-1}b,$$

and with right preconditioning, to the system

$$AB^{-1}u = b$$

in the new variable $u = Bx$. Both systems are equivalent to $Ax = b$.

In this section, we only describe left preconditioning. We then have to modify the GMRES method by replacing the matrix A by $B^{-1}A$ and the residual $r_0 = b - Ax_0$ by the new residual $q_0 := B^{-1}b - B^{-1}Ax_0 = B^{-1}r_0$. We then obtain the following algorithm instead of (6.17).

(6.39) **GMRES with left preconditioning (PGMRES)**

Given $\epsilon > 0$, x_0 with $r_0 := b - Ax_0 \neq 0$.

(0) Compute $q_0 := B^{-1}r_0$, $\beta := \bar{\gamma}_0 := \|q_0\|_2$, $v_1 := q_0/\beta$ and set $k := 1$.

(1) Compute $w := B^{-1}Av_k$.

(2) For $i = 1, 2, ..., k$:
Compute $h_{ik} := v_i^T w$, $w := w - h_{ik}v_i$.

(3) Compute $h_{k+1,k} := \|w\|_2$ and $\bar{\gamma}_{k+1}$ [cf. (6.34)].

(4) If $|\bar{\gamma}_{k+1}| > \epsilon$,
Compute $v_{k+1} := w/h_{k+1,k}$, set $k := k + 1$ and go to (1). Otherwise,

(5) Compute the solution y_k of (6.29) and $x_k := x_0 + V_k y_k$,
set $m := k$ and stop.

Conclusion: The Krylov spaces $K_k(q_0, B^{-1}A)$, $k = 1, 2, ..., k$ have the orthonormal bases $v_1, v_2, ..., v_k$ and the method computes the first vector $x_m \in x_0 + K_m(q_0, B^{-1}A)$ with

$$\|B^{-1}(b - Ax_m)\|_2 = \min_u \{\|B^{-1}(b - Au)\|_2 : \ u \in x_0 + K_m(q_0, B^{-1}A)\} \leq \epsilon.$$

Clearly, there are similar preconditioned versions of the truncated [cf. GMRES(N), (6.37)] and incomplete versions [see: QGMRES(l), (6.38)] of the GMRES algorithm.

6.4 Biorthogonalization method of Lanczos and the QMR algorithm

There are additional Krylov space methods for solving linear equations $Ax = b$ with arbitrary real or complex nonsingular $n \times n$ matrices A. These methods work with pairs of Krylov spaces:

$$\begin{aligned} K_k(v_1, A) &= span[v_1, Av_1, ..., A^{k-1}v_1] \\ K_k(w_1, A^T) &= span[w_1, A^T w_1, ..., (A^T)^{k-1}w_1], \end{aligned}$$

and not with single spaces, as the methods considered so far. Even though these methods are applicable to systems with a complex matrix A, we still assume that A is real.

Again, let x_0 be an initial approximate solution of $Ax = b$ with $r_0 = b - Ax_0 \neq 0$. Then the following biorthogonalization algorithm of Lanczos (1950, [95]) starts with the vectors

$$v_1 := r_0/\beta, \quad \beta := \|r_0\|_2,$$

and an arbitrary additional vector $w_1 \in \mathbb{R}^n$ with $\|w_1\|_2 = 1$ (a common vector is $w_1 = v_1$). The algorithm seeks to generate two, as long as possible, sequences $\{v_i\}_{i\geq 1}$ and $\{w_i\}_{i\geq 1}$ of linearly independent vectors that are biorthogonal, that is,

$$w_i^T v_j = \begin{cases} \delta_j & \text{for } i \neq j \\ 0 & \text{otherwise.} \end{cases}$$

and span Kylov spaces $K_k(v_1, A)$ and $K_k(w_1, A^T)$, $k \geq 1$, respectively,

$$span[v_1, Av_1, ..., A^{k-1}v_1] = K_k(v_1, A), \ \ span[w_1, A^T w_1, ..., (A^T)^{k-1}w_1] = K_k(w_1, A^T).$$

(6.40) **Biorthogonalization method of Lanczos**

Given $x_0 \in \mathbb{R}^n$ with $r_0 := b - Ax_0 \neq 0$, set $\beta := \|r_0\|_2$, $v_1 := r_0/\beta$, choose $w_1 \in \mathbb{R}^n$ with $\|w_1\|_2 = 1$, and let $v_0 := w_0 := 0$, $k := 1$.

(1) Compute $\delta_k := w_k^T v_k$. If $\delta_k = 0$, set $m := k - 1$ and stop. Otherwise,

(2) Compute $\alpha_k := w_k^T A v_k/\delta_k$, $\beta_1 := \epsilon_1 := 0$, and for $k > 1$,

$$\beta_k := \frac{\sigma_k \delta_k}{\delta_{k-1}}, \quad \epsilon_k := \frac{\rho_k \delta_k}{\delta_{k-1}},$$

and

$$\begin{aligned} \tilde{v}_{k+1} &:= Av_k - \alpha_k v_k - \beta_k v_{k-1}, \\ \tilde{w}_{k+1} &:= A^T w_k - \alpha_k w_k - \epsilon_k w_{k-1}. \end{aligned}$$

(3) Compute $\rho_{k+1} := \|\tilde{v}_{k+1}\|_2$, $\sigma_{k+1} := \|\tilde{w}_{k+1}\|_2$.
If $\rho_{k+1} = 0$ or $\sigma_{k+1} = 0$, set $m := k$ and stop. Otherwise,

(4) Compute $v_{k+1} := \tilde{v}_{k+1}/\rho_{k+1}$, $w_{k+1} := \tilde{w}_{k+1}/\sigma_{k+1}$.

(5) Set $k := k + 1$ and go to (1).

Theorem 6.4.1. *Let m be the break-off index of the algorithm* (6.40). *Then for all $1 \leq k \leq m$,*

$$span[v_1, v_2, ..., v_k] = K_k(v_1, A), \tag{6.41}$$
$$span[w_1, w_2, ..., w_k] = K_k(w_1, A^T).$$

and

$$w_k^T v_j = \begin{cases} \delta_j \neq 0 & \text{for } j = k \\ 0 & \text{for } j \neq k \end{cases} \qquad j, k = 1, 2, ..., m. \tag{6.42}$$

The vectors $v_1, v_2, ..., v_m$ and also the vectors $w_1, w_2, ..., w_m$ are linearly independent.

Proof. Steps (2) to (4) of the algorithm (6.40) imply immediately (6.41). The biorthogonality (6.42) is shown by mathematical induction. It is trivial true for $m = 0$, and if $m \geq 1$ also for $k = 1$. Assume inductively that for some k with $1 \leq k < m$ the following holds [cf. (6.42)]

$$w_i^T v_j = 0, \;\; v_i^T w_j = 0 \;\; 1 \leq i < j \leq k.$$

Because of $k < m$, it follows that $\delta_j \neq 0$ for all $j \leq k$, $\rho_{k+1} \neq 0$, $\sigma_{k+1} \neq 0$, and the vectors v_{k+1} and w_{k+1} are well-defined. We wish to show that also the vectors $v_1, v_2, ..., v_{k+1}$ and $w_1, w_2, ..., w_{k+1}$ are biorthogonal.

First, we show that $w_i^T v_{k+1} = 0$ for $i \leq k$. For $i = k$, this follows from the definition of $\tilde{v}_{k+1}$. the induction hypothesis, and the definition of α_k, since

$$\begin{aligned} w_k^T v_{k+1} &= \frac{1}{\rho_{k+1}}[w_k^T A v_k - \alpha_k w_k^T v_k - \beta_k w_k^T v_{k-1}] \\ &= \frac{1}{\rho_{k+1}}[w_k^T A v_k - \alpha_k w_k^T v_k] = 0. \end{aligned}$$

For $i \leq k - 1$, the induction hypothesis and the definition of $\tilde{w}_{k+1}$ give

$$\begin{aligned} w_i^T v_{k+1} &= \frac{1}{\rho_{k+1}}[w_i^T A v_k - \alpha_k w_i^T v_k - \beta_k w_i^T v_{k-1}] \\ &= \frac{1}{\rho_{k+1}}[w_i^T A v_k - \beta_k w_i^T v_{k-1}] \\ &= \frac{1}{\rho_{k+1}}[v_k^T(\tilde{w}_{i+1} + \alpha_i w_i + \epsilon_i w_{i-1}) - \beta_k w_i^T v_{k-1}] \\ &= \frac{1}{\rho_{k+1}}[\sigma_{i+1} v_k^T w_{i+1} + \alpha_i v_k^T w_i + \epsilon_i v_k^T w_{i-1} - \beta_k w_i^T v_{k-1}] = 0. \end{aligned}$$

Indeed, the induction hypothesis then implies $w_{i+1}^T v_k = 0$ for $i < k - 1$, and for $i = k - 1$, by the definition of β_k one has

$$\sigma_{i+1} w_{i+1}^T v_k = \beta_k \delta_{k-1} = \beta_k w_i^T v_{k-1}.$$

In the same way, one shows that $v_i^T w_{k+1} = 0$ for $i \leq k$. Finally, in terms of the matrices $V_k := [v_1, v_2, ..., v_k]$, $W_k := [w_1, w_2, ..., w_k]$, and the diagonal matrices $D_k := diag(\delta_1, \delta_2, ..., \delta_k)$, relation (6.42) is equivalent to the equation

$$W_m^T V_m = D_m.$$

The nonsingularity of D_m, and $W_m^T V_m = D_m$ then imply $rank\, V_m = rank\, W_m = m$. □

(6.43) **Transformation of recursion** (6.40) **in terms of matrices**

Similarly to Arnoldi's method (6.17), the recursions (6.40) can be expressed in terms of the matrices V_k, W_k, the tridiagonal Hessenberg matrices

$$(6.44) \qquad \bar{T}_k := \begin{bmatrix} \alpha_1 & \beta_2 & & & 0 \\ \rho_2 & \alpha_2 & \beta_3 & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & \beta_k \\ & & & \ddots & \alpha_k \\ 0 & & & & \rho_{k+1} \end{bmatrix}, \quad \bar{S}_k := \begin{bmatrix} \alpha_1 & \epsilon_2 & & & 0 \\ \sigma_2 & \alpha_2 & \epsilon_3 & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & \epsilon_k \\ & & & \ddots & \alpha_k \\ 0 & & & & \sigma_{k+1} \end{bmatrix},$$

and their submatrices T_k and S_k of order k obtained by deleting the last row of $\bar{T}_k$ and $\bar{S}_k$, respectively: as in the proof of (6.24), (6.25), one shows for $k < m$

$$\begin{aligned} AV_k &= V_{k+1}\bar{T}_k = V_k T_k + \tilde{v}_{k+1} e_k^T, \\ A^T W_k &= W_{k+1}\bar{S}_k = W_k S_k + \tilde{w}_{k+1} e_k^T. \end{aligned}$$

Because of $W_k^T V_k = D_k$, $W_k^T \tilde{v}_{k+1} = V_k^T \tilde{w}_{k+1} = 0$, this implies

$$W_k^T A V_k = D_k T_k, \quad V_k^T A^T W_k = D_k S_k,$$

and, by $W_k^T A V_k = (V_k^T A^T W_k)^T$, also

$$S_k^T = D_k T_k D_k^{-1}.$$

NB: The last identity can also be verified by means of the definitions of β_k and ϵ_k.

Remark 6.4.1. *The break-off behavior of the biorthogonalization algorithm* (6.40) *is more complicated than that of the Arnoldi's method* (6.17).
It can terminate in step (4) *if* $\rho_{k+1} = 0$ *and/or* $\sigma_{k+1} = 0$. *Because of* $k = rank\, V_k = dim\, K_k(v_1, A) = rank\, W_k = dim\, K_k(w_1, A^T)$, *this is equivalent to*

$$dim\, K_{k+1}(v_1, A) = k, \quad respectively \;\; dim\, K_{k+1}(w_1, A^T) = k,$$

that is, with the A-invariance of $K_k(v_1, A)$, *respectively, the* A^T*-invariance of* $K_k(w_1, A^T)$. *The columns of* V_k *(respectively,* W_k*) then provide a basis of the invariant Krylov space* $K_k(v_1, A)$ *(respectively,* $K_k(w_1, A^T)$*). Therefore, the break-off index* m *of* (6.40) *is at most equal to* n, $m \leq n$.
Unfortunately, Lanczos' algorithm (6.40) *can also stop in step* (1) *because of* $\delta_k = w_k^T v_k = 0$ *even though, in this case, both vectors* v_k *and* w_k *are nonzero. Then the method terminates "prematurely", that is, before invariant Krylov spaces have been found. This situation marks a so-called "serious breakdown" of the method.*

Remark 6.4.2. *In floating point calculations, already situations in which* $|\delta_k| \approx 0$ *becomes too small ("numerical breakdown") may cause a serious loss of accuracy of the quantities computed in* (6.40). *However, these dangerous situations can be (almost) always avoided by employing "look ahead techniques", which weaken the biorthogonality requirements* (6.42). *The QMR method of Freud and Nachtigal* (1991, [65]*) is such a variant of* (6.40) *which avoid (almost) all numerical breakdowns, and still provides bases* $v_1, v_2, ..., v_k$ *of* $K_k(v_1, A)$, *and bases* $w_1, w_2, ..., w_k$ *of* $K_k(w_1, A^T)$, *without changing the simple structure of matrices* $\bar{T}_k, \bar{S}_k$ *[see,* (6.44)*] too much. In their method, these matrices will become block tridiagonal matrices. Their block components* $\alpha_i, \beta_i, \rho_i, \epsilon_i, \sigma_i$ *are no longer numbers but simple matrices with a very small number of rows and columns. The dimension of these blocks are determined in such a way that too small* $|\delta_k|$ *are avoided. Since the details of the QMR method are fairly sophisticated, we refer the reader to Freud and Nachtigal* (1991, [65]*).*

In the interest of simplicity, we restrict ourselves in this representation to describe the QMR method in the basic situation when no numerical breakdown occurs. So, we assume that (6.40) never terminates in step (1) but only in step (4). Then, $\delta_k \neq 0$ for all $k \leq m+1$, the columns of V_k provide a basis of $K_k(v_1, A) = K_k(r_0, A)$, for $k \leq m$, and

$$AV_k = V_{k+1}\bar{T}_k = V_k T_k + \tilde{v}_{k+1} e_k^T.$$

Each $x \in x_0 + K_k(r_0, A)$ then has the form $x = x_0 + V_k y$ with a unique vector $y \in \mathbb{R}^k$, and, as in section 6.3,

$$\begin{aligned} \|b - Ax\|_2 &= \|b - Ax_0 - AV_k y\|_2 \\ &= \|r_0 - V_{k+1}\bar{T}_k y\|_2 \\ &= \|V_{k+1}(\beta \bar{e}_1 - \bar{T}_k y)\|_2, \end{aligned}$$

where $\bar{e}_1 := [1, 0, ..., 0]^T \in \mathbb{R}^{k+1}$. Instead of minimizing $\|b - Ax\|_2$ over $x_0 + K_k(r_0, A)$, one determines x_k as in quasi-minimal GMRES method (6.38): one computes the solution y_k of the least-squares problem

$$\min_{y \in \mathbb{R}^k} \|\beta \bar{e}_1 - \bar{T}_k y\|_2 \tag{6.45}$$

and sets $x_k = x_0 + V_k y_k$.

The calculations are as in the incomplete quasi-minimal GMRES method (QGMRES (6.38)), if we replace in (6.38) the Hessenberg matrix $\bar{H}_k$ of bandwidth l by the tridiagonal matrix $\bar{T}_k$ (a band matrix of bandwidth $l = 2$) of this section: We only have to choose $l = 2$ in (6.38), and to replace the vector h_k by

$$t_k = \begin{bmatrix} t_{1k} \\ \vdots \\ t_{k-1,k} \\ t_{k,k} \\ t_{k+1,k} \end{bmatrix} := \begin{bmatrix} 0 \\ \vdots \\ 0 \\ \beta_k \\ \alpha_k \\ \rho_{k+1} \end{bmatrix}, \tag{6.46}$$

the last column of $\bar{T}_k$.

In this way, we obtain a much simplified version of the QMR method, which, however, does not take the possibility of a serious or of a numerical breakdown (6.40) into account, which one should in practice.

(6.47) **QMR method**

Given $x_0 \in \mathbb{R}^n$ with $r_0 := b - Ax_0 \neq 0$ and $\epsilon > 0$.

Compute $\beta := \|r_0\|_2$, $v_1 = w_1 := r_0/\beta$ and set $k := 1$.

(1) Use (6.40) and (6.46) to determine $\alpha_k, \beta_k, \epsilon_k, \rho_{k+1}, \sigma_{k+1}, v_{k+1}, w_{k+1}$ and the last column t_k of $\bar{T}_k$.

(2) Compute $\tilde{r}_k := \Omega_{k-1}\Omega_{k-2}t_k$ (with $\Omega_{-1} = \Omega_0 := I$), the rotation parameters c_k, s_k of Ω_k as in (6.32), and $\gamma_k, \bar{\gamma}_{k+1}$ as in (6.34).

(3) Compute the vectors

$$\begin{aligned} \bar{r}_k &= \begin{bmatrix} r_{1k} \\ \vdots \\ r_{kk} \\ 0 \end{bmatrix} := \Omega_k \tilde{r}_k \\ p_k &:= \frac{1}{r_{kk}} \left(v_k - \sum_{i=k-2}^{k-1} r_{ik} p_i \right). \end{aligned}$$

(4) Compute $x_k := x_{k-1} + \gamma_k p_k$.

(5) If $|\bar{\gamma}_{k+1}| \leq \epsilon$, then stop. Otherwise,
set $k := k+1$ and go to (1).

Remark 6.4.3. *As in the GMRES algorithm, it is possible to improve the efficiency of the QMR method by incorporing preconditioning techniques [see Freud and Nachtigal* (1991, [65]*) for details].*

Remark 6.4.4. *(QMR approach with preconditioning* [65]*)*
Let M be a given nonsingular $n \times n$ matrix which approximates in some sense the coefficients matrix A of the linear system (6.1)*, $Ax = b$. Moreover, assume that M is decomposed in the form*

$$M = M_1 M_2. \tag{6.48}$$

Instead of solving the original linear system (6.1)*, we apply the QMR algorithm to the equivalent linear system*

$$A^{'} y = b^{'}, \;\; where \; A^{'} = M_1^{-1} A M_2^{-1}, \;\; b^{'} = M_1^{-1}(b - Ax_0), \;\; y = M_2(x - x_0).$$

Here x_0 denotes some initial guess for the solution of (6.1).

- ***SSOR***

The SSOR preconditioner is based on the decomposition of the matrix A into a nonsingular diagonal matrix D, a strictly lower triangular matrix L, and a strictly upper triangular matrix U, such that $A = D + L + U$. D might have to be block diagonal to ensure it is nonsingular. The preconditioner matrix M is given by

$$M = (D + L)D^{-1}(D + U).$$

- ***ILUT(k)***

The Incomplete LU decomposition is based on the LU decomposition of the coefficient matrix A into a unit lower triangular matrix L and an upper triangular matrix U. The full LU decomposition of A would result in factors L and U which, in general, have far more nonzero elements than A. The incomplete LU factorization aims to reduce this additional fill-in in the factors L and U.

In ILUT(k), we use a strategy due to Saad [$115^{(1)}$] *for dropping nonzero elements which would fill-in L and U. Each row of L and U is constructed subject to the restriction that only a small amount of fill-in, k more elements for each, is allowed beyond the number of elements of A already present in that row (in the lower and upper part, respectively). Furthermore, elements which are deemed to make only an insignificant contribution to the decomposition are also dropped. For example, this means that if $nmax_L$ is the maximum number of elements allowed for some row of L, n_L is the actual number of elements of that row computed by the elimination process, and ctol is the cutoff tolerance, then the algorithm orders the n_L elements in decreasing order of magnitude, and keeps only up to $min(n_L; nmax_L)$ elements, or until the elements reach the level ctol, whichever cutoff comes first. The resulting matrices L and U can be used either as $M_1 = L$ and $M_2 = U$ in* (6.48)*, or in $M_2 = LU$ respectively $M_1 = LU$ for right respectively left preconditioning.*

The variant of ILU used is different from the standard one. For a Hermitian matrix A, the standard ILU preconditioner [99] *preserves the sparsity structure of the matrix, i.e., for $k = 0$, the preconditioner matrices have nonzero elements only in those locations where A itself has nonzero elements. In* [99] *it is shown that this strategy does produce a good preconditioner, provided that A is a Hermitian M-matrix. For a general non-Hermitian matrix, there is no reason to preserve the sparsity structure of A. Instead, the ILUT(k) variant discards elements subject only to the constraints of fill-in and size, without regard to the sparsity structure of A. However, this does mean that if A is Hermitian, we do not recover the standard ILU preconditioner.*

6.5 Bi-CG and BI-CGSTAB algorithms

The biconjugate gradient (Bi-CG, also BCG) algorithm for solving a system $Ax = b$ with a nonsymmetric (real) $n \times n$ matrix A is a direct generalization of the classical cg-method (6.3) of Hestenes and Stiefel (see [80]). It is related to the biorthogonalization algorithm (6.40) and is due to Lanczos (1950, [95]) and Fletcher (1976, [63]).
In what follows, (v, w) and $\|v\|_2$ always denote the usual scalar product: $(v, w) = v^T w$ and the Euclidian norm $\|v\|_2 = (v, v)^{1/2}$, respectively.

(6.49) **Bi-CG algorithm**

Initialization: Given $x_0 \in \mathbb{R}^n$ with $r_0 := b - Ax_0 \neq 0$. Choose $\hat{r}_0 \in \mathbb{R}^n$ with $(\hat{r}_0, r_0) \neq 0$ and set $p_0 := r_0$, $\hat{p}_0 = \hat{r}_0$.

For $k = 1, 2, \ldots$:
Compute

(1) $a_k := \frac{(\hat{r}_k, r_k)}{(\hat{p}_k, Ap_k)}$, $x_{k+1} := x_k + a_k p_k$,
$r_{k+1} := r_k - a_k A p_k$, $\hat{r}_{k+1} := \hat{r}_k - a_k A^T \hat{p}_k$.

(2) $b_k := \frac{(\hat{r}_{k+1}, r_{k+1})}{(\hat{r}_k, r_k)}$,
$p_{k+1} := r_{k+1} + b_k p_k$ $\hat{p}_{k+1} := \hat{r}_{k+1} + b_k \hat{p}_k$.

Remark 6.5.1. *The algorithm is well-defined as long as $(\hat{r}_k, r_k)$ and $(\hat{p}_k, Ap_k)$ remain nonzero. Its theoretical properties are comparable to those of the cg-algorithm [cf. Theorem 6.2.1].*

Theorem 6.5.1. *Let A be any real nonsingular $n \times n$ matrix and $b \in \mathbb{R}^n$. Then, to any starting vectors $x_0 \in \mathbb{R}^n$, $\hat{r}_0$ with $(\hat{r}_0, r_0) \neq 0$, $r_0 := b - Ax_0$, the vectors $x_k, p_k, \hat{p}_k, r_k, \hat{r}_k$ generated by (6.49) have the following properties:*

There is a first index $m \leq n$ such that $(\hat{r}_m, r_m) = 0$ or $(\hat{p}_m, Ap_m) = 0$, and all assertions (1) to (6) of (A_m) hold:

(A_m):

(1) $(\hat{p}_i, r_j) = (\hat{r}_j, p_i) = 0$ *for* $i < j \leq m$,

(2) $(\hat{r}_i, r_i) \neq 0$ *for* $i \leq m$,
$(\hat{r}_i, p_i) = (\hat{r}_i, r_i) = (\hat{p}_i, r_i) \neq 0$ *for* $i \leq m$,

(3) $(\hat{p}_i, Ap_j) = (A^T \hat{p}_j, p_i) = 0$ *for* $i < j \leq m$,
$(\hat{p}_i, Ap_i) \neq 0$ *for* $i < m$,

(4) $(\hat{r}_i, r_j) = (\hat{r}_j, r_i) = 0$ *for* $i < j \leq m$,

(5) $r_i = b - Ax_i$ *for* $i \leq m$,

(6) *For* $i \leq m$

$$\begin{aligned} span[r_0, r_1, ..., r_i] = span[p_0, p_1, ..., p_i] &= K_{i+1}(r_0, A), \\ span[\hat{r}_0, \hat{r}_1, ...\hat{r}_i] = span[\hat{p}_0, \hat{p}_1, ..., \hat{p}_i] &= K_{i+1}(\hat{p}_0, A^T). \end{aligned}$$

Proof. The proof is by mathematical induction and essentially the same as the proof of Theorem 6.2.1: Property (A_0) is trivially true, and for any $k \geq 0$ the implication

$$(A_k), \quad (\hat{p}_k, Ap_k) \neq 0, \ (\hat{r}_k, r_k) \neq 0 \Rightarrow (A_{k+1})$$

holds. Since $(A_k)(2)(4)$ imply that the vectors $r_i, \hat{r}_i$ are nonzero for $i < k$ and biorthogonal

$$(\hat{r}_i, r_j) = (\hat{r}_j, r_i) = 0 \ \text{ for } i < j < k,$$

then the vectors r_i, $i = 0, 1, 2, ..., k-1$, and $\hat{r}_i$, $i = 0, 1, ..., k-1$, must be linearly independent vectors in $\mathbb{R}^n$. Hence $k \leq n$, so that there is a first index $m \leq n$ such that $(\hat{r}_m, r_m) = 0$ or $(\hat{p}_m, Ap_m) = 0$ holds. □

The iterates x_i exist for $i = 0, 1, 2, ..., m$, but there have no minimization property with respect to the set $x_0 + K_i(r_0, A)$, but only the Galerkin property

$$(w, b - Ax_i) = 0, \quad \text{for all } \ w \in K_{i-1}(\hat{r}_0, A^T).$$

This follows at once form $(A_m)(1)(6)$.

The break-off behavior of the algorithm is related to but even more complicated than that of the Lanczos biorthogonalization algorithm (6.49). First, the algorithm stops if $(\hat{p}_m, Ap_m) = 0$ but both $\hat{p}_m$ and p_m are nonzero: one can show that this happens exactly if there is no $x_{m+1} \in x_0 + K_{m+1}(r_0, A)$ with the Galerkin property. But the algorithm stops also if $(\hat{r}_m, r_m) = 0$ even though the vectors $\hat{r}_m$ and r_m are nonzero: this happens exactly if the Lanczos biorthogonalization algorithm (6.49), when started with $v_1 := r_0/\|r_0\|_2$, $w_1 := \hat{r}_0/\|\hat{r}_0\|_2$, stops because of a "serious break-down" [see section 6.4].

A further drawback of the algorithm is that the sizes $\|r_i\|_2$ of the residuals may behave quite erratically as i increases: usually they fluctuate very much before settling down. Moreover, the accuracy of the computed vectors $r_k, \hat{r}_k, p_k, \hat{p}_k$ and x_k suffer badly due to round-off if a near break down occurs when some of the crucial quantities

$$\frac{(\hat{r}_k, r_k)}{\|r_k\|_2 \|\hat{r}_k\|_2}, \quad \frac{(\hat{p}_k, Ap_k)}{\|Ap_k\|_2 \|\hat{p}_k\|_2}$$

become small.

However, the "convergence" of the residuals r_k and their erratic behavior can be much improved by using techniques proposed by Van der Vorst (1992, [168]) (on the basis of results found by Sonneveld (1989, [147])) in his Bi-CGSTAB algorithm that stabilizes the Bi-CG method. For a description of this method we need some further properties of the vectors generated by the Bi-CG algorithm (6.49).

Proposition 6.5.1. *There are polynomials $R_k(\mu)$, $P_k(\mu)$, $k = 1, 2, ..., m$, of degree k with $R_k(0) \equiv P_k(0) \equiv 1$ satisfying*

$$\left.\begin{array}{l} r_k = R_k(A) r_0, \ \ \hat{r}_k = R_k(A^T)\hat{r}_0 \\ p_k = P_k(A) r_0, \ \ \hat{p}_k = P_k(A^T)\hat{r}_0 \end{array}\right\} \ k = 0, 1, 2, ..., m,$$

and the recursions

$$\left.\begin{array}{l} R_{k+1}(\mu) = R_k(\mu) - a_k \mu P_k(\mu) \\ P_{k+1}(\mu) = R_{k+1}(\mu) + b_k \mu P_k(\mu) \end{array}\right\} \ k = 0, 1, 2, ..., m-1. \tag{6.50}$$

As a consequence of these recursions, the highest order terms of these polynomials are known for $k = 0, 1, ...m$:

$$\begin{array}{l} R_k(\mu) = (-1)^k a_0 a_1 ... a_{k-1} \mu^k + \ \text{ lower order terms}\,, \\ P_k(\mu) = (-1)^k a_0 a_1 ... a_{k-1} \mu^k + \ \text{ lower order terms}\,. \end{array} \tag{6.51}$$

Moreover, property $(A_m)(4)$ of Theorem 6.5.1 implies the orthogonality relation

$$(R_i(A^T)\hat{r}_0, R_j(A)r_0) = (\hat{r}_0, R_i(A)R_j(A)r_0) \text{ for } i < j \leq m. \tag{6.52}$$

We now introduce new vectors

$$\begin{array}{l} \bar{r}_k := Q_k(A)R_k(A)r_0 = Q_k(A)r_k \\ \bar{p}_k := Q_k(A)P_k(A)r_0 = Q_k(A)p_k \end{array}, \quad k = 0, 1, 2, ...,$$

defined by the choice of real polynomials $Q_k(\mu)$ of degree k of the form

$$Q_k(\mu) = (1 - w_1\mu)(1 - w_2\mu)...(1 - w_k\mu),$$

that satisfy the recursion

$$Q_{k+1}(\mu) = (1 - w_{k+1}\mu)Q_k(\mu). \tag{6.53}$$

It will turn out that the vectors $\bar{r}_k$ and $\bar{p}_k$ (and the associated vectors $\bar{x}_k$ with residual $b - A\bar{x}_k = \bar{r}_k$) can be computed directly without using the vectors defined by the Bi-CG algorithm. Moreover, the parameter w_k of Q_k can be chosen such that the size of the new residual $\bar{r}_k$ becomes as small as possible. To see this, we note first that the recursions (6.50) and (6.53) lead to a recursion for $\bar{r}_k$, $\bar{p}_k$:

$$\begin{array}{rcl} \bar{r}_{k+1} & = & Q_{k+1}(A)R_{k+1}(A)r_0 \\ & = & (1 - w_{k+1}A)Q_k(A)[R_k(A) - a_kAP_k(A)]r_0 \\ & = & [Q_k(A)R_k(A) - a_kAQ_k(A)P_k(A)]r_0 \\ & - & w_{k+1}A[Q_k(A)R_k(A) - a_kAQ_k(A)P_k(A)]r_0 \\ & = & \bar{r}_k - a_kA\bar{p}_k - w_{k+1}A(\bar{r}_k - a_kA\bar{p}_k). \end{array} \tag{6.54}$$

$$\begin{array}{rcl} \bar{p}_{k+1} & = & Q_{k+1}(A)P_{k+1}(A)r_0 \\ & = & Q_{k+1}(A)[R_{k+1}(A) + b_kP_k(A)]r_0 \\ & = & \bar{r}_{k+1} + (1 - w_{k+1}A)b_kQ_k(A)P_k(A)r_0 \\ & = & \bar{r}_{k+1} + b_k(\bar{p}_k - w_{k+1}A\bar{p}_k). \end{array} \tag{6.55}$$

Next, we show that the quantities a_k and b_k can be expressed in terms of the vectors $\bar{r}_j$ and $\bar{p}_j$. For this purpose, we introduce new quantities ρ_k and $\bar{\rho}_k$ by:

$$\begin{array}{c} \rho_k := (\hat{r}_k, r_k), \\ \bar{\rho}_k := (\hat{r}_0, \bar{r}_k) = (\hat{r}_0, Q_k(A)R_k(A)r_0) = (Q_k(A^T)\hat{r}_0, R_k(A)r_0). \end{array} \tag{6.56}$$

Now, the highest order term of $Q_k(\mu)$ is

$$(-1)^k w_1 w_2 \ldots w_k \mu^k$$

and, by (6.51), each power μ^i with $i < k$ $(\leq m)$ can be expressed as a linear combination of the polynomials $R_j(\mu)$ with $j < k$. Therefore, the orthogonality relations (6.52) and (6.56) give

$$\bar{\rho}_k = (-1)^k w_1 w_2 ... w_k \left((A^T)^k \hat{r}_0, R_k(A)r_0\right).$$

This implies, using the same orthogonality arguments and (6.51),

(6.57)
$$\begin{aligned}\rho_k &= (\hat{r}_k, r_k)\\ &= (R_k(A^T)\hat{r}_0, R_k(A)r_0)\\ &= (-1)^k a_0 \dots a_{k-1}\left((A^T)^k\hat{r}_0, R_k(A)r_0\right) = \bar{\rho}_k \frac{a_0}{w_1}\cdot\dots\cdot\frac{a_{k-1}}{w_k}.\end{aligned}$$

Therefore $b_k = (\hat{r}_{k+1}, r_{k+1})/(\hat{r}_k, r_k)$ can also be computed as

(6.58)
$$b_k = \frac{\bar{\rho}_{k+1}}{\bar{\rho}_k}\cdot\frac{a_k}{w_{k+1}}.$$

We reexpress $(\hat{p}_k, Ap_k)$ using $(A_m)(3)$, (6.49) and (6.51):

$$\begin{aligned}(\hat{p}_k, Ap_k) &= (\hat{r}_k + b_{k-1}\hat{p}_{k-1}, Ap_k)\\ &= (\hat{r}_k, Ap_k) = (R_k(A^T)\hat{r}_0, AP_k(A)r_0)\\ &= (-1)^k a_0 a_1 \dots a_{k-1}((A^T)^k\hat{r}_0, AP_k(A)r_0).\end{aligned}$$

On the other hand, again using $(A_m)(3)$,

$$\begin{aligned}(\hat{r}_0, A\bar{p}_k) &= (\hat{r}_0, AQ_k(A)P_k(A)r_0)\\ &= (Q_k(A^T)\hat{r}_0, AP_k(A)r_0)\\ &= (-1)^k w_1 w_2 \dots w_k((A^T)^k\hat{r}_0, AP_k(A)r_0)\end{aligned}$$

holds so that

$$(\hat{p}_k, Ap_k) = \frac{a_0}{w_1}\cdot\dots\cdot\frac{a_{k-1}}{w_k}(\hat{r}_0, A\bar{p}_k).$$

Together with (6.57), this gives an alternative formula for

$$a_k = (\hat{r}_k, r_k)/(\hat{p}_k, Ap_k),$$

namely

(6.59)
$$a_k = \frac{(\hat{r}_0, \bar{r}_k)}{(\hat{r}_0, A\bar{p}_k)} = \frac{\bar{\rho}_k}{(\hat{r}_0, A\bar{p}_k)}.$$

So far, the choice of w_{k+1} was left opened: since we wish to achieve the small residuals $\bar{r}_i$, it is reasonable to choose w_{k+1} such that the norm $\|\bar{r}_{k+1}\|_2$ of [cf. (6.54)], that is,

$$\bar{r}_{k+1} = s_k - w_{k+1}t_k,$$

where

$$s_k := \bar{r}_k - a_k A\bar{p}_k, \quad t_k := As_k,$$

becomes minimal. This leads to the choice

$$w_{k+1} := \frac{(s_k, t_k)}{(t_k, t_k)}.$$

Finally, if $\bar{r}_k = b - A\bar{x}_k$ is the residual of $\bar{x}_k$, then (6.54) shows that $\bar{r}_{k+1}$ is the residual of

(6.60)
$$\bar{x}_{k+1} := \bar{x}_k + a_k\bar{p}_k + w_{k+1}(\bar{r}_k - a_k A\bar{p}_k).$$

Combining the formulas (6.54), (6.55) and (6.60) lead to the Bi-CGSTAB algorithm:

(6.61) **Bi-CGSTAB algorithm**

Initialization: Given $\bar{x}_0 \in \mathbb{R}^n$ with $\bar{r}_0 := b - A\bar{x}_0 \neq 0$. Choose $\hat{r}_0 \in \mathbb{R}^n$ so that $(\hat{r}_0, \bar{r}_0) \neq 0$ and set $\bar{p}_0 := \bar{r}_0$.

For $k = 0, 1, 2, ...$:
Compute

(1)

$$\begin{aligned} v &:= A\bar{p}_k; \quad a_k := \frac{(\hat{r}_0, \bar{r}_k)}{(\hat{r}_0, v)}; \\ s &:= \bar{r}_k - a_k v; \quad t := As. \end{aligned}$$

(2)

$$\begin{aligned} w_{k+1} &:= \frac{(s,t)}{(t,t)}; \\ \bar{x}_{k+1} &:= \bar{x}_k + a_k\bar{p}_k + w_{k+1}s; \quad \bar{r}_{k+1} := s - w_{k+1}t. \end{aligned}$$

Stop, if $\|\bar{r}_{k+1}\|_2$ is small enough. Otherwise,

(3)

$$\begin{aligned} b_k &:= \frac{(\hat{r}_0, \bar{r}_{k+1})}{(\hat{r}_0, \bar{r}_k)} \frac{a_k}{w_{k+1}}; \\ \bar{p}_{k+1} &:= \bar{r}_{k+1} + b_k(\bar{p}_k - w_{k+1}v). \end{aligned}$$

Remark 6.5.2. *In step k, $(\hat{r}_0, \bar{r}_k)$ need not be determined, since it has already been computed in step $k-1$. An operation count shows that each iteration of Bi-CGSTAB requires the computation of two matrix-vector products with the $n \times n$ matrix A, four inner products, and $12n$ additional floating point operations to update various vectors of length n.*

Bi-CGSTAB is a powerful algorithm for solving even very large systems $Ax = b$ with a sparse nonsymmetric matrix A. It is possible to increase its efficiency still further by preconditioning techniques [see Van der Vorst (1992, [168])*].*

However, even though the stability of Bi-CGSTAB is much better than that of Bi-CG, it will break down whenever the underlying Bi-CG method breaks down. Compared with the QMR method, Bi-CGSTAB is much simpler but not so stable: unlike QMR, Bi-CGSTAB takes no precautions against the danger of "serious" or "nearly serious" break-downs and is also affected if the Galerkin condition defines some iterates only badly.

6.6 Multigrid methods

Multigrid methods belong to the most efficient methods for the solution of those linear equations that result from the discretization of differential equations. As these methods are very flexible, there are many variants of them. Here we wish to explain only the basic ideas behind these powerful methods, and do this in a rather simple situation, which however, already reveals their typical properties. For a detailed treatment, we have to refer the reader to the special literature, for instance Brandt (1977, [20]), Hackbusch and Trottenberg (1982, [79]), and the monographs of Hackbusch (1985, [78]), Bramble (1993, [19]), and Braess (1997, [18]). Our treatment follows the elementary exposition of Briggs (1987, [25]). Instead of boundary value problems for partial differential equations, where multigrid methods have their greatest impact, we only consider their application to the boundary value problem

$$\begin{cases} -y''(x) = f(x) & \text{for } x \in \Omega := (0, \pi), \\ y(0) = y(\pi) = 0 \end{cases} \tag{6.62}$$

for an ordinary differential equation, which can be viewed as the one-dimensional analog of the two-dimensional model problem (6.79). The standard discretization with the grid size $h = \pi/n$ leads to a one-dimensional grid $\Omega_h = \{x_j = jh : \ j = 1, 2, ..., n-1\} \subset \Omega$ and the following set of linear equations for a vector $u_h = [u_{h,1}, u_{h,2}, ..., u_{h,n-1}]^T$ of the approximations $u_{h,j} \approx y(x_j)$ for the exact solution y on the grid Ω_h:

$$A_h u_h = f_h, \ \ A_h := \frac{1}{h^2} \begin{bmatrix} 2 & -1 & & 0 \\ -1 & 2 & \ddots & \\ & \ddots & \ddots & -1 \\ 0 & & -1 & 2 \end{bmatrix}, \ \ f_h := \begin{bmatrix} f(x_1) \\ f(x_2) \\ \vdots \\ f(x_{n-1}) \end{bmatrix}. \tag{6.63}$$

The index h also indicates that u_h and f_h can be viewed as functions on the grid Ω_h. Therefore, we will sometimes write the j-th component $u_{h,j}$ of u_h as the value of a grid function $u_h(x)$ for $x = x_j \in \Omega_h$, $u_{h,j} = u_h(x_j)$. The matrix A_h is a matrix of order $n-1$, for which the eigenvalues $\lambda_h^{(k)}$ and the eigenvectors $z_h^{(k)}$ are known explicitly:

$$\begin{aligned} z_h^{(k)} &:= [\sin kh, \sin 2kh, ..., \sin(n-1)kh]^T, \\ \lambda_h^{(k)} &:= \tfrac{1}{h^2} 4 \sin^2 \tfrac{kh}{2} = \tfrac{2}{h^2}(1 - \cos kh), \qquad k = 1, 2, ..., n-1. \end{aligned} \tag{6.64}$$

This is easily verified by checking $A_h z_h^{(k)} = \lambda_h^{(k)} z_h^{(k)}$, $k = 1, 2, ..., n-1$. The vectors $z_h^{(k)}$ have the Euclidean norm $\|z_h^{(k)}\| = \sqrt{n/2}$ and are orthogonal to each other.

If we consider for fixed k the components $\sin jkh = \sin(jk\pi/n)$ of the eigenvector $z_h^{(k)}$ at the grid points x_j of Ω_h for $j = 1, 2, ..., n-1$, we see that the grid function $z^{(k)} = z_h^{(k)}$ describes a wave on Ω_h with "frequency" k and "wavelength" $2\pi/k$: the number k just gives the number of half-waves on Ω_h.

In order to simplify the notation, we omit the index h occasionally, if it is clear from the context to which grid size h and grid points Ω_h the vectors $u = u_h$, $f = f_h$ and the matrix $A = A_h$ belong.

One motivation for multigrid methods is connected with the convergence behavior of the standard iterative methods, the Jacobi method and the Gauss-seidel method, for solving linear equations $Au = f$. We study this in more detail for the Jacobi method. The usual decomposition

$$A_h = D_h(I - J_h), \quad D_h = \frac{2}{h^2} I,$$

of $A = A_h$ leads to the matrix of order $n-1$

$$J = J_h = I - \frac{h^2}{2} A_h = \frac{1}{2} \begin{bmatrix} 0 & 1 & & 0 \\ 1 & 0 & \ddots & \\ & \ddots & \ddots & 1 \\ 0 & & 1 & 0 \end{bmatrix}$$

and the iteration of the Jacobi method

$$v^{(i+1)} = Jv^{(i)} + \frac{h^2}{2} f.$$

The errors $e^{(i)} := v^{(i)} - u$ of the iterates $v^{(i)}$ then satisfy the recursion

$$e^{(i+1)} = Je^{(i)} = J^{i+1} e^{(0)}.$$

Clearly, the iteration matrix $J = J_h = I - \frac{h^2}{2} A_h$ has the eigenvalues

$$\mu^{(k)} = \mu_h^{(k)} = 1 - \frac{h^2}{2} \lambda_h^{(k)} = \cos kh, \quad k = 1, ..., n-1,$$

but still the same eigenvectors $z^{(k)} = z_h^{(k)}$ as A_h. In order to analyze the behavior of the error e under an iteration step $e \Rightarrow \bar{e} = Je$, we write e as a linear combination of the eigenvectors $z^{(k)} = z_h^{(k)}$ of $J = J_h$:

$$e = \sum_{j=1}^{n-1} \rho_j z^{(j)}.$$

The weight measures the influence of frequency k in e. Because of

$$\bar{e} = Je = \sum_{j=1}^{n-1} \rho_j \mu^{(j)} z^{(j)}$$

and $1 > \mu^{(1)} > \mu^{(2)} > \cdots > \mu^{(n-1)} = -\mu^{(1)} > -1$, we see that all frequencies $k = 1, 2, ..., n-1$ are damped in $\bar{e}$, but to a different extent. The central frequencies $k \approx n/2$ are damped most, the extreme frequencies $k = 1$ and $k = n-1$ only slightly.

The damping of the large frequencies k with $n/2 \leq k \leq n-1$ can be much improved by introducing a suitable relaxation factor ω into the iteration matrix. To this end, we consider a slightly more general decomposition of A defined by: $A = A_h = B - (B - A)$ with $B := (1/\omega)D$, which leads to the damped Jacobi method with the iteration rule

$$v^{(i+1)} = J(\omega)v^{(i)} + \frac{\omega}{2}h^2 f \tag{6.65}$$

in terms of the matrix $J_h(\omega) = J(\omega) := (1-\omega)I + \omega J$. The original Jacobi method corresponds to $\omega = 1$, $J(1) = J$. Clearly, the eigenvalues $\mu^{(k)}(\omega) = \mu_h^{(k)}(\omega)$ of $J(\omega)$ are given by

$$\mu_h^{(k)}(\omega) = \mu^{(k)}(\omega) := 1 - \omega + \omega\mu^{(k)} = 1 - 2\omega \sin^2 \frac{kh}{2}, \;\; k = 1, ..., n-1, \tag{6.66}$$

and they belong to the same eigenvectors $z^{(k)} = z_h^{(k)}$ as before.

Now, an iteration step transforms the error as follows:

$$e = \sum_{k=1}^{n-1} \rho_k z^{(k)} \Rightarrow \bar{e} = J(\omega)e = \sum_{k=1}^{n-1} \rho_k \mu^{(k)}(\omega) z^{(k)}. \tag{6.67}$$

Since $|\mu^{(k)}(\omega)| < 1$ for all $0 < \omega \leq 1$, $k = 1, 2, ..., n-1$, all frequencies k will be damped if $0 < \omega \leq 1$. However, by a suitable choice of ω it is possible to damp the high frequencies $n/2 \leq k \leq n-1$ most heavily. In particular,

$$\max_{n/2 \leq k \leq n-1} |\mu^{(k)}(\omega)|$$

becomes minimal for the choice $\omega = \omega_0 := 2/3$, and then $|\mu^{(k)}(\omega)| \leq 1/3$ for $n/2 \leq k \leq n-1$: the method acts as a "smoother," as the high-frequency oscillations are smoothed out. Note that the damping factor $1/3$ for the high frequencies does not depend on h, but the overall damping factor $\max_k |\mu^{(k)}(\omega)| = \mu^{(1)}(\omega) = 1 - 2\omega \sin^2 \frac{h}{2} = 1 - O(h^2)$ converges to 1 as $h \downarrow 0$, so that the convergence rate of the damped Jacobi method deteriorates as h tends to zero.

A drawback of the damped Jacobi method is that it depends on a parameter ω which has to be chosen properly in order to ensure the damping property. In practice, one prefers

parameter free damping methods. Such a method is the Gauss-seidel method that belongs to the decomposition

$$A_h = D_h - E_h - F_h, \;\; E_h = F_h^T := \frac{1}{h^2}\begin{bmatrix} 0 & 0 & \dots & \dots & 0 \\ 1 & \ddots & & & 0 \\ 0 & \ddots & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & 1 & 0 \end{bmatrix},$$

of A_h. With this method, the new iterate $v^{(i+1)}$ is obtained from $v^{(i)}$ as the solution of the linear equations

$$(A_h - E_h)v^{(i+1)} - F_h v^{(i)} = f_h.$$

One can show that the Gauss-Seidel method has similar damping properties as the damped Jacobi method. Since The damped Jacobi method is easier to analyze we restrict the further discussion to this method.

After relatively few steps of the damped Jacobi method one finds an iterate $v^{(i)} = v_h^{(i)}$ with an error

$$e_h^{(i)} = v_h^{(i)} - u_h = \sum_{j=1}^{n-1} \rho_j^{(i)} z_h^{(j)}$$

containing almost no high frequencies anymore:

$$\max_{n/2 \le k \le n-1} |\rho_k^{(i)}| << \max_{1 \le k < n/2} |\rho_k^{(i)}|.$$

Now, there is a new consideration that comes into play: the vector $e_h^{(i)}$ is the exact solution of the system $A_h e_h = -r_h^{(i)}$, where $r_h^{(i)} = f_h - A_h v_h^{(i)}$ is the residual of $v_h^{(i)}$; hence the decomposition of $r_h^{(i)} = -\sum_{k=1}^{n-1} \rho_k^{(i)} \lambda_h^{(k)} z_h^{(k)}$ essentially contains only contributions of the lower frequencies. But a long wave grid function g_h on Ω_h can be approximated quite well by a grid function g_{2h} on the coarser grid $\Omega_{2h} = \{j \cdot 2h : \; j = 1, 2, ..., (n/2) - 1\}$ (here we assume that n is even) by means of a projection operator I_h^{2h}:

$$g_{2h} := I_h^{2h} g_h, \quad I_h^{2h} := \frac{1}{4}\begin{bmatrix} 1 & 2 & 1 & & & & \\ & & 1 & 2 & 1 & & \\ & & & & \ddots & \ddots & \\ & & & & & 1 & 2 & 1 \end{bmatrix}.$$

Here, I_h^{2h} is an $((n/2) - 1) \times (n - 1)$ matrix. The coarse-grid function g_{2h} on Ω_{2h} is obtained from the fine-grid function g_h on Ω_h by averaging:

$$g_{2h}(j \cdot 2h) = \frac{1}{4} g_h((2j-1) \cdot h) + \frac{2}{4} g_h(2j \cdot h) + \frac{1}{4} g_h((2j+1) \cdot h), \;\; j = 1, ..., \frac{n}{2} - 1.$$

Instead of forming averages, one could also use the simple restriction operator

$$I_h^{2h} := \begin{bmatrix} 0 & 1 & 0 & & & & \\ & & 0 & 1 & 0 & & \\ & & & & \ddots & \ddots & \\ & & & & & 0 & 1 & 0 \end{bmatrix}$$

for the projection. Then the function $g_{2h} = I_h^{2h} g_h$ would be just the restriction of the function g_h on Ω_h to Ω_{2h},

$$g_{2h}(j \cdot 2h) := g_h(2j \cdot h), \;\; j = 1, 2, ..., \frac{n}{2} - 1.$$

We do not pursue this possibility further.

Conversely, interpolation operators I_{2h}^h can be used to extend a grid function g_{2h} on the coarse grid Ω_{2h} to a grid function $g_h = I_{2h}^h g_{2h}$ on the fine grid Ω_h, by defining,

$$I_{2h}^h := \frac{1}{2}\begin{bmatrix} 1 & & & \\ 2 & & & \\ 1 & 1 & & \\ & 2 & & \\ & 1 & \vdots & \\ & & \vdots & 1 \\ & & & 2 \\ & & & 1 \end{bmatrix}.$$

Here I_{2h}^h is an $(n-1) \times ((n/2)-1)$ matrix, and the function g_h is obtained from g_{2h} by interpolation ($g_{2h}(0) = g_{2h}(\pi) := 0$), that is, for $j = 1, 2, ..., n-1$,

$$g_h(jh) := \begin{cases} g_{2h}\left(\frac{j}{2} \cdot 2h\right) & \text{if } j \text{ is even,} \\ \frac{1}{2} g_{2h}\left(\frac{j-1}{2} \cdot 2h\right) + \frac{1}{2} g_{2h}\left(\frac{j+1}{2} \cdot 2h\right) & \text{otherwise} \end{cases}$$

Now an elementary form of a multigrid method runs as follows: A given approximate solution $v_h^{(i)}$ of $A_h u_h = f_h$ is first transformed by a finite number of steps of the damped Jacobi method (6.65) into a new approximate solution $w_h^{(i)}$ of $A_h u_h = f_h$ with error $e_h^{(i)}$ and residual $r_h^{(i)}$. Then the residual $r_h^{(i)}$ is projected to the coarse grid Ω_{2h}: $r_h^{(i)} \to r_{2h}^{(i)} := I_h^{2h} r_h^{(i)}$, and the linear "coarse-grid equation"

$$A_{2h} e_{2h} = -r_{2h}^{(i)}$$

is solved. Its solution $e_{2h}^{(i)}$ is then extended to the fine grid: $e_{2h}^{(i)} \to \tilde{e}_h^{(i)} := I_{2h}^h e_{2h}^{(i)}$ by interpolation. We expect that $\tilde{e}_h^{(i)}$ is a good approximation for the exact solution $e_h^{(i)}$ of $A_h e_h = -r_h^{(i)}$, since $e_h^{(i)}$ is a low-frequency grid function. Therefore, $v_h^{(i+1)} := v_h^{(i)} - \tilde{e}_h^{(i)}$ will presumably be a much better approximation to u_h than $v_h^{(i)}$.

In this way, we obtain the two-grid method, whose basic step $v_h^{(i)} \to v_h^{(i+1)} := TGM(v_h^{(i)})$ is defined by a mapping TGM according to the following rules.

(6.68) **Two-Grid method**

Let v_h be a grid vector on Ω_h.

(1) Perform υ steps of the damped Jacobi method (6.65), with $\omega = \omega_0 := 2/3$ and the starting vector v_h, which results in the vector w_h with the residual $r_h := f_h - A_h w_h$ (smoothing step).

(2) Compute $r_{2h} := I_h^{2h} r_h$ (projecting step).

(3) Solve $A_{2h} e_{2h} = -r_{2h}$ (coarse-grid solution).

(4) Set $\mathrm{TGM}(v_h) := w_h - I_{2h}^h e_{2h}$ (interpolation and fine-grid correction step).

It is relatively easy to analyze the behavior of the error

$$e_h := v_h - u_h \to \bar{e}_h := \bar{v}_h - u_h$$

during one iteration step $v_h \to \bar{v}_h := TGM(v_h)$ of (6.68) in the case of our simple model problem. Because of (6.67), after the υ smoothing steps, the error $d_h := w_h - u_h$ of w_h satisfies

$$d_h = J(\omega_0)^\upsilon e_h, \quad r_h = -A_h d_h = -A_h J(\omega_0)^\upsilon e_h.$$

Further, by (6.68),

$$\begin{aligned} A_{2h}e_{2h} = -r_{2h} &= -I_h^{2h} r_h = I_h^{2h} A_h d_h, \\ \bar{e}_h = \bar{v}_h - u_h &= d_h - I_{2h}^h e_{2h}, \end{aligned}$$

and we find the formula

$$\bar{e}_h = \left(I - I_{2h}^h A_{2h}^{-1} I_h^{2h} A_h\right) d_h = C_h \cdot J(\omega_0)^v e_h \tag{6.69}$$

where C_h is the $(n-1) \times (n-1)$ matrix

$$C_h := I - I_{2h}^h A_{2h}^{-1} I_h^{2h} A_h.$$

In order to study the propagation of the frequencies contained in e_h, we need explicit formulas for the maps $C_h z_h^{(k)}$ of the eigenvectors $z_h^{(k)}$ of A_h. Using the abbreviation $c_k := \cos^2(kh/2)$, $s_k := \sin^2(kh/2)$, and $k' := n - k$, a short direct calculation shows

$$I_h^{2h} z_h^{(k)} = \begin{cases} c_k z_{2h}^{(k)} & \text{if } k = 1, 2, ..., (n/2) - 1, \\ -s_{k'} z_{2h}^{(k')} & \text{for } k = n/2, ..., n-1. \end{cases} \tag{6.70}$$

Here, the vector $z_{2h}^{(k)}$, $1 \le k < n/2$, are just the eigenvectors of A_{2h} for the eigenvalues

$$\lambda_{2h}^{(k)} = \frac{4}{(2h)^2} \sin^2 kh = \frac{1}{h^2} \sin^2 kh$$

[see (6.64)], so that

$$A_{2h}^{-1} z_{2h}^{(k)} = \frac{1}{\lambda_{2h}^{(k)}} z_{2h}^{(k)}, \quad k = 1, 2, ..., \frac{n}{2} - 1.$$

Again, by a simple direct calculation one verifies

$$I_{2h}^h z_{2h}^{(k)} = c_k z_h^{(k)} - s_k z_h^{(k')}, \quad k = 1, 2, ..., \frac{n}{2} - 1. \tag{6.71}$$

Combining these results, one has for $k = 1, 2, ..., \frac{n}{2} - 1$

$$\begin{aligned} I_{2h}^h A_{2h}^{-1} I_h^{2h} A_h z_h^{(k)} &= \lambda_h^{(k)} I_{2h}^h A_{2h}^{-1} I_h^{2h} z_h^{(k)} \\ &= \lambda_h^{(k)} c_k I_{2h}^h A_{2h}^{-1} z_{2h}^{(k)} \\ &= \frac{\lambda_h^{(k)}}{\lambda_{2h}^{(k)}} c_k I_{2h}^h z_{2h}^{(k)} \\ &= \frac{\lambda_h^{(k)}}{\lambda_{2h}^{(k)}} c_k \left(c_k z_h^{(k)} - s_k z_h^{(k')}\right). \end{aligned}$$

Using that

$$\frac{\lambda_h^{(k)}}{\lambda_{2h}^{(k)}} = \frac{\frac{4}{h^2} \sin^2(kh/2)}{\frac{1}{h^2} \sin^2(kh)} = \frac{4s_k}{\sin^2(kh)}, \quad c_k s_k = \frac{1}{4} \sin^2(kh),$$

we finally obtain for $k = 1, 2, ..., \frac{n}{2} - 1$

$$C_h z_h^{(k)} = \left(I - I_{2h}^h A_{2h}^{-1} I_h^{2h} A_h\right) z_h^{(k)} = s_k z_h^{(k)} + s_k z_h^{(k')}. \tag{6.72}$$

Similarly, the maps of the high-frequency vectors $z_h^{(k')}$ are given by

$$C_h z_h^{(k')} = c_k z_h^{(k)} + c_k z_h^{(k')}, \;\; k = 1, 2, ..., \frac{n}{2}. \tag{6.73}$$

We are now able to show the following theorem.

Theorem 6.6.1. *Let $\upsilon = 2$, and $\omega_0 = 2/3$, and suppose that one step of the two-grid method (6.68) transforms the vector v_h into $\bar{v}_h := TGM(v_h)$. Then the errors $e_h := v_h - u_h$, $\bar{e}_h := \bar{v}_h - u_h$ of v_h and $\bar{v}_h$ satisfy*

$$\|\bar{e}_h\|_2 \leq 0.782 \|e_h\|_2.$$

Thus, the two-grid method generates a sequence $v_h^{(j+1)} = TGM(v_h^{(j)})$, $j = 1, 2, ...$, whose errors $e_h^{(j)} = v_h^{(j)} - u_h$ converges to zero with a linear rate of convergence that is independent of h,

$$\|e_h^{(j)}\|_2 \leq (0.782)^j \|e_h^{(0)}\|_2.$$

This is quite remarkable: the convergence rate of the iterative method considered depends on h and deteriorates as $h \downarrow 0$.

Proof. We start with the decomposition of the error $e_h := v_h - u_h$,

$$e_h = \sum_{j=1}^{n-1} \rho_j z_h^{(j)}.$$

We have already seen, in (6.66), that the vectors $z_h^{(k)}$ are also the eigenvectors of $J(\omega_0)$ belonging to the eigenvalues $\mu_h^{(k)}(\omega_0) = 1 - 2\omega_0 s_k$, $k = 1, 2, ..., n-1$. The choice of ω_0 guarantees, for $k = 1, 2, ..., \frac{n}{2}$, $k' := n - k$, that

$$|\mu_h^{(k)}(\omega_0)| < 1, \;\; |\mu_h^{(k')}(\omega_0)| \leq \frac{1}{3}.$$

Next, (6.67), (6.72) and (6.73) imply for $k = 1, 2, ..., \frac{n}{2}$

$$\begin{aligned} C_h J(\omega_0)^\upsilon z_h^{(k)} &= (\mu_h^{(k)}(\omega_0))^\upsilon \left(s_k z_h^{(k)} + s_k z_h^{(k')}\right) := \alpha_k (z_h^{(k)} + z_h^{(k')}), \\ C_h J(\omega_0)^\upsilon z_h^{(k')} &= (\mu_h^{(k')}(\omega_0))^\upsilon \left(c_k z_h^{(k)} + c_k z_h^{(k')}\right) := \beta_k (z_h^{(k)} + z_h^{(k')}), \end{aligned}$$

where the constants α_k and β_k are estimated by

$$|\alpha_k| < s_k \leq \frac{1}{2}, \;\; |\beta_k| \leq \frac{1}{3^\upsilon} \;\text{ for } k = 1, 2, ..., \frac{n}{2}.$$

We thus obtain the following formula for the errors $\bar{e}_h$:

$$\bar{e}_h = C_h J(\omega_0)^\upsilon e_h = \sum_{k=1}^{n/2} \delta_k (\rho_k \alpha_k + \rho_{k'} \beta_k)(z_h^{(k)} + z_h^{(k')}),$$

where we have used the abbreviations $\delta_k := 1$ for $k < n/2$ and $\delta_{n/2} := 1/2$. Finally, the orthogonality of the vectors $z_h^{(k)}$ and $\|z_h^{(k)}\|^2 = n/2$ imply that

$$\begin{aligned}
\|\bar{e}_h\|_2^2 &= n\left[\sum_{k=1}^{n/2}\delta_k(\rho_k^2\alpha_k^2 + \rho_{k'}^2\beta_k^2 + 2\rho_k\rho_{k'}\alpha_k\beta_k)\right] \\
&\leq n\left[\sum_{k=1}^{n/2}\delta_k(\rho_k^2\alpha_k^2 + \rho_{k'}^2\beta_k^2 + (\rho_k^2 + \rho_{k'}^2)|\alpha_k\beta_k|)\right] \\
&\leq n\left(\frac{1}{4} + \frac{1}{2\cdot 3^\upsilon}\right)\sum_{k=1}^{n/2}\delta_k(\rho_k^2 + \rho_{k'}^2) \\
&= \left(\frac{1}{2} + \frac{1}{3^\upsilon}\right)\|e_h\|_2^2.
\end{aligned}$$

For $\upsilon = 2$ we obtain the estimate of the theorem. □

With the two-grid method there remains the problem of how to solve the linear equations $A_{2h}e_{2h} = -r_{2h}$ "on the coarse grid" Ω_{2h} in step (3) of (6.68). Here, the idea suggests itself to use the two-grid method again, thereby reducing this problem to the problem of solving further linear equations on the coarser grid Ω_{4h}, etc.... In this way, we obtain multigrid methods proper. From among the many variants of these methods we only describe the so-called multigrid **V**-cycle, which is an essential ingredient of all such methods. In order to solve $A_h u_h = f_h$ on the grid Ω, the multigrid **V**-cycle visits all grids

$$\Omega_h \to \Omega_{2h} \to \cdots \to \Omega_{2^j h} \to \Omega_{2^{j-1}h} \to \cdots \to \Omega_h$$

between the finest grid Ω_h and the coarsest grid $\Omega_{2^j h}$ in the indicated order: it first descends from the finest to the coarsest grid, and then ascends again to the finest grid, which also explains the name of the method. During one **V**-cycle, an approximate solution v_h of the fine-grid equation $A_h u = f_h$ is replaced by a new approximate solution

$$v_h \leftarrow MV_h(v_h, f_h)$$

of the same equation, where the function $MV_h(v_h, f_h)$ is recursively defined by the following:

(6.74) **Multigrid V-cycle.**

Suppose v_h, f_h are given vectors on Ω_h. Put $H := h$.

(1) By υ steps of the damped Jacobi method (6.65) with $\omega_0 = 2/3$, transform the approximate solution v_H of $A_H u = f_H$ into another approximate solution, again denoted by v_H.

(2) If $H = 2^j h$ goto (4). Otherwise put

$$f_{2H} := I_H^{2H}(f_H - A_H v_H), \quad v_{2H} := MV_{2H}(0, f_{2H}).$$

(3) Compute $v_H := v_H + I_{2H}^H v_{2H}$.

(4) Apply the damped Jacobi method (6.65) υ times with $\omega_0 = 2/3$ to transform the approximate solution v_H of $A_H u = f_H$ into another approximate solution of these equations, again denoted by v_H.

Remark 6.6.1. *The most efficient of these methods require only $O(N)$ operations to compute an approximate solution v_h of a system $A_h u_h = f_h$ with N unknowns, which is sufficiently accurate in the following sense. The error $\|v_h - u_h\| = O(h^2)$ has the same order of magnitude as the truncation error $\max_{x \in \Omega_h} \|y(x) - u_h(x)\| = \tau(h) = O(h^2)$ of the underlying discretization method. Since the exact solution u_h of the discretized equation $A_h u_h = f_h$ differs from the exact solution $y(x)$ of the boundary value problem (6.62) by the truncation error $\tau(h)$ anyway, it makes no sense to compute an approximation v_h to u_h with $\|v_h - u_h\| << \tau(h)$.*

For the simple two-grid method (6.68), Theorem 6.6.1 implies only a weaker result: because of $N = n - 1$ and $h^2 = \pi^2/n^2$, this method requires $j = O(ln\, N)$ iterations to compute an approximate solution $v_h^{(j)}$ of $A_h u_h = f_h$ with $\|v_h^{(j)} - u_h\| = O(h^2)$, if we start with $v_h^{(0)} = 0$. Since the tridiagonal system in step (3) of (6.68) can be solved with $O(N)$ operations, the two-grid method requires altogether $O(N\, ln\, N)$ operations in order to find an approximate solution v_h of acceptable accuracy.

6.7 Comparison of methods

In this section, we determine the respective computational efforts required by the methods discussed in this chapter when applied to the following boundary value problems for partial differential equations:

$$(6.75) \qquad \begin{cases} -u_{xx} - u_{yy} + \gamma x u_x + \gamma y u_y + \delta u = f; & \delta, \gamma \text{ constants} \\ u(x,y) = 0 & \text{for } (x,y) \in \partial\Omega \\ \Omega := \{(x,y) : \ 0 \le x, y \le 1\} \end{cases}$$

on the unit square Ω of $\mathbb{R}^2$.

We approximate problem (6.75) by discretization. We choose a step size $h = \frac{1}{N+1}$, grid points $x_i = ih$, $y_j = jh$ with $i, j = 0, 1, 2, ..., N+1$. We assume that the function u is four times continuously differentiable on Ω, i.e., $u \in \mathbf{C}^4(\Omega)$. Then, by the Taylor expansion of $u(x_i \pm h, y_j \pm h)$ about (x_i, y_j), one obtains by setting: $u_x^{(k)}(x,y) = u_{\underbrace{x...x}_{k+1}}(x,y)$, $u_y^{(k)}(x,y) = u_{\underbrace{y...y}_{k+1}}(x,y)$ and $u(x_l, y_k) = u_{lk}$:

$$(1) \ u(x_{i+1}, y_j) = u(x_i, y_j) + h u_x^{(0)}(x_i, y_j) + \frac{h^2}{2!} u_x^{(1)}(x_i, y_j) + \frac{h^3}{3!} u_x^{(2)}(x_i, y_j) + \frac{h^4}{4!} u_x^{(3)}(x_i + \theta_1^+ .h, y_j)$$

$$(2) \ u(x_{i-1}, y_j) = u(x_i, y_j) - h u_x^{(0)}(x_i, y_j) + \frac{h^2}{2!} u_x^{(1)}(x_i, y_j) - \frac{h^3}{3!} u_x^{(2)}(x_i, y_j) + \frac{h^4}{4!} u_x^{(3)}(x_i + \theta_1^- .h, y_j)$$

$$(3) \ u(x_i, y_{j+1}) = u(x_i, y_j) + h u_y^{(0)}(x_i, y_j) + \frac{h^2}{2!} u_y^{(1)}(x_i, y_j) + \frac{h^3}{3!} u_y^{(2)}(x_i, y_j) + \frac{h^4}{4!} u_y^{(3)}(x_i, y_j + \theta_2^+ .h)$$

$$(4) \ u(x_i, y_{j-1}) = u(x_i, y_j) - h u_y^{(0)}(x_i, y_j) + \frac{h^2}{2!} u_y^{(1)}(x_i, y_j) - \frac{h^3}{3!} u_y^{(2)}(x_i, y_j) + \frac{h^4}{4!} u_y^{(3)}(x_i, y_j + \theta_2^- .h)$$

$$(5) \ u(x_{i+1}, y_j) = u(x_i, y_j) + h u_x^{(0)}(x_i, y_j) + \frac{h^2}{2!} u_x^{(1)}(x_i + \theta_3^+ .h, y_j)$$

$$(6) \ u(x_{i-1}, y_j) = u(x_i, y_j) - h u_x^{(0)}(x_i, y_j) + \frac{h^2}{2!} u_x^{(1)}(x_i + \theta_3^- .h, y_j)$$

$$(7) \ u(x_i, y_{j+1}) = u(x_i, y_j) + h u_y^{(0)}(x_i, y_j) + \frac{h^2}{2!} u_y^{(1)}(x_i, y_j + \theta_4^+ .h)$$

$$(8)\quad u(x_i, y_{j-1}) = u(x_i, y_j) - hu_y^{(0)}(x_i, y_j) + \frac{h^2}{2!}u_y^{(1)}(x_i, y_j + \theta_4^-.h)$$

with $|\theta_k^{\pm}| < 1$.
It follows respectively from $(1)-(2)$, $(3)-(4)$, $(5)-(6)$ and $(7)-(8)$ that

$$\begin{aligned} -\Delta_{xx}u_{ij} &= \frac{-u_{i+1,j} + 2u_{ij} - u_{i-1,j}}{h^2} \\ -\Delta_{yy}u_{ij} &= \frac{-u_{i,j+1} + 2u_{ij} - u_{i,j-1}}{h^2} \\ \Delta_x u_{ij} &= \frac{u_{i+1,j} - u_{i-1,j}}{2h} \\ \Delta_y u_{ij} &= \frac{u_{i,j+1} - u_{i,j-1}}{2h} \end{aligned}$$

When replacing the differential operators u_{xx}, u_{yy}, u_x, u_y at the grid points (x_i, y_j), $i, j = 1, 2, ..., N$ by the corresponding difference quotients: $\Delta_{xx}u_{ij}$, $\Delta_{yy}u_{ij}$, $\Delta_x u_{ij}$, $\Delta_y u_{ij}$, one has:

$$(6.76)\qquad -u_{xx}(x_i, y_j) - u_{yy}(x_i, y_j) \approx \frac{-u_{i+1,j} - u_{i,j+1} + 4u_{ij} - u_{i-1,j} - u_{i,j-1}}{h^2}$$

$$(6.77)\qquad \gamma x_i u_x(x_i, y_j) + \gamma y_j u_y(x_i, y_j) \approx \frac{\gamma x_i u_{i+1,j} + \gamma y_j u_{i,j+1} - \gamma x_i u_{i-1,j} - \gamma y_j u_{i,j-1}}{2h}$$

Since $u \in \mathbf{C}^4(\Omega)$, then $u^{(4)}$ is still continuous over Ω, it follows from this that the truncation error is

$$\begin{aligned} \tau_{ij}(u) &:= u_{xx}(x_i, y_j) + u_{yy}(x_i, y_j) - \Delta_{xx}u_{ij} - \Delta_{yy}u_{ij} \\ &= \frac{h^2}{12}[u_x^{(3)}(x_i + \theta_1.h, y_j) + u_y^{(3)}(x_i, y_j + \theta_2.h)] \text{ for some } |\theta_k| < 1. \end{aligned}$$

The scheme to three points defined by relations $(6.76)-(6.77)$ produces a system of N^2 linear equations

$$(6.78)\qquad Az = b$$

for the vector

$$z := [z_{1,1}, z_{2,1}, ..., z_{N,1}, ..., z_{1,N-1}, ..., z_{N,N-1}, z_{1,N}, ...z_{N,N}]^T$$

of N^2 unknowns z_{ij}: $i, j = 1, 2, ..., N$, by which we approximate $u_{ij} := u(x_i, y_j)$.

The matrix A depends on the choice of γ and δ. For $\gamma = \delta = 0$, it is (up to factor h^2) a positive definite matrix A associated with the model problem:

$$(6.79)\qquad \begin{cases} -u_{xx} - u_{yy} = f \\ u(x, y) = 0 & \text{for } (x, y) \in \partial\Omega \\ \Omega := \{(x, y) :\ 0 \le x, y \le 1\} \end{cases}$$

For $\gamma = 0$ and all δ, the matrix A is still symmetric, but it becomes indefinite if δ is decreased to sufficiently negative values. Finally, for $\gamma \neq 0$, the matrix A is nonsymmetric.

For the group of tests, we use the linear equations (6.78) with a nonsymmetric matrix A (corresponding to choices $\gamma \neq 0$ and $\delta << 0$) in order to compare the Krylov space methods (GMRES, QMR, Bi-CGSTAB) that are able to handle general linear equations.

Next, we consider all the Krylov space methods studied in this chapter. The test results for these methods were obtained with the help of **MATLAB** using, with the exception of the incomplete QGMRES(l) method (6.38), the **MATLAB** functions PCG, GMRES, QMR and BICGSTAB realizing the corresponding Krylov space methods.

a) The matrices involved in the below tests are in general Toeplitz matrices or close to Toeplitz matrices.

b) The associated preconditioners belong to the circulant algebra, Tau algebra or in a Trigonometric algebra. In the case of circulant algebra, the preconditioners are optimal or natural.

c) The reason of the test is to show in general the behavior of preconditioning techniques in the Krylov subspace methods

Since the conjugate gradient method (6.3) is only applicable to positive definite systems, we used also here the linear equations (6.78) belonging to the model problem ($\gamma = \delta = 0$) for numerical tests. The vector $b := Ae$, $e := [1, 1, ..., 1]^T \in \mathbb{R}^{N^2}$ was chosen as right hand side of (6.78), so that $z = e$ is the exact solution of (6.78), and $z^{(0)} := 0$ was chosen as starting point. we describe the test results obtained for $N = 50$ [i.e., the number of unknowns in (6.78) is $N^2 = 2500$] in the form of a table which lists sizes.

$$red_i := \frac{\|Az^{(i)} - b\|_2}{\|Az^{(0)} - b\|_2} \tag{6.80}$$

of the relative residuals versus the iteration number i.

The slow convergence of the unpreconditioned conjugate gradient algorithm is explained by the estimate (6.9) and the relatively large condition number

$$c = cond_2(A) \doteq \frac{4}{\pi^2}(N+1)^2 \approx 1054$$

of the matrix A [see for instance chapter 9, section 9.3: case of Toeplitz or block Toeplitz matrices]. This slow convergence is very much improved by preconditioning (see also chapter 9, subsection 9.7.2). On the other hand, one step of the preconditioned algorithm is more expensive, but not much. The following small table compares the number of floating point operations (flops) per iteration if one uses the SSOR preconditioner:

	no preconditioning	with preconditioning
flops/iteration	$34.5N^2$	$47.5N^2$

The number of iterations (it) and of floating point operations (flops) needed to reduce the size of the initial residual by a factor 10^{-7} is as follows:

(6.81)

	no preconditioning	with preconditioning
it	765	56
$flops$	$26378N^2$	$2662N^2$

In this context belongs a theoretical result of Axelsson (1977, [4]): he shown that the conjugate-gradient method using the SSOR preconditioner (defined in (6.11)) requires $O(N^{2.5}logN)$ operations to find an approximate solution $\bar{z}$ of the linear equations (6.78) for the model problem (6.79) that is sufficiently accurate, $\|\bar{z} - u\|_2 = O(h^2)$.

The behavior of the remaining three Krylov space methods GMRES, QMR, and Bi-CGSTAB described in this chapter is illustrated by similar tables. Since these methods work also for systems of linear equations with nonsymmetric matrices, they were tested using the nonsymmetric system (6.78) resulting from problem (6.75) for the choice $\delta = -100$, $\gamma = 40$, $N = 50$, and the starting point $z^{(0)} = 0$, but only for the preconditioned versions of these algorithms [the SSOR-preconditioner (6.11) with $w = 1$ was used for left-preconditioning].

The arithmetical expenses for the restarted GMRES(25) method (6.37), the incomplete QGMRES(30) method (6.38), the QMR method, and the Bi-CGSTAB method are described

in Table 6.82, where *it* and *flops* denote the number of iterations and operations needed to reduce the initial residual by a factor 10^{-9}:

(6.82)

	GMRES(25)	QGMRES(30)	QMR	Bi-CGSTAB
it	202	179	73	101
flops	$25199N^2$	$37893N^2$	$6843N^2$	$4928N^2$
flops/it	$124.7N^2$	$211.7N^2$	$93.7N^2$	$48.8N^2$

Remark 6.7.1. *The numerical results for the Krylov space methods show the following: The QMR method and Bi-CGSTAB are clearly superior to both GMRES(restart), the restart version* (6.37) *of GMRES, and the incomplete QGMRES(l) method* (6.38)*. If the parameters restart or l are small, each step of these methods is relatively inexpensive, but these methods then are not able to reduce an initial residual substantially within a reasonable number of iterations. For larger values of restart or l, the methods are able to compute accurate solutions, but each iteration then becomes too expensive. In our tests, QMR needed about* 40% *more operations but* 30% *fewer iterations to compute a high accuracy solution than Bi-CGSTAB. The results also confirm that the residuals of the iterates converge more smoothly with QMR than with Bi-CGSTAB, where larger fluctuations are observed.*

Remark 6.7.2. *(Important for preconditioning studied in chapter* 8*)*

- *The test results and Remark* 6.4.4 *show the efficiency of the QMR method for solving the preconditioned collocation linear systems approximating elliptic boundary value problems.*
- *In fact, the collocation matrices are in general neither Toeplitz matrices and nor symmetric matrices (see chapter* 8*). According to Remark* 6.4.4*, for a Hermitian matrix A, the standard Incomplete LU preconditioner* [99] *preserves the sparsity structure of the matrix, that is, the preconditioner matrices have nonzero elements only in those locations where A itself has nonzero elements. For a general non-Hermitian matrix, there is no reason to preserve the sparsity structure of A. Instead, the ILUT(k) variant discards elements subject only to the constraints of fill-in and size, without regard to the sparsity structure of A. However, this does mean that if A is Hermitian, we do not recover the standard Incomplete LU preconditioner. As it is shown in chapter* 8 *the collocation matrices are not symmetric, so the preconditioners can be symmetric. Hence, the above arguments allow us to conclude that the QMR method is a powerful algorithm to solve the preconditioned collocation linear systems (involved in chapter* 8*) with symmetric preconditioners which belong to the Tau algebra.*

Remark 6.7.3. *Some final remarks on the methods considered in this chapter are in order. All Krylov space methods are general purpose methods that may, in principle, be applied to the solution of sparse linear systems $Ax = b$ of an arbitrary origin. By contrast, the ADI methods, Buneman's algorithm and the Fourier methods are methods for solving only special linear equations which result from the discretization of a restricted class of particular boundary value problems for partial differential equations.*

Remark 6.7.4. *Compared to these specialized methods, the convergence of the classical methods is too slow. Because of their particular damping properties, the latter, however, are still being used as part of the "smoothing step" of modern multigrid methods. In general, multigrid methods are the methods of choice if one wishes to solve boundary value problems for partial differential equations by discretization techniques: in rather general situations, multigrid methods will yield a sufficiently accurate solution of the resulting linear equations with a number of operations that grows only linearly or almost linearly with the number of unknowns. In this respect multigrid methods are comparable to the ADI methods and Buneman's algorithm, but may be applied to much more general problems.*

Remark 6.7.5. *The natural realm of application of Krylov space methods is the solution of general sparse linear systems $Ax = b$, in particular those which are not connected with the discretization of partial differential equations. Such system arise e.g. with the treatment of network problems. But Krylov space methods are also used within multigrid methods e.g. for computing "coarse grid solutions". For the efficiency of Krylov space methods, it is important to know and use good preconditioners.*

Conclusion

In this Chapter, we have studied the Krylov space methods and we have provided a general idea of Multigrid methods. The comparison of methods shows the practical importance of Krylov space methods for solving the linear equations $Ax = b$ even with matrices A of large dimension. We didn't do a detailed study of Multigrid methods in this Thesis but, reserve it for future investigations.

Chapter Seven

Regularizing Preconditioning g-Toeplitz Sequences via g-Circulants

Abstract

For a given nonnegative integer g, a matrix A_n of size n is called g-Toeplitz if its entries obey the rule $A_n = [a_{r-gs}]_{r,s=0}^{n-1}$. Analogously, a matrix A_n again of size n is called g-circulant if $A_n = \left[a_{(r-gs)\,mod\,n}\right]_{r,s=0}^{n-1}$. In a recent work we studied the asymptotic properties, in term of spectral distribution, of both g-circulant and g-Toeplitz sequences in the case where $\{a_k\}_k$ is the sequence of Fourier coefficients of a function $f \in L^1(-\pi,\pi)$. Here we are interested in the preconditioning problem 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 nontrivial result is that the preconditioned sequence $\{\mathcal{P}_n\}_n = \{P_n^{-1}A_n\}_n$, where $\{P_n\}_n$ is the sequence of preconditioner, cannot be clustered at 1 so that the case of $g = 1$ is exceptional. However, while a standard preconditioning cannot be achieved, the result has a positive implication since there exist choices of g-circulant sequences which can be used as basic regularizing preconditioning sequences for the corresponding g-Toeplitz structures. Generalizations to the block and multilevel case are also considered, where g is a vector with nonnegative integer entries. Few numerical experiments, related to specific applications in signal and image restoration, are presented and discussed.

Keywords: circulants, Toeplitz, g-circulants, g-Toeplitz, spectral distributions, clustering, preconditioning, multigrid methods.
AMS SC: 65F10, 15A18.

7.1 Introduction

A matrix A_n of size n is called g-Toeplitz if its entries obey the rules $A_n = [a_{r-gs}]_{r,s=0}^{n-1}$, where g is a given nonnegative integer. A matrix A_n of size n is called g-circulant if $A_n = [a_{(r-gs)mod\,n}]_{r,s=0}^{n-1}$: for an introduction and for the algebraic properties of such matrices refer to section 5.1 of the classical book by Davis [51], while new additional results can be found in [161] and references therein. On the other hand, such structured matrices are of interest in many fields such as e.g. multigrid methods [78, 57], wavelet analysis [50], and subdivision algorithms or, equivalently, in the associated refinement equations, see [58] and references therein. Furthermore, it is interesting to remind that Gilbert Strang [150] has shown rich connections between dilation equations in the wavelets context and multigrid algorithms [78, 162], when constructing the restriction/prolongation operators [1] with various boundary conditions. It is worth noticing that the use of different boundary conditions is quite natural when dealing with signal/image restoration problems or differential equations, see [129, 126].

In a recent paper [104] we addressed the problem of characterizing the singular values 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 $\{a_k\}_k$, 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, multilevel 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 $p \times q$.

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 $\{\mathcal{P}_n\}_n$ cannot be clustered at 1 so that the case of $g = 1$ is exceptional and, by the way, widely studied in the literature (see e.g. [36, 38] for the one-level case, [123] for the multilevel case, and [124] for the multilevel block case). However, while the optimal preconditioning cannot be achieved, the result has a positive implication since there exists choices of g-circulant sequences which are regularizing preconditioning sequences for the corresponding g-Toeplitz structures. Generalizations to the block and multilevel cases are also considered.

The paper is organized as follows. In section 7.2 we introduce useful definitions and well-known results concerning the notion of spectral distribution, while section 7.3 is devoted to some preparatory and general results on preconditioning and clustering. In section 7.4 we report distribution results on g-circulants and g-Toeplitz sequences. Section 7.5 is devoted to the preconditioning analysis both in the standard and regularizing senses, while in section 7.6 we discuss the generalization of the results when we deal with the multilevel block case. The aim of section 7.7 is to draw conclusions and to indicate the future lines of research.

7.2 General definitions and tools from spectral distribution theory

For any function F defined on $\mathbb{R}_0^+$ and for any $m \times n$ matrix A, the symbol $\sum_\sigma(F, A)$ stands for the mean

$$\sum\nolimits_\sigma(F, A) := \frac{1}{\min\{m,n\}} \sum_{j=1}^{\min\{m,n\}} F(\sigma_j(A)) = \frac{1}{\min\{m,n\}} \sum_{\sigma \in Sval(A)} F(\sigma) \tag{7.1}$$

Throughout this chapter we speak also about matrix sequences $\{A_k\}_k$ 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 work, we will not need the extra parameter k and we will consider simply matrix sequences of the form $\{A_n\}_n$.

Concerning the case of matrix-sequences, an important notion is that of spectral distribution in eigenvalue or singular value sense, linking the collective behavior of the eigenvalues or singular values of all the matrices in the sequence to a given function (or to a measure). The notion 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 [16] where many classical results by the authors, Szegö, Avram, Parter, Widom Tyrtyshnikov, and many other can be found, and more recent results in [14, 15, 71, 93, 146, 167, 156, 157]). Here we treat the notion of spectral distribution only in the singular value sense since our analysis is devoted to singular values: regarding eigenvalues the analysis, both in the preconditioned and even non-preconditioned case, is substantially trickier given the inherent non-normality of the involved structures.

Definition 7.2.1. *Let $\mathcal{C}_0(\mathbb{R}_0^+)$ be the set of continuous functions with bounded support defined over the nonnegative real numbers, d a positive integer, and θ a complex-valued measurable function defined on a set $G \subset \mathbb{R}^d$ of finite and positive Lebesgue measure $\mu(G)$. Here G will be often equal to $(-\pi, \pi)^d$ so that $e^{i\overline{G}} = \mathbb{T}^d$ with $\mathbb{T}$ denoting the complex unit circle. A*

matrix sequence $\{A_k\}_k$ *is said to be distributed (in the sense of the singular values) as the pair* (θ, G) *or to have the distribution function* θ *(*$\{A_k\}_k \sim_\sigma (\theta, G)$*), if,* $\forall F \in \mathcal{C}_0(\mathbb{R}_0^+)$*, the following limit relation holds*

$$\lim_{k\to\infty} \Sigma_\sigma(F, A_k) = \frac{1}{\mu(G)} \int_G F(|\theta(t)|)dt, \quad t = (t_1, \ldots, t_d). \tag{7.2}$$

When considering θ *taking values in* $\mathcal{M}_{pq}$*, where* $\mathcal{M}_{pq}$ *is the space of* $p \times q$ *matrices with complex entries and a function is considered to be measurable if and only if the component functions are, we say that* $\{A_k\}_k \sim_\sigma (\theta, G)$ *when for every* $F \in \mathcal{C}_0(\mathbb{R}_0^+)$ *we have*

$$\lim_{k\to\infty} \Sigma_\sigma(F, A_k) = \frac{1}{\mu(G)} \int_G \frac{\sum_{j=1}^{\min\{p,q\}} F(\sigma_j(\theta(t)))}{\min\{p, q\}} dt, \quad t = (t_1, \ldots, t_d),$$

with $\sigma_j(\theta(t)) = \sqrt{\lambda_j(\theta(t)^\star\theta(t))} = \lambda_j\sqrt{\theta(t)^\star\theta(t)}$*. Finally we say that two sequences* $\{A_k\}_k$ *and* $\{B_k\}_k$ *are equally distributed in the sense of singular values* (σ) *if,* $\forall F \in \mathcal{C}_0(\mathbb{R}_0^+)$*, we have*

$$\lim_{k\to\infty} \left[\sum\nolimits_\sigma(F, B_k) - \sum\nolimits_\sigma(F, A_k)\right] = 0.$$

Definition 7.2.2. [39]. *Consider a matrix sequences* $\{A_n\}_n$*, where* A_n *is of size* d_n*, and a set* M *in the nonnegative real line. Take* $\epsilon > 0$ *and denote by* M_ϵ *the* ϵ*-extension of* M*, i.e., the union of all real* ϵ*-balls encircling* $M's$ *points. For any* n*, let* $\gamma_n(\epsilon)$ *be the number of those singular values of* A_n *not belonging to* M_ϵ*. Then* M *is called a general singular value cluster if* $\forall\epsilon > 0$

$$\lim_{n\to\infty} \frac{\gamma_n(\epsilon)}{d_n} = 0$$

and M *is called a proper singular value cluster if* $\forall\epsilon > 0$

$$\gamma_n(\epsilon) \le c(\epsilon),$$

where $c(\epsilon)$ *is independent of* n*. In the case where* $M = \{p\}$ *then we simply say that* $\{A_n\}_n$ *is clustered at* p *with respect to the singular values.*

Proposition 7.2.1. [125, 128]. *If* $\{A_n\}_n$*,* $\{B_n\}_n$*, and* $\{Q_n\}_n$ *are sequences of matrices of strictly increasing dimensions* $\{d_n\}_n$*, such that* $\{A_n\}_n \sim_\sigma (\theta, G)$*,* $\{B_n\}_n \sim_\sigma (0, G)$ *and* $\|Q_n\| \le M$ *for some nonnegative constant* M *independent of* n*, then*

$$\begin{aligned} \{A_n + B_n\}_n &\sim_\sigma (\theta, G), \\ \{A_nQ_n\}_n &\sim_\sigma (0, G), \\ \{Q_nA_n\}_n &\sim_\sigma (0, G). \end{aligned}$$

7.3 General definitions and tools from preconditioning theory

When preconditioning is 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 such that $\mathcal{P}_n - I$ is clustered at 0, then the difference between the original sequence and the sequence of preconditioners 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 in subsection 7.3.1.

7.3.1 Tools and machineries

In this subsection, first we give some basic definitions and we introduce some general tools for the spectral analysis of matrix sequences. as already mentioned in the previous section, by $\{d_n\}_n$ we denote an increasing sequence of natural numbers.

Definition 7.3.1. *A sequence of matrices $\{X_n\}_n$, with X_n of size d_n, is said to be sparsely vanishing if there exists a nonnegative function $x(s)$ with $\lim_{s\to 0} x(s) = 0$ so that $\forall \epsilon > 0\ \exists N_\epsilon \in \mathbb{N}$ such that $\forall n > N_\epsilon$*

$$\frac{1}{d_n}\sharp\{i : \sigma_i^{(n)} \leq \epsilon\} \leq x(\epsilon),$$

where $\{\sigma_i^{(n)} :\ i = 1, 2, ..., d_n\}$ denotes the complete set of the singular values of X_n.

Moreover $\{X_n\}_n$ is defined as sparsely unbounded if there exists a nonnegative function $x(s)$ with $\lim_{s\to 0} x(s) = 0$ so that $\forall \epsilon > 0\ \exists N_\epsilon \in \mathbb{N}$ such that $\forall n > N_\epsilon$

$$\frac{1}{d_n}\sharp\left\{i : \sigma_i^{(n)} \leq \frac{1}{\epsilon}\right\} \leq x(\epsilon).$$

It is worth stressing that the reason of the previous definition is due to the notion of sparsely vanishing Lebesgue-measurable functions introduced by Tyrtyshnikov as those functions whose set of zeros has zero Lebesgue measure. In fact, a sequence $\{X_n\}_n$ spectrally distributed as a sparsely vanishing function is sparsely vanishing in the sense of Definition 7.3.1 and a sequence of matrices $\{X_n\}_n$ spectrally distributed as a sparsely unbounded function is sparsely unbounded also in the sense of Definition 7.3.1. In proposition 7.3.1 we prove the above statements.

Proposition 7.3.1. *Let $\{A_n\}_n$, $A_n \in \mathbb{C}^{n\times n}$, be a sequence of matrices spectrally distributed as a sparsely vanishing (sparsely unbounded) function f. Then the sequence $\{A_n\}_n$ is sparsely vanishing (sparsely unbounded).*

Proof. First, we consider the case of a sparsely vanishing function f. For any $\epsilon > 0$ define the nonnegative test function

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

Now, since

$$\frac{1}{n}\sum_{i=1}^{n} G_\epsilon\left(\sigma_i^{(n)}\right) = \frac{1}{n}\left[\sum_{i\in\{j:\sigma_j^{(n)}\leq\epsilon\}} 1 + \sum_{i\in\{j:\epsilon<\sigma_j^{(n)}\leq 2\epsilon\}} G_\epsilon\left(\sigma_i^{(n)}\right)\right]$$

it holds that

$$\frac{1}{n}\sharp\{i : \sigma_i^{(n)} \leq \epsilon\} \leq \frac{1}{n}\sum_{i=1}^{n} G_\epsilon\left(\sigma_i^{(n)}\right)$$

Moreover,

$$\begin{aligned} \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n} G_\epsilon\left(\sigma_i^{(n)}\right) &= \frac{1}{m\{K\}}\int_K G_\epsilon(|f(t)|)dt \\ &\leq \frac{1}{m\{K\}} m\{x \in K : |f(x)| \leq 2\epsilon\}. \end{aligned}$$

The thesis follows by recalling that the assumption f sparsely vanishing implies that

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

Now, we consider the case of a sparsely unbounded function f. For any $\epsilon > 0$ define the nonnegative test function

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

Now, since

$$\frac{1}{n}\sum_{i=1}^{n} F_\epsilon\left(\sigma_i^{(n)}\right) = \frac{1}{n}\left[\sum_{i\in\{j:\,\sigma_j^{(n)}\le\frac{1}{2\epsilon}\}} 1 + \sum_{i\in\{j:\,\frac{1}{2\epsilon}<\sigma_j^{(n)}\le\frac{1}{\epsilon}\}} G_\epsilon\left(\sigma_i^{(n)}\right)\right]$$

it holds that

$$\frac{1}{n}\sharp\left\{i : \sigma_i^{(n)} < \frac{1}{\epsilon}\right\} \ge \frac{1}{n}\sum_{i=1}^{n} F_\epsilon\left(\sigma_i^{(n)}\right)$$

Moreover,

$$\begin{aligned} \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n} F_\epsilon\left(\sigma_i^{(n)}\right) &= \frac{1}{m\{K\}}\int_K F_\epsilon(|f(t)|)dt \\ &\ge \frac{1}{m\{K\}} m\left\{x \in K : |f(x)| \le \frac{1}{2\epsilon}\right\}. \end{aligned}$$

The thesis follows by recalling that

$$\frac{1}{n}\sharp\left\{i : \sigma_i^{(n)} \ge \frac{1}{\epsilon}\right\} = 1 - \frac{1}{n}\sharp\left\{i : \sigma_i^{(n)} < \frac{1}{\epsilon}\right\}$$

and that the assumption f sparsely unbounded implies that

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

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 function F_ϵ and G_ϵ : in fact it is enough to take new test functions $\hat{T}_\epsilon = \hat{T}_\epsilon(y)$ that coincides with $T_\epsilon(y)$ if the argument y is nonnegative and coincides with $T_\epsilon(-y)$ 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 7.3.1. *Consider a sequence of matrices $\{X_n\}_n$, X_n of size d_n. The following are equivalent.*

- *The sequence $\{X_n\}_n$ is sparsely unbounded.*

- *There exists a nonnegative function $x(s)$ with $\limsup_{s\to 0} x(s) = 0$ so that $\forall \epsilon > 0\ \exists N_\epsilon \in \mathbb{N}$ such that $\forall n \geq N_\epsilon$ it holds that $X_n = B_n + L_n$, where $\|B_n\|_2 < \frac{1}{\epsilon}$ and $rank(L_n) \leq x(\epsilon) d_n$.*

Proof. The result trivially follows by using the singular value decomposition properties of the involved matrices and the singular values interlacing properties [72]. □

The following technical lemmas will be useful for performing the spectral analysis of preconditioned matrices in section 7.5.

Lemma 7.3.2. *Let $\{X_n\}_n$ and $\{Y_n\}_n$, X_n, Y_n of size d_n, be two sparsely unbounded matrix sequences. Then $\{X_n Y_n\}_n$ is sparsely unbounded.*

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

$$\begin{aligned} X_n &= \hat{B}_n + \hat{L}_n \\ Y_n &= \tilde{B}_n + \tilde{L}_n \end{aligned}$$

where $\forall \hat{\delta} > 0$, $\exists N_{\hat{\delta}} \in \mathbb{N}$ such that $\forall n \geq N_{\hat{\delta}}$, it holds $\|\hat{B}_n\|_2 < \frac{1}{\hat{\delta}}$ and $rank(\hat{L}_n) \leq \hat{x}(\hat{\delta}) d_n$ with $\lim_{\hat{\delta}\to 0} \hat{x}(\hat{\delta}) = 0$ and where $\forall \tilde{\delta} > 0$, $\exists N_{\tilde{\delta}} \in \mathbb{N}$ such that $\forall n \geq N_{\tilde{\delta}}$, it holds $\|\tilde{B}_n\|_2 < \frac{1}{\tilde{\delta}}$ and $rank(\tilde{L}_n) \leq \tilde{x}(\tilde{\delta}) d_n$ with $\lim_{\tilde{\delta}\to 0} \tilde{x}(\tilde{\delta}) = 0$. Therefore, the matrix $X_n Y_n$ can be written as

$$X_n Y_n = B_n + L_n$$

with

$$\begin{aligned} B_n &= \tilde{B}_n \hat{B}_n \\ L_n &= \tilde{L}_n(\hat{B}_n + \hat{L}_n) + \tilde{B}_n \hat{L}_n, \end{aligned}$$

where, for n large enough, we find

$$\begin{aligned} \|B_n\|_2 &< \frac{1}{(\tilde{\delta}\hat{\delta})} \\ rank(L_n) &\leq (\tilde{x}(\tilde{\delta}) + \hat{x}(\hat{\delta})) d_n. \end{aligned}$$

For the arbitrariness of $\hat{\delta}$ and $\tilde{\delta}$ the claimed thesis follows by virtue of Lemma 7.3.1 □

Lemma 7.3.3. *Let $\{X_n\}_n$ be a sequence of invertible matrices, with X_n of size d_n. If the sequence $\{X_n\}_n$ is sparsely vanishing then the sequence $\{X_n^{-1}\}_n$ is sparsely unbounded and vice versa.*

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

Remark 7.3.1. *The assumption of invertibility can be removed by considering the pseudo-inverse of Moore-Penrose* [100, 112] *instead of the usual inverse matrix.*

Lemma 7.3.4. *Let $\{X_n\}_n$ and $\{Y_n\}_n$ be two sparsely vanishing matrix sequences, with X_n, Y_n of size d_n. Then $\{X_n Y_n\}_n$ is sparsely vanishing. The same is true for sparsely unboundary sequences. In addition, the notion sparsely unbounded sequence is also stable under linear combinations: of course this is not true for the notion of sparsely vanishing sequence.*

Proof. The first part trivially follows from Lemma 7.3.2 by recalling Lemma 7.3.3 and Remark 7.3.1. The rest is a simple variation on the theme. □

Lemma 7.3.5. *Let $\{X_n\}_n$ and $\{Y_n\}_n$ be two matrix sequences, with X_n, Y_n of size d_n. Suppose that the sequence $\{X_n\}_n$ is sparsely unbounded and the sequence $\{Y_n\}_n$ is clustered at 0. Then both the sequences $\{X_nY_n\}_n$ and $\{Y_nX_n\}_n$ are clustered at 0.*

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

$$X_n = B_n + L_n$$

where $\|B_n\|_2 < 1/\hat{\epsilon}$ and $rank(L_n) \leq x(\hat{\epsilon})d_n$ with $\lim_{s\to 0} x(s) = 0$ and $\forall \epsilon > 0\ \exists N_\epsilon \in \mathbb{N}$ such that $\forall n \geq N_\epsilon$ we have

$$Y_n = N_n + R_n$$

where $\|N_n\|_2 \leq \epsilon$ and $rank(R_n) \leq y(\epsilon)d_n$ with $\lim_{s\to 0} y(s) = 0$. Now, by splitting the matrices as

$$X_nY_n = \tilde{N}_n + \tilde{R}_n$$

with

$$\begin{aligned} \tilde{N}_n &= B_nN_n \\ \tilde{R}_n &= B_nR_n + L_n(N_n + R_n), \end{aligned}$$

where

$$\begin{aligned} \|\tilde{N}_n\|_2 &< \epsilon/\hat{\epsilon} \\ rank(\tilde{R}_n) &= (x(\hat{\epsilon}) + y(\epsilon))d_n \end{aligned}$$

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

Lemma 7.3.6. *Consider a sequence $\{A_n\}_n$, where A_n is of size d_n. Then the following are equivalent.*

- *There exists a sequence $\{D_n\}_n$ so that $\|A_n - D_n\|_F^2 = o(d_n)$ and $rank(D_n) = o(d_n)$.*
- *There exists a sequence $\{D_n\}_n$ so that $\forall p \in [1, \infty)$ it holds $\|A_n - D_n\|_{S,p}^p = o(d_n)$, $rank(D_n) = o(d_n)$.*
- *There exists a function $x(s)$ such that $\lim_{s\to 0} x(s) = 0$ so that $\forall \epsilon > 0\ \exists N_\epsilon \in \mathbb{N}$ such that $\forall n \geq N_\epsilon$ it holds $A_n = N_n + R_n$, with $\|N_n\|_2 \leq \epsilon$ and $rank(R_n) \leq x(\epsilon)d_n$.*
- *The sequence $\{A_n\}_n$ is clustered at zero (refer Definition 7.2.2).*
- *The sequence $\{A_n\}_n$ is spectrally distributed as the identically null function (refer to Definition 7.2.1).*

Proof. It is a direct check by making a clever use of the singular value decomposition [72]. □

Lemma 7.3.7. *Consider two sequences $\{A_n\}_n$ and $\{B_n\}_n$, where A_n, B_n are of size d_n. If there exists a sequence $\{D_n\}_n$ so that $\|A_n - B_n - D_n\|_F^2 = o(d_n)$ and $rank(D_n) = o(d_n)$, then the sequence $\{A_n - B_n\}_n$ is spectrally distributed as the identically null function (in the sense of Definition 7.2.1) and the sequences $\{A_n\}_n$ and $\{B_n\}_n$ are equally distributed (in the sense of Definition 7.2.1). 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 7.3.6 we get that $\{A_n - B_n\}_n \sim_\sigma 0$. The equal distribution of the sequences $\{A_n\}_n$ and $\{B_n\}_n$ was proved by Tyrtyshnikov [164]. 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 7.3.1. *Let $\{X_n\}_n$ and $\{P_n\}_n$ be two sequences of matrices, with X_n, P_n of size d_n. Let $\{I_n\}_n$ be the sequence of identity matrices of size d_n. Suppose that the sequence $\{X_n\}_n$ is sparsely unbounded, the matrices P_n are all invertible and the sequence $\{P_n^{-1}X_n - I_n\}_n$ is clustered at zero. Then $\{X_n - P_n\}_n \sim_\sigma 0$ and the sequences $\{X_n\}_n$ and $\{P_n\}_n$ are equally distributed. In addition, if the sequence $\{X_n\}_n$ is distributed as a function then the sequence $\{P_n\}_n$ has the same distribution.*

Finally, if $\{X_n - P_n\}_n \sim_\sigma 0$ then $\{P_n^{-1}X_n - I_n\}_n$ is clustered at 0, under the condition that $\{P_n^{-1}\}_n$ is sparsely unbounded that is $\{P_n\}_n$ is sparsely vanishing.

Proof. From the third assumption, by putting $Y_n = X_n - P_n$, we have $\{P_n^{-1}X_n - I_n\}_n = \{P_n^{-1}Y_n\}_n \sim_\sigma 0$. Therefore, by virtue of Lemma 7.3.6, there exists a function $\tilde{x}(s)$ such that $\lim_{s\to 0}\tilde{x}(s) = 0$ so that $\forall \epsilon > 0\ \exists N_\epsilon \in \mathbb{N}$ such that $\forall n \geq N_\epsilon$ we have $P_n^{-1}Y_n = \tilde{N}_n + \tilde{R}_n$, with $\|\tilde{N}_n\|_2 \leq \epsilon/2$ and $rank(\tilde{R}_n) \leq \tilde{x}(\epsilon)d_n$. Consequently an explicit computation implies

$$P_n^{-1}X_n = I_n + \tilde{N}_n + \tilde{R}_n,$$

that is

$$X_n = P_n(I_n + \tilde{N}_n) + P_n\tilde{R}_n$$

and finally

$$P_n - X_n = X_nN_n + R_n,$$

with

$$\begin{aligned} N_n &= (I_n + \tilde{N}_n)^{-1} - I_n, \\ R_n &= -P_n\tilde{R}_n(I_n + \tilde{N}_n)^{-1} \end{aligned}$$

where ($\epsilon < 1$)

$$\begin{aligned} \|N_n\|_2 &\leq \epsilon \\ rank(R_n) &= \tilde{x}(\epsilon)d_n. \end{aligned}$$

Since the sequence $\{X_n\}_n$ is sparsely unbounded we deduce that $\{X_nN_n\}_n \sim_\sigma 0$ by virtue Lemma 7.3.5 and therefore, by Lemmas 7.3.6 and 7.3.7, we deduce that $\{Y_n\}_n = \{X_n - P_n\}_n \sim_\sigma 0$ and that the sequences $\{X_n\}_n$ and $\{P_n\}_n$ are equally distributed. Now, if the sequence $\{X_n\}_n$ is distributed as a function then the definition of equally distributed implies that the sequence $\{P_n\}_n$ 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 7.3.5 implies $\{P_n^{-1}X_n - I_n\}_n \sim_\sigma 0$ if $\{X_n - P_n\}_n \sim_\sigma 0$ and $\{P_n^{-1}\}_n$ is sparsely unbounded (which is the same as $\{P_n\}_n$ is sparsely vanishing given the invertibility of each P_n and thanks to Lemma 7.3.3). □

Remark 7.3.2. *Lemma 7.3.4 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 7.3.5 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.*

Remark 7.3.3. *Theorem 7.3.1 has a "philosophical" meaning. If we think to the matrices P_n as preconditioners then Theorem 7.3.1 states that a good preconditioning sequence $\{P_n\}_n$ inherits from the original sequence $\{X_n\}_n$ the distribution, if any. Moreover if the sequence $\{X_n\}_n$ is sparsely unbounded (sparsely vanishing) then the same is true for the sequence $\{P_n\}_n$.*

Remark 7.3.4. *The sparsely unboundedness assumption of $\{X_n\}_n$ is necessary and cannot be removed as far as we are concerned with Theorem* 7.3.1. *For instance, take $X_n = (n+1)I_n$ and $P_n = nI_n$. Then the sequence $\{P_n^{-1}X_n - I_n\}_{n\in\mathbb{N}^\star} = \{\frac{1}{n}I_n\}_{n\in\mathbb{N}^\star}$ is clustered at* 0, *but $\{X_n - P_n\}_n = \{I_n\}_n$ is not. However $\{X_n\}_n \sim_\sigma \{P_n\}_n$ since they are both distributed as the constant function ∞.*

Theorem 7.3.2. *Let $\{X_n\}_n$, $\{Y_n\}_n$ and $\{P_n\}_n$ be three sequences of matrices, with X_n, Y_n, P_n of size d_n and P_n invertible for any n. Let $\{I_n\}_n$ be the sequence of identity matrices of of size d_n. Suppose that*

1. *The sequence $\{X_n\}_n$ is sparsely vanishing,*

2. *the sequence $\{X_n - Y_n\}_n$ is clustered at* 0,

3. *the sequence $\{P_n^{-1}X_n - I_n\}_n$ is clustered at* 0.

Then the sequence $\{P_n^{-1}Y_n - I_n\}_n$ is clustered at 0.

Proof. The matrices $P_n^{-1}Y_n - I_n$ can clearly be split as

$$P_n^{-1}Y_n - I_n = (P_n^{-1}X_n - I_n) + P_n^{-1}(Y_n - X_n), \tag{7.3}$$

where the sequence $\{P_n^{-1}X_n - I_n\}_n$ is clustered at 0 by virtue of the assumption 3. Moreover the sequence $\{P_n\}_n$ is sparsely vanishing since the sequence $\{X_n\}_n$ is sparsely vanishing (see Remark 7.3.3). Therefore the application of Lemmas 7.3.3 and 7.3.5 proves that the sequence $\{P_n^{-1}(Y_n - X_n)\}_n$ is clustered at 0. As a final statement, by virtue of equation (7.3), the sequence $\{P_n^{-1}Y_n - I_n\}_n$ is expressed as the sum of two matrix sequences that are clustered at 0, so that the proof is concluded. □

7.4 Singular value distribution of g-circulants and g-Toeplitz sequences

Let f be a Lebesgue integrable function defined on $(-\pi,\pi)^d$ and taking values in $\mathcal{M}_{pq}$, for given positive integers p and q. Then, for d-indices $r = (r_1, r_2, ..., r_d)$, $j = (j_1, j_2, ..., j_d)$, $n = (n_1, n_2, ..., n_d)$, $e = (1, 1, ..., 1)$, $\underline{0} = (0, 0, ..., 0)$, the Toeplitz matrix $T_n(f)$ of size $p\hat{n} \times q\hat{n}$, $\hat{n} = n_1.n_2...n_d$, is defined as follows $T_n(f) = [\tilde{f}_{r-j}]_{r,j=\underline{0}}^{n-e}$, where $\tilde{f}_k$ are the Fourier coefficients of f defined by equation

$$\tilde{f}_k = \tilde{f}_{(k_1,...,k_d)} = \frac{1}{(2\pi)^d}\int_{[-\pi,\pi]^d} f(t_1, ..., t_d)e^{-\hat{i}(k_1t_1+...+k_dt_d)}dt_1...dt_d, \quad \hat{i}^2 = -1, \tag{7.4}$$

for integers k_l such that $-\infty < k_l < \infty$ for $1 \le l \le d$. Since f is a matrix-valued function of d variables whose component functions are all integrable, then the $(k_1, k_2, ..., k_d)$-th Fourier coefficient is considered to be the matrix whose (u, v)-th entry is the $(k_1, k_2, ..., k_d)$-th Fourier coefficient of the function $(f(t_1, ..., t_d))_{u,v}$.

According to this multi-index block notation we can define general multi-level block g-Toeplitz and g-circulants. Of course, in this multidimensional setting, g denotes a d-dimensional vector of nonnegative integers that is, $g = (g_1, g_2, ..., g_d)$. In that case $A_n = [a_{r-g\circ s}]_{r,s=\underline{0}}^{n-e}$ where the $\circ$ operation is the componentwise Hadamard product between vector or matrices of the same size. A matrix A_n of size $p\hat{n} \times q\hat{n}$ is called g-circulant if $A_n = [a_{(r-g\circ s) mod\, n}]_{r,s=\underline{0}}^{n-e}$, where

$$(r - g \circ s) mod\, n = ((r_1 - g_1 \cdot s_1) mod\, n_1, (r_2 - g_2 \cdot s_2) mod\, n_2, ..., (r_d - g_d \cdot s_d) mod\, n_d)\,.$$

7.4.1 The singular value distribution result for g-Toeplitz sequences

We consider the general multilevel case, where f is allowed to be both Lebesgue integrable over Q^d and matrix-valued, $Q = (-\pi, \pi)$. We have

$$\{T_{n,g}\}_n \sim_\sigma (\theta_f, Q^d \times [0,1]^d) \tag{7.5}$$

where

$$\theta_f(x,t) = \begin{cases} \sqrt{\widehat{|f|^{(2)}}(x)} & \text{if } t \in [\underline{0}, 1/g], \\ 0 & \text{for } t \in (1/g, e], \end{cases} \tag{7.6}$$

with

$$\widehat{|f|^{(2)}}(x) = \frac{1}{\hat{g}} \sum_{j=\underline{0}}^{g-e} |f|^2 \left(\frac{x + 2\pi j}{g} \right) \tag{7.7}$$

and where all the arguments are modulus 2π and all the operations are intended componentwise, that is $t \in [\underline{0}, \frac{1}{g}]$ means that $t_k \in \left[0, \frac{1}{g_k}\right]$, $k = 1, 2, ..., d$ and $t \in \left(\frac{1}{g}, e\right]$ means that $t_k \in \left(\frac{1}{g_k}, 1\right]$, $k = 1, 2, ..., d$. The writing $\frac{x+2\pi j}{g}$ defines the d-dimensional vector whose k-th component is $\frac{x_k + 2\pi j_k}{g_k}$, $k = 1, 2, ..., d$ and $\hat{g} = g_1 g_2 ... g_d$. Moreover, if the vector g is degenerate namely there exists an index $s \in \{1, 2, ..., d\}$ for which $g_s = 0$ then the function $\sqrt{\widehat{|f|^{(2)}}(x)}$ becomes identically zero so that

$$\{T_{n,g}\}_n \sim_\sigma (0, G)$$

for every admissible set G. For some concrete examples of g-circulant and g-Toeplitz sequences and related spectra, where some of the entries of g vanish, see [103]. Interestingly enough, if g is the vector of all ones that is we are in standard Toeplitz multilevel context, then $T_{n,g} = T_n(f)$, $\sqrt{\widehat{|f|^{(2)}}(x)}$ reduces to $|f(x)|$, and the variable $t \in [0,1]^d$ becomes useless so that

$$\{T_n(f)\}_n \sim_\sigma (f, Q^d \times [0,1]^d)$$

which is the same as the classical Szegö-Tyrtyshnikov-Tilli result [167, 156]

$$\{T_n(f)\}_n \sim_\sigma (f, Q^d).$$

We finally mention that the technique for obtaining formula (7.5), as in Locally Toeplitz setting [155, 130], strongly relies on the notion of approximating class of sequences [125] which was aimed to develop a basic approximation theory, when the spectral distribution of matrix sequences is considered.

7.4.2 The singular value distribution result for g-circulant sequences

Following the analysis in [104], for g fixed vector and n increasing sequence of vectors we do not find a joint distribution. Assuming $\{C_n\}_n \sim_\sigma (h, Q^d)$ with $\{C_n\}_n$ standard sequence of multilevel circulants (that is g-circulants where g is the vectors of all ones), and assuming that the sequence is chosen so that $\gamma_i = (n_i, g_i)$, $i = 1, 2, ..., d$, are d fixed numbers, we find

$$\{C_{n,g}\}_n \sim_\sigma (\eta_h, Q^d \times [0,1]^d) \tag{7.8}$$

where

$$\eta_h(x,t) = \begin{cases} \sqrt{\widehat{|h|^{(3)}}(x)} & \text{if } t \in \left[\underline{0}, \frac{1}{\gamma}\right], \\ 0 & \text{for } t \in \left(\frac{1}{\gamma}, e\right], \end{cases} \tag{7.9}$$

with

$$\widehat{|h|^{(3)}}(x) = \hat{\gamma}\widehat{|h|^{(2)}}(x) = \sum_{j=\underline{0}}^{\gamma-e} |h|^2 \left(\frac{x+2\pi j}{\gamma}\right) \tag{7.10}$$

and $\hat{\gamma} = \gamma_1\gamma_2...\gamma_d$.

7.5 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 and of the analysis of section 7.3. Then we consider the preconditioning in a regularizing context.

7.5.1 Consequences of the distribution results on preconditioning of g-Toeplitz sequences

We study the possibility of a standard preconditioning in the light of the distribution results and of the analysis of section 7.3.

First of all, Theorem 7.3.1 tells one that $\{P_n\}_n$ is a good preconditioning sequence for $\{X_n\}_n$ (that is $\{P_n^{-1}X_n - I_n\}_n \sim_\sigma 0$) if and only if $\{X_n - P_n\}_n \sim_\sigma 0$ and $\{P_n\}_n$ is sparsely vanishing, with the matrices P_n all invertible. The consequences below are of paramount importance:

- The vector g has to be strictly positive; otherwise the original problem $T_{n,g}\mathbf{x} = \mathbf{b}$ is substantially ill-posed since $\{T_{n,g}\}_n \sim_\sigma 0$ and in addition $C_{n,g}$ is singular and indeed $\{C_{n,g}\}_n \sim_\sigma 0$ which violates the crucial condition of Theorem 7.3.1 that $\{P_n\}_n$ is sparsely vanishing with $P_n = C_{n,g}$.

- Even in the case that g is strictly positive, relations (7.5), (7.6) and (7.7) imply that $\{X_n\}_n$ with $X_n = T_{n,g}$ is sparsely vanishing if and only if f is sparsely vanishing and $g_i = 1$ (or more generally $g_i = \pm 1$), $i = 1, 2, ..., d$. In other words, again by Theorem 7.3.1, a good preconditioning can be achieved only in standard case of multilevel Toeplitz sequences and in fact the latter is a case widely studied in the literature [36, 38, 123] (for $d = 1$ also with strong clustering when f is continuous [123], while for $d > 1$ the clustering is necessary weak due to the computational barrier proven in [143]).

- In any case the condition required by Theorem 7.3.1 that the sequences $\{X_n\}_n$ and $\{P_n\}_n$ with $X_n = T_{n,g}$, $P_n = C_{n,g}$, share the same distribution symbol is quite tricky. By comparing (7.5), (7.6), (7.7) and (7.8), (7.9), (7.10) we have to choose $h = f$.

In conclusion, a good preconditioning can be reached only in the standard multilevel Toeplitz setting. However, if we look at the preconditioning in a different sense something can be said.

7.5.2 Regularizing preconditioning

Suppose that $\{X_n\}_n$ is a sequence of matrices and there exists a sequence of subspaces $\{\mathcal{S}_n\}_n$ of dimension $r_n = [cd_n]$, $c \in (0,1)$ for which $\forall \epsilon > 0$, $\exists N_\epsilon$ and

$$\|X_n v\| \leq \|v\|, \quad \forall v \in \mathcal{S}_n, \ \forall n \geq N_\epsilon.$$

This situation naturally arises when $\{X_n\}_n \sim_\sigma (\theta, G)$ with θ vanishing on $\hat{G} \subset G$ with $m(\hat{G})/m(G) = c$, $m(\cdot)$ being the Lebesgue measure and $|\theta| > 0$ almost everywhere in the complement $G - \hat{G}$. Under such conditions we look for a preconditioning $\{J_n\}_n$ already in inverse such that

$$\begin{aligned} \|J_n X_n v\| &\leq \epsilon \|v\|, \quad \forall v \in \mathcal{S}_n, \ \forall n \geq N_\epsilon, \\ \|J_n X_n v - v\| &\leq \epsilon \|v\|, \quad \forall v \in \mathcal{S}_n^\perp, \ \forall n \geq N_\epsilon. \end{aligned}$$

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 subsection 7.5.4.

7.5.3 Some preparatory tools

Since the notations can become quite heavy, for the sake of simplicity and at the beginning, we start with the case $d = p = q = 1$. Several generalizations are given in section 7.6. We observe that also the case of nonpositive g can be taken into consideration and can be reduced to the case of a nonnegative g. In fact, the role of circulants will be played by (-1)-circulant matrices (called also anti-circulants or skew-circulants), [51]: 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.

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 the identity matrix of order t, while the quantities n_g and $\check{g}$ are defined respectively as $n_g = \frac{n}{(n,g)}$ and $\check{g} = \frac{g}{(n,g)}$.

If we denote by C_n the classical circulant matrix (i.e. with $g = 1$) and by $C_{n,g}$ the g-circulant matrix generated by its elements, for generic n and g one immediately finds $C_{n,g} = C_n Z_{n,g}$, where

$$Z_{n,g} = [\delta_{r-gs}]_{r,s=0}^{n-1}, \qquad \delta_k = \begin{cases} 1 & \text{if } k \equiv 0 (mod\, n); \\ 0 & \text{otherwise.} \end{cases} \tag{7.11}$$

The following preparatory results are straightforward. The detailed proofs are reported in [103]; see also [51].

Lemma 7.5.1. [103]. *Let n be any integer greater than 2 such that*

$$Z_{n,g} = [\underbrace{\tilde{Z}_{n,g} | \tilde{Z}_{n,g} | \ldots | \tilde{Z}_{n,g}}_{(n,g)\ times}] \tag{7.12}$$

where $Z_{n,g}$ is the matrix defined in (7.11) *and $\widetilde{Z}_{n,g} \in \mathbb{C}^{n \times n_g}$ is the submatrix of $Z_{n,g}$ obtained by considering only its first n_g columns, that is*

$$\widetilde{Z}_{n,g} = Z_{n,g} \left[\begin{array}{c} I_{n_g} \\ \hline 0 \end{array}\right]. \tag{7.13}$$

Moreover

$$\widetilde{Z}_{n,g} = \widetilde{Z}_{n,(n,g)} Z_{n_g,\check{g}}, \tag{7.14}$$

where $Z_{n_g,\check{g}}$ *is the matrix defined in* (7.11) *of dimension* $n_g \times n_g$. *Therefore*

$$Z_{n_g,\check{g}} = \left[\widehat{\delta}_{r-\check{g}s}\right]_{r,s=0}^{n_g-1}, \ \widehat{\delta}_k = \begin{cases} 1 & \text{if } k \equiv 0\,(mod\, n_g), \\ 0 & \text{otherwise.} \end{cases} \tag{7.15}$$

Finally, if $g \geq n$ *then* $Z_{n,g} = Z_{n,g^o}$, *where* $g^o = g(mod\, n)$ *and* $Z_{n,g}$ *is defined in* (7.11), *so that*

$$C_{n,g} = C_n Z_{n,g} = C_n Z_{n,g^o} = C_{n,g^o}$$

and

$$C_{n,g} = F_n D_n F_n^\star Z_{n,g}, \tag{7.16}$$

$$D_n = diag(\sqrt{n} F_n^\star \underline{a}), \tag{7.17}$$

$$F_n = \frac{1}{\sqrt{n}} \left[e^{-\hat{\imath}\frac{2\pi jk}{n}}\right]_{j,k=0}^{n-1}, \ \textit{Fourier matrix}, \tag{7.18}$$

$$\underline{a} = [a_0, a_1, ..., a_{n-1}]^T, \ \textit{first column of the matrix } A_n.$$

Lemma 7.5.2. [103]. *Let* F_n *be the Fourier matrix of order* n *defined in* (7.18) *and* $\widetilde{Z}_{n,g} \in \mathbb{C}^{n\times n_g}$ *be the matrix represented in* (7.13). *Then*

$$F_n \widetilde{Z}_{n,g} = \frac{1}{\sqrt{(n,g)}} I_{n,g} F_{n_g} Z_{n_g,\check{g}}, \tag{7.19}$$

where $I_{n,g} \in \mathbb{C}^{n\times n_g}$ *and*

$$I_{n,g} = \left.\begin{bmatrix} I_{n_g} \\ \hline I_{n_g} \\ \hline \vdots \\ \hline I_{n_g} \end{bmatrix}\right\} (n,g) \textit{ times}$$

with I_{n_g} *being the identity matrix of size* n_g *and* $Z_{n_g,\check{g}}$ *as in* (7.15). *Therefore* $\widetilde{Z}_{n,g}^T \widetilde{Z}_{n,g} = I_{n_g}$. *Finally if* $\widehat{Z}_{n,g} \in \mathbb{C}^{n\times\mu_g}$, $\mu_g = \lceil \frac{n}{g} \rceil$, *denotes the matrix* $Z_{n,g}$ *by considering only the* μ_g *first columns, then* $1 \leq (n,g) \leq g$, $\mu_g \leq n_g \leq n$, *and*

$$\widetilde{Z}_{n,g}^T \hat{Z}_{n,g} = \left[\frac{I_{\mu_g}}{0}\right]. \tag{7.20}$$

Remark 7.5.1. *In Lemma* 7.5.2, *if* $(n,g) = g$, *we have* $n_g = \frac{n}{(n,g)} = \frac{n}{g}$ *and* $\check{g} = \frac{g}{(n,g)} = 1$; *so the matrix* $Z_{n_g,\check{g}} = Z_{n_g,1}$, *appearing in* (7.19), *is the identity matrix of dimension* $n_g \times n_g$. *The relation* (7.19) *becomes*

$$F_n \widetilde{Z}_{n,g} = \frac{1}{\sqrt{g}} I_{n,g} F_{n_g}.$$

Remark 7.5.2. *If* $(n,g) = 1$, *Lemma* 7.5.2 *is trivial, because* $n_g = \frac{n}{(n,g)} = n$, $\check{g} = \frac{g}{(n,g)} = g$, *and so* $\widetilde{Z}_{n,g} = Z_{n,g}$. *The relation* (7.19) *becomes*

$$\begin{aligned} F_n \widetilde{Z}_{n,g} &= I_{n,g} F_{n_g} Z_{n_g,\check{g}} \\ &= F_n Z_{n,g}, \end{aligned}$$

since the matrix $I_{n,g}$ *reduces by its definition to the identity matrix of order* n.

Remark 7.5.3. *Lemma* 7.5.2 *is true also if, instead of* F_n *and* F_{n_g}, *we put* $F_n^\star$ *and* $F_{n_g}^\star$, *respectively, because* $F_n^\star = \overline{F}_n$. *In fact there is no transposition, but only conjugation.*

7.5.4 The analysis of regularizing preconditioners when $p = q = d = 1$ and n chosen s.t. $(n, g) = 1$

According to the very concise analysis in subsection 7.5.2, we will prove that a proper choice of the matrix sequence $\{C_{n,g}\}_n$ leads to a satisfactory regularizing preconditioning for $\{T_{n,g}\}_n$, at least the entries of $T_{n,g}$ comes from the Fourier coefficients of a sparsely vanishing function f.

Theorem 7.5.1. *Let $\{T_{n,g}\}_n$ be a sequence of g-Toeplitz matrices generated by a sparsely vanishing function $f \in L^1(Q)$ then the sequence $\{C_{n,g}^{-1}\}_n$, where $\{C_{n,g}\}_n = \{C_n Z_{n,g}\}_n$, C_n is the Frobenius distance minimizer of $T_n(f)$ in the standard circulant algebra and $Z_{n,g}$ defined as in (7.11), is a regularizing preconditioning for $\{T_{n,g}\}_n$.*

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 some sparsely vanishing function f in $L^1(Q)$, with $Q = (-\pi, \pi)$ and by $T_{n,g}$ the g-Toeplitz matrix generated by the same function

$$T_{n,g} = [a_{r-gc}]_{r,c=0}^{n-1}, \tag{7.21}$$

where the quantities $r - gc$ are not reduced modulus n, one verifies immediately for n and g generic that

$$\begin{aligned} T_{n,g} = \left[\widehat{T}_{n,g} | \mathfrak{T}_{n,g}\right] &= \left[T_n \widehat{Z}_{n,g} | \mathfrak{T}_{n,g}\right] \\ &= T_n \left[\ \widehat{Z}_{n,g} \ \middle|\ 0\ \right] + \left[\ 0 \mid \mathfrak{T}_{n,g}\ \right], \end{aligned} \tag{7.22}$$

where $\hat{T}_{n,g} \in \mathbb{C}^{n \times \mu_g}$, $\mu_g = \lceil \frac{n}{g} \rceil$, is the matrix $T_{n,g}$ defined in (7.21) by considering only the μ_g first columns, $\mathfrak{T}_{n,g} \in \mathbb{C}^{n \times (n-\mu_g)}$ is the matrix $T_{n,g}$ defined in (7.21) by considering only the $n - \mu_g$ last columns, and $\hat{Z}_{n,g} \in \mathbb{C}^{n \times \mu_g}$ is the matrix defined by

(7.23)

$$\hat{Z}_{n,g} = [\delta_{r-gs}];\ \ r = 0, 1, ..., n-1;\ \ s = 0, 1, ..., \mu_g - 1, \text{ where } \quad \delta_k = \begin{cases} 1 & \text{if } k \equiv 0 (mod\, n), \\ 0 & \text{otherwise.} \end{cases}$$

(for the proof of relation (7.22) see [104] page 12). Regarding the second addend in (7.22), in [104] section 4.2.2, it was shown that

$$\{[0|\mathfrak{T}_{n,g}]\}_n \ \sim_\sigma \ (0, Q). \tag{7.24}$$

Now we consider the g-circulant matrix

$$C_{n,g} = [a_{(r-gc) mod\, n}]_{r,c=0}^{n-1} = C_n Z_{n,g}, \tag{7.25}$$

where C_n is the classical circulant matrix generated from elements of the first column of $C_{n,g}$ and

$$Z_{n,g} = [\delta_{r-gs}]_{r,s=0}^{n-1}; \qquad \delta_k = \begin{cases} 1 & \text{if } k \equiv 0 (mod\, n), \\ 0 & \text{otherwise.} \end{cases} \tag{7.26}$$

(we observe that $\hat{Z}_{n,g}$ in (7.23) is the matrix $Z_{n,g}$ defined in (7.26) by considering only the μ_g first columns) and we suppose that C_n is nonsingular and $(n, g) = gcd(n, g) = 1$, so that $Z_{n,g}$ is a permutation matrix (see Lemma 7.5.2). More in detail the last item in subsection

7.5.1 prescribes that the symbol h should be chosen to be equal to f, the symbol of the standard Toeplitz sequence. For this reason we choose $\{C_n\}_n$ with C_n the Frobenius distance minimizer of $T_n(f)$ in the standard circulant algebra (the one proposed by Tony Chan in the one-level setting [38]). By the analysis in [131], for $f \in L^1(Q^d)$, we have $\{C_n\}_n \sim_\sigma (f, Q^d)$ so that $\{C_{n,g}\}_n \sim_\sigma (f, Q^d)$ whenever $(n_i, g_i) = 1$, $i = 1, 2, ..., d$, because $Z_{n,g}$ is a permutation matrix (here we are for the moment interested only in the case where $d = 1$). As also observed in subsection 7.5.1, it is necessary to assume that f is sparsely vanishing: this condition is essential for the general clustering results in section 7.3, but it will be also essential in the rest of the proof.

If we consider the product $C_{n,g}^{-1}T_{n,g}$, from (7.22) and since $Z_{n,g}^T Z_{n,g} = I_n$ we have that

$$C_{n,g}^{-1}T_{n,g} = Z_{n,g}^T C_n^{-1} T_n[\hat{Z}_{n,g}|0] + C_{n,g}^{-1}[0|\mathfrak{T}_{n,g}],$$

and, if $\{C_n^{-1}T_n\}_n \sim_\sigma 1$ or more precisely if $\{C_n^{-1}T_n - I_n\}_n \sim_\sigma 0$, i.e.,

$$C_n^{-1}T_n = I_n + E_n, \quad \text{with} \quad \{E_n\}_n \sim_\sigma 0,$$

we obtain

$$\begin{aligned} C_{n,g}^{-1}T_{n,g} &= Z_{n,g}^T C_n^{-1} T_n[\hat{Z}_{n,g}|0] + C_{n,g}^{-1}[0|\mathfrak{T}_{n,g}] \\ &= Z_{n,g}^T [I_n + E_n][\hat{Z}_{n,g}|0] + C_{n,g}^{-1}[0|\mathfrak{T}_{n,g}] \\ &= Z_{n,g}^T [\hat{Z}_{n,g}|0] + Z_{n,g}^T E_n[\hat{Z}_{n,g}|0] + C_{n,g}^{-1}[0|\mathfrak{T}_{n,g}] \\ &= \left[\begin{array}{c|c} I_{\mu_g} & 0 \\ \hline 0 & 0 \end{array}\right] + Z_{n,g}^T E_n[\hat{Z}_{n,g}|0] + C_{n,g}^{-1}[0|\mathfrak{T}_{n,g}]. \end{aligned}$$

Now, from Lemma 4.7 in [104], since $\|Z_{n,g}^T\| = 1$ and $\|[\hat{Z}_{n,g}|0]\| = 1$ (indeed the first is a permutation matrix and the second is an "incomplete" permutation matrix), and since $\{E_n\}_n \sim_\sigma 0$, we infer that $\{Z_{n,g}^T E_n[\hat{Z}_{n,g}|0]\}_n \sim_\sigma 0$. Finally, since $\{C_{n,g}^{-1}\}_n$ is sparsely unbounded (in fact $\{C_n\}_n$, $\{C_{n,g}\}_n \sim_\sigma (f, Q)$ with f sparsely vanishing), we deduce $\{C_{n,g}^{-1}[0|\mathfrak{T}_{n,g}]\}_n \sim_\sigma 0$ and the proof is concluded. □

Remark 7.5.4. *In Theorem* 7.5.1 *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.*

7.6 Generalizations

With all the constraints of subsection 7.5.4, we can allow to have $d > 1$ that is $n = (n_1, n_2, ..., n_d)$ sequence of positive integer vectors with $(n_i, g_i) = 1$, $i = 1, 2, ..., d$, so that $Z_{n,g}$ is still a permutation matrix. The proof reported in subsection 7.5.4 is identical with the only observation that the cluster of $\{C_{n,g}^{-1}T_{n,g} - I_n\}_n$ is weak and not strong, due to the computational barrier proven in [143]. More precisely, under the assumption of positivity and continuity of $|f|$, by using the Korovkin Theory [123] and the Tony Chan preconditioners, we find that the number of outliers of $\{C_{n,g}^{-1}T_{n,g} - I_n\}_n$ grows asymptotically as $\hat{n}\left(\sum_{j=1}^{d} n_j\right)$, $\hat{n} = \prod_{j=1}^{d} n_j$. Moreover the weak clustering can be achieved by using the mild assumption that f is only Lebesgue integrable and sparsely vanishing (see [131]).

Furthermore, by following the approach in [124], 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 $(n_i, g_i) = 1$, $i = 1, 2, ..., d$, 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 $\hat{n}_g$ with $\mu_g \leq n_g < n$, $\mu_g = \lceil \frac{n}{g} \rceil$ (see Lemma 7.5.2, where all the objets $n, g, \mu_g, n_g, (n, g)$ are d-dimensional vectors of positive integers and the inequalities are componentwise). In this case a good preconditioner already in inverse form is

$$J_n = Z_{n,g} C_n^{-1}$$

with C_n the usual Tony Chan preconditioner (refer subsection 7.5.2,). Since $\mu_g \leq n_g < n$ (because $1 < (n, g) \leq g$) by Lemma 7.5.2 we find

$$\widetilde{Z}_{n,g}^T \hat{Z}_{n,g} = \left[\frac{I_{\mu_g}}{0} \right].$$

As a consequence the proof given in subsection 7.5.4 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 $Z_{n,g} C_n^{-1}$ so that $\{C_n^{-1}[0|\mathcal{T}_{n,g}]\}_n \sim_\sigma 0$ because $\{C_n\}_n \sim_\sigma (f, Q^d)$ with f sparsely vanishing and $\{[0|\mathcal{T}_{n,g}]\}_n \sim_\sigma 0$ and finally $\{Z_{n,g}^T C_n^{-1}[0|\mathcal{T}_{n,g}]\}_n \sim_\sigma 0$ because of Proposition 7.2.1, where $Z_{n,g}^T$ plays the role of Q_n and $C_n^{-1}[0|\mathcal{T}_{n,g}]$ plays the role of A_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 Korovkin Theory (see chapter 2). However, any other alternative and successful preconditioner for standard Toeplitz structures can be employed thanks to Theorem 7.3.2, which states a kind of useful transitive property.

7.7 Conclusion

In this paper 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 generalization to the multilevel block setting has been sketched. 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 $\{C_{n,g}\}$. In chapter 9, section 9.9, we have presented and discussed various 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.

Chapter Eight

Preconditioning of Collocation Matrices Approximating Elliptic Boundary Value Problems

Throughout this chapter, we study the preconditioning of collocation matrices approximating elliptic boundary value problems and we provide an asymptotic analysis of spectral radii. First, we present a general idea on the Perron-Frobenius theory and some results on the Weyl-Tyrtyshnikov equal distribution.

8.1 Definitions and results

The purpose of this section is to recall some definitions and main results of linear algebra which are useful in the study of collocation matrices.

8.1.1 Definitions and Perron Frobenius theory

Throughout this subsection, we recall the Perron-Frobenius theory.

Definition 8.1.1. *Let $A = (a_{jk})$ and $B = (b_{jk})$ be two $n \times r$ matrices. Then, $A \geq B$ ($A > B$) if $a_{jk} \geq b_{jk}$ ($a_{jk} > b_{jk}$) for all $j = 1, 2, ..., n$ and $k = 1, 2, ..., r$.*

Definition 8.1.2. *$A \in \mathbb{R}^{n\times r}$ is said to be nonnegative (positive) matrix if $A \geq 0$ ($A > 0$).*

Definition 8.1.3. *Let $B = (b_{ij}) \in \mathbb{C}^{n\times r}$, then $|B|$ denotes the matrix with entries $|b_{ij}|$.*

Definition 8.1.4. *A matrix $A \in \mathbb{R}^{n\times n}$ is said to be reducible if there exists a permutation matrix P such that*

$$C = PAP^T = \begin{bmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{bmatrix}$$

where $A_{11} \in \mathbb{R}^{r\times r}$, $A_{22} \in \mathbb{R}^{(n-r)\times(n-r)}$ and $A_{12} \in \mathbb{R}^{r\times(n-r)}$, $0 < r < n$.

Definition 8.1.5. *A matrix $A \in \mathbb{R}^{n\times n}$ is said to be irreducible if it is not reducible.*

Theorem 8.1.1. [175]. *For every matrix $A \in \mathbb{R}^{n\times n}$ there exists a permutation matrix $P \in \mathbb{R}^{n\times n}$ such that*

$$C = PAP^T = \begin{bmatrix} A_{11} & A_{12} & \dots & A_{1r} \\ 0 & A_{22} & \dots & A_{2r} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & A_{rr} \end{bmatrix}$$

where each block A_{ii} is square matrix and either irreducible or 1×1 null matrix.

Proof. The proof is given by considering the 2×2 block form of a reducible matrix. If A_{11} or A_{22} is reducible, we choose the associated permutation matrix to split it again to its 2×2 block form, and so on. Obviously, if A is irreducible then $r = 1$. The Frobenius normal form is unique, up to a permutation. □

Theorem 8.1.2. [9, 175]. *Let $A \geq 0$ be an irreducible $n \times n$ matrix. Then,*

1. *A has a positive real eigenvalue equal to its spectral radius $\rho(A)$.*

2. *To $\rho(A)$ there corresponds an eigenvector $x > 0$.*

3. *$\rho(A)$ increases when any entry of A increases.*

4. *$\rho(A)$ is a simple eigenvalue of A.*

5. *There is not other nonnegative eigenvector of A different from x.*

Definition 8.1.6. *The associated directed graph, $G(A)$ of an $n \times n$ matrix A, consists of n vertices (nodes) $P_1, P_2, ..., P_n$ where an edge leads from P_i to P_j if and only if $a_{ij} \neq 0$.*

Definition 8.1.7. *A directed graph G is strongly connected if for any ordered pair (P_i, P_j) of vertices of G, there exists a sequence of edges (a path), $(P_i, P_{l_1}), (P_{l_1}, P_{l_2}), (P_{l_2}, P_{l_3}), ..., (P_{l_{r-1}}, P_j)$ which leads from P_i to P_j. We shall say that such a path has length r.*

Theorem 8.1.3. [9, 175]. *An $n \times n$ matrix A is irreducible if and only if $G(A)$ is strongly connected.*

Proof. Let A be an irreducible matrix. Looking for contradiction, suppose that $G(A)$ is not strongly connected. So there exists an ordered pair of nodes (P_i, P_j) for which there is not connection from P_i to P_j. We denote by S_1 the set of nodes that are connected to P_j and by S_2 the set of remaining nodes. Obviously, there is no connection from any node $P_l \in S_2$ to any node of $P_q \in S_1$, since otherwise $P_l \in S_1$ by definition. Both sets are nonempty since $P_j \in S_1$ and $P_i \in S_2$. Suppose that r and $n-r$ are their cardinalities. Consider a permutation transformation $C = PAP^T$ which reorders the nodes of $G(A)$, such that $P_1, P_2, ..., P_r \in S_1$ and $P_{r+1}, P_{r+2}, ..., P_n \in S_2$. This means that $c_{kl} = 0$ for all $k = r+1, r+2, ..., n$ and $l = 1, 2, ..., r$, which constitutes a contradiction since A is irreducible.
Conversely, suppose that A is reducible. Following the above proof in the reverse order we prove that $G(A)$ is not strongly connected. □

Theorem 8.1.4. [9, 175]. *Let $A = [a_{ij}]_{i,j=1}^n \geq 0$ be an irreducible $n \times n$ matrix, and B be $n \times n$ complex matrix with $|B| \leq A$. If β is any eigenvalue of B, then*

$$|\beta| \leq r = \sup\{r_x : x \in \mathbb{C}^n, x \geq 0 \text{ and } x \neq 0\}, \tag{8.1}$$

where

$$r_x = \min_{x_i > 0} \left\{ \frac{\sum_{j=1}^n a_{ij} x_j}{x_i} \right\} \tag{8.2}$$

Moreover, equality is valid in (8.1)*, i.e, $\beta = re^{\hat{i}\phi}$, if and only if $|B| = A$, and where B has the form*

$$B = e^{\hat{i}\phi} DAD^{-1}, \tag{8.3}$$

and D is a diagonal matrix with diagonal entries of modulus unity.

Proof. If $\beta y = By$, $y \neq 0$, then for all $i = 1, 2, ..., n$,

$$\beta y_i = \sum_{j=1}^{n} b_{ij} y_j \Rightarrow |\beta||y_i| = \left| \sum_{j=1}^{n} b_{ij} y_j \right| \leq \sum_{j=1}^{n} |b_{ij}||y_j| \leq \sum_{j=1}^{n} a_{ij}|y_j|$$

this means that

$$|\beta||y| \leq |B||y| \leq |A||y|, \tag{8.4}$$

which implies that $|\beta| \leq r_{|y|} \leq r$, proving (8.1).
If $|\beta| = r$ then $|y|$ is an extremal vector and consequently it is a positive eigenvector of A corresponding to the eigenvalue r. Thus,

$$|\beta||y| = |B||y| = A|y|, \tag{8.5}$$

and since $|y| > 0$ and $|B| \leq A$, we conclude that

$$|B| = A \tag{8.6}$$

□

Theorem 8.1.5. [9, 175]. *If $A = [a_{ij}]_{i,j=1}^{n} \geq 0$ is an irreducible matrix, then either*

$$\sum_{j=1}^{n} a_{ij} = \rho(A) \quad \forall i = 1, 2, ..., n \tag{8.7}$$

or

$$\min_{1 \leq i \leq n} \left(\sum_{j=1}^{n} a_{ij} \right) < \rho(A) < \max_{1 \leq i \leq n} \left(\sum_{j=1}^{n} a_{ij} \right). \tag{8.8}$$

Proof. First suppose that all the row sums are equal. Then, the vector e of all ones is an eigenvector of A : $Ae = \left(\sum_{j=1}^{n} a_{ij} \right) e$. Since $e > 0$, from the Perron Frobenius theorem (see **Theorem** 8.1.10), it follows that $\rho(A) = \sum_{j=1}^{n} a_{ij}$.
If all the row sums are not equal. then, we construct a nonnegative matrix B by decreasing certain positive entries of A, so that for all $k = 1, 2, ..., n$,

$$\sum_{j=1}^{n} b_{kj} = \min_{1 \leq i \leq n} \left(\sum_{j=1}^{n} a_{ij} \right),$$

where $0 \leq B \leq A$ and $B \neq A$. Then, from the Perron Frobenius theorem (see **Theorem** 8.1.11), we get $\rho(B) = \min_{1 \leq i \leq n} \left(\sum_{j=1}^{n} a_{ij} \right) < \rho(A)$. Similarly, we construct an irreducible matrix C by increasing certain positive entries of A, so that for all $k = 1, 2, ..., n$,

$$\sum_{j=1}^{n} c_{kj} = \max_{1 \leq i \leq n} \left(\sum_{j=1}^{n} a_{ij} \right),$$

where $0 \leq A \leq C$ and $C \neq A$. Then, from the Perron Frobenius theorem (**Theorem** 8.1.11), we get $\rho(C) = \max_{1 \leq i \leq n} \left(\sum_{j=1}^{n} a_{ij} \right) > \rho(A)$. □

Theorem 8.1.6. [9; 39]. *Let $A \in \mathbb{R}^{n\times n}$ be a nonnegative matrix and $B \in \mathbb{C}^{n\times n}$ be a complex matrix such that $0 \leq |B| \leq A$. Then*

$$\rho(B) \leq \rho(A) \tag{8.9}$$

Proof. If A is irreducible then the result follows from **Theorem** 8.1.4. If A is reducible, we apply the same permutation transformation P to A and B such that PAP^T be the Frobenius normal form of A. It is obvious that the inequality $0 \leq |B| \leq A$ is invariant under permutation transformation. Then, apply **Theorem** 8.1.4 to submatrices $|B_{ii}|$ and A_{ii} to obtain the result. □

Definition 8.1.8. *A matrix $A \in \mathbb{R}^{n\times n}$ possesses the Perron-Frobenius property if its dominant eigenvalue λ_1 is positive and the corresponding eigenvector $x^{(1)}$ is nonnegative.*

Definition 8.1.9. *A matrix $A \in \mathbb{R}^{n\times n}$ possesses the strong Perron-Frobenius property if its dominant eigenvalue λ_1 is positive, simple ($\lambda_1 > |\lambda_j|, j = 2, 3, ..., n$) and the corresponding eigenvector $x^{(1)}$ is positive.*

Definition 8.1.10. *A matrix $A \in \mathbb{R}^{n\times n}$ is said to be eventually positive (eventually nonnegative) if there exists a positive integer k_0 such that $A^k > 0$ ($A^k \geq 0$) for all $k \geq k_0$.*

Theorem 8.1.7. [101, 105, 152]. *For a symmetric matrix $A \in \mathbb{R}^{n\times n}$, the following properties are equivalent:*

(i) *A possesses the strong Perron-Frobenius property.*

(ii) *A is an eventually positive matrix.*

Proof. $(i \Rightarrow ii)$: $\lambda_1 = \rho(A) > |\lambda_2| \geq \lambda_3| \geq ... \geq \lambda_n|$, where λ_1 is a simple eigenvalue with the eigenvector $x^{(1)} \in \mathbb{R}^n$ being positive. Choose the i-th column $a^{(i)} \in \mathbb{R}^n$ of A.
Expand $a^{(i)}$: $a^{(i)} = \sum_{j=1}^{n} c_j x^{(j)}$ (where $\{x^{(1)}, x^{(2)}, ..., x^{(n)}\}$ is an orthonormal basis of $\mathbb{R}^n$).
$c_j = (a^{(i)}, x^{(j)})$, $j = 1, 2, ..., n$. So, $c_1 = (a^{(i)}, x^{(1)}) = \lambda_1 x_i^{(1)} > 0$. Apply power method: $\lim_{k\to\infty} A^k a^{(i)} > 0 \Rightarrow A^k a^{(i)} > 0 \ \forall k > m$. Choose $m_0 = \min\{m : \ A^k a^{(i)} > 0 \ \forall k \geq m\}$, then, $A^k > 0 \ \forall k \geq k_0 = m_0 + 1$. So, A is an eventually positive matrix.
$(ii \Rightarrow i)$: From the Perron-Frobenius theory of nonnegative matrices, the assumption $A^k > 0$ means that the dominant eigenvalue of A^k is positive and the only one in the circle while the corresponding eigenvector is positive. It is well known that the matrix A has as eigenvalues the k-th roots of those of A^k with the same eigenvectors. Since this happens $\forall k \geq k_0$, A possesses the strong Perron-Frobenius property. □

Theorem 8.1.8. [101, 105, 152]. *For a matrix $A \in \mathbb{R}^{n\times n}$ the following properties are equivalent:*

i. *Both matrices A and A^T possess the strong Perron-Frobenius property.*

ii. *A is an eventually positive matrix.*

iii. *A^T is an eventually positive matrix.*

Proof. $(i \Rightarrow ii)$: Let $A = XDX^{-1}$ be the Jordan canonical form of the matrix A. We assume that the eigenvalue $\lambda_1 = \rho(A)$ is the first diagonal entry of D. So the Jacobi canonical form can be written as

$$A = [x^{(1)}|X_{n,n-1}] \left[\begin{array}{c|c} \lambda_1 & 0 \\ \hline 0 & D_{n-1,n-1} \end{array}\right] \left[\begin{array}{c} y^{(1)^T} \\ \hline Y_{n-1,n} \end{array}\right] \tag{8.10}$$

where $y^{(1)^T}$ and $Y_{n-1,n}$ are the first row and the matrix formed by the last $n-1$ rows of X^{-1}, respectively. Since A possesses the strong Perron-Frobenius property, the eigenvector $x^{(1)}$ is positive. From (8.10), the block form of A^T is

$$A^T = [y^{(1)}|Y_{n,n-1}^T] \left[\begin{array}{c|c} \lambda_1 & 0 \\ \hline 0 & D_{n-1,n-1}^T \end{array}\right] \left[\begin{array}{c} x^{(1)^T} \\ \hline X_{n-1,n}^T \end{array}\right] \tag{8.11}$$

The matrix $D_{n-1,n-1}^T$ is the block diagonal matrix formed by the transpose of the Jordan blocks except λ_1. It is obvious that there exists a permutation matrix $P \in \mathbb{R}^{(n-1)\times(n-1)}$ such that the associated permutation transformation on the matrix $D_{n-1,n-1}^T$ transposes all the Jordan blocks.
Thus, $D_{n-1,n-1} = P^T D_{n-1,n-1}^T P$ and relation (8.11) takes the form:

$$\begin{aligned} A^T &= [y^{(1)}|Y_{n,n-1}^T] \left[\begin{array}{c|c} 1 & 0 \\ \hline 0 & P \end{array}\right] \left[\begin{array}{c|c} 1 & 0 \\ \hline 0 & P^T \end{array}\right] \left[\begin{array}{c|c} \lambda_1 & 0 \\ \hline 0 & D_{n-1,n-1}^T \end{array}\right] \left[\begin{array}{c|c} 1 & 0 \\ \hline 0 & P \end{array}\right] \\ &\times \left[\begin{array}{c|c} 1 & 0 \\ \hline 0 & P^T \end{array}\right] \left[\begin{array}{c} x^{(1)^T} \\ \hline X_{n-1,n}^T \end{array}\right] = [y^{(1)}|Y_{n-1,n}^{\prime T}] \left[\begin{array}{c|c} \lambda_1 & 0 \\ \hline 0 & D_{n-1,n-1} \end{array}\right] \left[\begin{array}{c} x^{(1)^T} \\ \hline X_{n,n-1}^{\prime T} \end{array}\right] \end{aligned}$$

where $Y_{n-1,n}^{\prime T} = Y_{n,n-1}^T P$ and $X_{n,n-1}^{\prime T} = P^T X_{n-1,n}^T$. The last relation is the Jordan canonical form of A^T which means that $y^{(1)}$ is the eigenvector corresponding to the dominant eigenvalue λ_1. Since A^T possesses the strong Perron-Frobenius property, $y^{(1)}$ is a positive vector or a negative one. Since $y^{(1)^T}$ is the first row of X^{-1}, we have that $(y^{(1)}, x^{(1)}) = 1$ implying that $y^{(1)}$ is a positive vector.
We return now to the Jordan canonical form (8.10) of A and form the power A^k.

$$A^k = [x^{(1)}|X_{n,n-1}] \left[\begin{array}{c|c} \lambda_1^k & 0 \\ \hline 0 & D_{n-1,n-1}^k \end{array}\right] \left[\begin{array}{c} y^{(1)^T} \\ \hline Y_{n-1,n} \end{array}\right]$$

then

$$\frac{1}{\lambda_1^k} A^k = [x^{(1)}|X_{n,n-1}] \left[\begin{array}{c|c} 1 & 0 \\ \hline 0 & \frac{1}{\lambda_1^k} D_{n-1,n-1}^k \end{array}\right] \left[\begin{array}{c} y^{(1)^T} \\ \hline Y_{n-1,n} \end{array}\right]$$

Since λ_1 is the dominant eigenvalue, the only one of modulus λ_1, we get that $\lim_{k\to\infty} \frac{1}{\lambda_1^k} D_{n-1,n-1}^k = 0$. Thus

$$\lim_{k\to\infty} \frac{1}{\lambda_1^k} A^k = x^{(1)} y^{(1)^T} > 0.$$

The last relation means that there exists an integer $k_0 > 0$ such that $A^k > 0$ for all $k \geq k_0$. So, A is an eventually positive matrix and the first part of theorem is proved.
$(ii \Leftrightarrow iii)$: Obvious from Definition 8.1.10.
$(ii \Rightarrow i)$: The proof is the same as that of Theorem 8.1.7, by considering that A and A^T are both eventually positive matrices. □

Theorem 8.1.9. [29, 152]. *Let $A \in \mathbb{R}^{n\times n}$ be an eventually nonnegative matrix. Then, both matrices A and A^T possess the Perron-Frobenius property.*

Proof. Analogous to the proof of the part $(ii \Rightarrow i)$ of Theorem 8.1.8. □

Theorem 8.1.10. [101, 105, 152]. *If $A^T \in \mathbb{R}^{n\times n}$ possesses the Perron-Frobenius property, then either*

$$\sum_{j=1}^{n} a_{ij} = \rho(A) \quad \forall i = 1, 2, ..., n, \tag{8.12}$$

or

$$\min_{1\leq i\leq n}\left(\sum_{j=1}^{n}a_{ij}\right) \;\leq\; \rho(A)\leq \max_{1\leq i\leq n}\left(\sum_{j=1}^{n}a_{ij}\right) \tag{8.13}$$

Moreover, if A^T possesses the strong Perron-Frobenius property, then both inequalities in (8.13) *are strict.*

Proof. Let $(\rho(A), y)$ be the Perron-Frobenius eigenpair of the matrix A^T and e be the vector of ones. Then,

$$\begin{aligned} y^TAe &= y^T\begin{pmatrix}\sum_{j=1}^{n}a_{1j}\\ \sum_{j=1}^{n}a_{2j}\\ \vdots\\ \sum_{j=1}^{n}a_{nj}\end{pmatrix} = \sum_{i=1}^{n}\left(y_i\sum_{j=1}^{n}a_{ij}\right)\leq \max_{1\leq i\leq n}\left(\sum_{j=1}^{n}a_{ij}\right)\sum_{i=1}^{n}y_i,\\ y^TAe &= \sum_{i=1}^{n}\left(y_i\sum_{j=1}^{n}a_{ij}\right)\geq \min_{1\leq i\leq n}\left(\sum_{j=1}^{n}a_{ij}\right)\sum_{i=1}^{n}y_i. \end{aligned}$$

On the other hand, we get

$$y^TAe = e^TA^Ty = \rho(A)e^Ty = \rho(A)\sum_{j=1}^{n}y_j.$$

Combining the relations above, we get our result. Obviously, equality holds if the row sums are equal. If A^T possesses the strong Perron-Frobenius property, then $y>0$ and the inequalities become strict. □

Corollary 8.1.1. [101, 105, 152]. *If $A\in\mathbb{R}^{n\times n}$ possesses the Perron-Frobenius property, then either*

$$\sum_{i=1}^{n}a_{ij} \;=\; \rho(A)\quad \forall j=1,2,...,n, \tag{8.14}$$

or

$$\min_{1\leq j\leq n}\left(\sum_{i=1}^{n}a_{ij}\right) \;\leq\; \rho(A)\leq \max_{1\leq j\leq n}\left(\sum_{i=1}^{n}a_{ij}\right). \tag{8.15}$$

Moreover, if A possesses the strong Perron-Frobenius property, then both inequalities in (8.15) *are strict.*

Theorem 8.1.11. [101, 105, 152]. *If the matrices $A, B\in\mathbb{R}^{n\times n}$ are such that $A\leq B$, and both A and B^T possess the Perron-Frobenius property (or both A^T and B possess the Perron-Frobenius property), then*

$$\rho(A) \;\leq\; \rho(B). \tag{8.16}$$

Moreover, if the above matrices possess the strong Perron-Frobenius property and $A\neq B$ then, the inequality (8.16) *becomes strict.*

Proof. Let $x, y \geq 0$ be the Perron right and left eigenvectors of A and B associated with the dominant eigenvalues λ_A and λ_B, respectively. Then the following equalities hold

$$y^T Ax = \lambda_A y^T x, \;\; y^T Bx = \lambda_B y^T x.$$

Since $A \leq B$, $B = A + C$, where $C \geq 0$. So,

$$\lambda_B y^T x = y^T Bx = y^T (A + C)x = y^T Ax + y^T Cx \geq y^T Ax = \lambda_A y^T x.$$

Assuming that $y^T x > 0$, the above relations imply that $\lambda_B \geq \lambda_A$. The case where $y^T x = 0$ is covered by using a continuity argument and perturbation technique. It is also obvious that the inequality becomes strict in the case where the associated Perron-Frobenius properties are strong. □

8.1.2 The Weyl-Tyrtyshnikov equal distribution

This part recalls some definitions on the distribution of matrix sequences. Furthermore, some tools to evaluate the strength of the equal distribution and equal localization that are based upon estimes of the singular values and involve the Frobenius norm. We denote by $\mathcal{M}_s(\mathbb{C})$ the linear space of all the square complex matrices of dimension $s \times s$, and we equippe this linear space by the Frobenius norm defined by:

$$\|A\|_F = \left[\sum_{j=1}^{s} \sigma_j(A)^2\right]^{\frac{1}{2}} = \left[\sum_{i=1}^{s}\sum_{j=1}^{s} |a_{ij}|^2\right]^{\frac{1}{2}}$$

where $A = [a_{ij}]_{i,j=1}^{s} \in \mathcal{M}_s(\mathbb{C})$ and $\sigma_j(A)$ denotes the j-th singular value of A. The first motivation is "practical" in the sense that, in the approximation of matrix sequences of increasing dimension in the simpler space of matrices, this is the only Shatten p-norm whose calculation is computationally not expensive. The second motivation is theoretical: actually the Frobenius norm is the only Shatten p-norm induced by an inner product which makes the space $\mathcal{M}_s(\mathbb{C})$ into a Hilbert space. More specifically, setting $< A, B >=\text{trace}(A^\star B)$, we deduce that $\|A\|_F =< A, A >^{\frac{1}{2}}$.

Definition 8.1.11. *Two real sequences $\{a_i^{(n)}\}_{i \leq d_n}$, $\{b_i^{(n)}\}_{i \leq d_n}$ $(d_n < d_{n+1})$ are equally distributed (ED) if and only if, for any real-valued continuous function F with bounded support, the following relation holds:*

$$\lim_{n \to \infty} \frac{1}{d_n} \sum_{i=1}^{d_n} \left(F(a_i^{(n)}) - F(b_i^{(n)}) \right) \;\; = \;\; 0. \tag{8.17}$$

*When the previous limit goes to zero as $O(d_n^{-1})$ and F is Lipschitz continuous, we say that there is **strong equal distribution** (SED). The same definition applies to the case of sequences of matrices $\{A_n\}_n$ and $\{B_n\}_n$ of dimension $d_n \times d_n$: in this case $\{a_i^{(n)}\}_{i \leq d_n}$ and $\{b_i^{(n)}\}_{i \leq d_n}$ are the sets of their singular values (or eigenvalues if the involved matrices are Hermitian).*

Notation $\{A_n\}_n \simeq_D \{B_n\}_n$ means that the matrix sequences $\{A_n\}_n$ and $\{B_n\}_n$ are equally distributed.

Definition 8.1.12. *Two real sequences $\{a_i^{(n)}\}_{i \leq d_n}$, $\{b_i^{(n)}\}_{i \leq d_n}$ $(d_n < d_{n+1})$ are equally localized (EL) if and only if, for any nontrivial interval $[\alpha, \beta]$ $(\alpha < \beta)$, the following relation holds:*

$$\lim_{n\to\infty}\frac{1}{d_n}\Big(card\{i: a_i^{(n)}\in[\alpha,\beta]\}-card\{i: b_i^{(n)}\in[\alpha,\beta]\}\Big) \;=\; 0. \tag{8.18}$$

When the previous limit goes to zero as $O(d_n^{-1})$, we say that there is ***strong equal localization*** *(SEL). The same definition applies to the case of matrix sequences $\{A_n\}_n$ and $\{B_n\}_n$ of dimension $d_n\times d_n$: in this case $\{a_i^{(n)}\}_{i\le d_n}$ and $\{b_i^{(n)}\}_{i\le d_n}$ are the sets of their singular values (or eigenvalues if the involved matrices are Hermitian).*

Definition 8.1.13. *Two real sequences $\{a_i^{(n)}\}_{i\le d_n}$, $\{b_i^{(n)}\}_{i\le d_n}$ $(d_n<d_{n+1})$ are ϵ-equally localized (ϵ-EL) if and only if, for any $\epsilon>0$, the following relation holds:*

$$\lim_{n\to\infty}\frac{1}{d_n}\Big(card\{i: |a_i^{(n)}-b_i^{(n)}|>\epsilon\}\Big) \;=\; 0. \tag{8.19}$$

When the previous limit goes to zero as $O(d_n^{-1})$, we say that there is ϵ-strong equal localization (ϵ-SEL). The same definition applies to the case of sequences of matrices $\{A_n\}_n$ and $\{B_n\}_n$ of dimension $d_n\times d_n$: in this case $\{a_i^{(n)}\}_{i\le d_n}$ and $\{b_i^{(n)}\}_{i\le d_n}$ are the ordered sets of their singular values (or eigenvalues if the involved matrices are Hermitian).

Definition 8.1.14. *We say that the sequence $\{a_i^{(n)}\}_{i\le d_n}$ is essentially bounded if there exists an interval $M=[\alpha,\beta]$ so that M is a general cluster for it. If M is a proper cluster, then we say that $\{a_i^{(n)}\}_{i\le d_n}$ is properly bounded.*

Definition 8.1.15. *Given a sequence $\{a_i^{(n)}\}_{i\le d_n}$, we say that $p\in\mathbb{R}$ is a sub-cluster point for $\{a_i^{(n)}\}_{i\le d_n}$ if and only if*

$$\lim_{\epsilon\to0}\frac{1}{d_n}\limsup_{n\to\infty}\Big\{card\{i: a_i^{(n)}\in(p-\epsilon,p+\epsilon)\}\Big\} \;=\; c>0. \tag{8.20}$$

A sequence $\{a_i^{(n)}\}_{i\le d_n}$ without sub-cluster point is called regular.

Theorem 8.1.12. [136, 142, 145]. *Let $\{a_i^{(n)}\}_{i\le d_n}$ and $\{b_i^{(n)}\}_{i\le d_n}$ $(d_n<d_{n+1})$ be two ordered real sequences. The following facts hold true.*

1. *SED implies ED, SEL implies EL and ϵ-SEL implies ϵ-EL. These implications cannot be reversed.*
2. *EL implies ED.*
3. *SEL does not imply SED.*
4. *SED does not imply EL.*
5. *ϵ-EL implies ED.*
6. *ϵ-SEL does not imply SED.*
7. *SED does not imply ϵ-EL.*
8. *ϵ-SEL does not imply EL.*
9. *SEL does not imply ϵ-EL.*

Lemma 8.1.1. [136, 142, 145]. *Let $\{A_n\}_n$ and $\{B_n\}_n$ be two sequences of $d_n\times d_n$ matrices.*

1. *Assume that* $rank(A_n - B_n) = o(d_n)$. *Then the sequences* $\{A_n\}_n$ *and* $\{B_n\}_n$ *are equally localized (EL) and equally distributed (ED).*

2. *If* $rank(A_n - B_n) = O(1)$. *Then the sequences* $\{A_n\}_n$ *and* $\{B_n\}_n$ *are strongly equally localized (SEL) and strongly equally distributed (SED).*

Proof. 1. Let $r_n = rank(A_n - B_n)$. As a consequence of the Cauchy interlace theorem (cf. Theorem 1.3.2) we have $\sigma_{i-2r_n}(B_n) \geq \sigma_i(A_n) \geq \sigma_{i+2r_n}(B_n)$ for $i = 2r_n + 1, ..., d_n - 2r_n$. Therefore, for any interval $[\alpha, \beta]$ we have

$$card\{i : \sigma_i(A_n) \in [\alpha, \beta]\} = card\{i : \sigma_i(B_n) \in [\alpha, \beta]\} + e_n \quad |e_n| \leq 4r_n. \tag{8.21}$$

Consequently $r_n = o(d_n)$ and then the sequences $\{A_n\}_n$ and $\{B_n\}_n$ are equally localized (EL). The use of part 2 of Theorem 8.1.12 leads to the equal distribution (ED).
2. If $r_n = O(1)$, then there is SEL by (8.21). For the proof of the last part, recall that F is Lipschitz continuous with bounded support contained in $M = [\alpha, \beta]$. Owing to its Lipschitzness, F is of bounded variation ($F \in BV$) too. Therefore it can be expressed as the sum of two monotone functions. By linearity it is enough to focus our attention on the monotone functions restricted to M. Let $S(A_n)$ and $S(B_n)$ be the sets of the singular values ordered nonincreasingly. Let q be an integer number and let $S(B_n, q)$ be such that $(S(B_n, q))_i = (S(B_n))_{i+q}$, $i = 1, 2, ..., d_n$, where $(S(B_n))_j = \min\{\alpha, (S(B_n))_{d_n}\}$ if $j \geq d_n + 1$ and $(S(B_n))_j = \max\{\beta, (S(B_n))_1\}$ if $j \leq 0$. Now, supposing that $r_n = O(1)$ *i.e*, $r_n \leq k$ for some positive k, we find that $S(B_n, -2k) \geq S(B_n), S(A_n) \geq S(B_n, 2k)$, where " $\geq$ " is intended componentwise. Finally, by monotonicity we deduce that

$$\begin{aligned}
\left|\sum_{i=1}^{d_n} (F(\sigma_i(A_n)) - F(\sigma_i(B_n)))\right| &\leq \left|\sum_{i=1}^{d_n} (F(\sigma_i(S(B_n, -2k))) - F(\sigma_i(S(B_n, 2k))))\right| \\
&= \left|\frac{1}{d_n} \sum_{i=1-2k,...,2k, j=d_n-2k+1,...,d_n+2k} (F(\sigma_i(S(B_n))) - F(\sigma_j(B_n)))\right| \\
&= O(d_n^{-1})
\end{aligned}$$

and the proof is complete. □

Lemma 8.1.2. [136, 142, 145]. *Let* $\{A_n\}_n$ *and* $\{B_n\}_n$ *be two sequences of* $d_n \times d_n$ *matrices.*

1. *If* $\|A_n - B_n\|_F^2 = o(d_n)$ *or* $\|A_n - B_n\|_\infty = o(1)$, *then the sequences* $\{A_n\}_n$ *and* $\{B_n\}_n$ *are* ϵ*-equally localized (*ϵ*-EL) and equally distributed (ED).*

2. *When* $\|A_n - B_n\|_F = O(1)$ *or* $\|A_n - B_n\|_\infty = O(d_n^{-1})$, *then the matrix sequences* $\{A_n\}_n$ *and* $\{B_n\}_n$ *are* ϵ*-strongly equally localized (*ϵ*-SEL).*

3. *If* $\|A_n - B_n\|_1 = O(1)$, *then* $\{A_n\}_n$ *and* $\{B_n\}_n$ *are strongly equally distributed (SED).*

Where $\|A\|_1 = \max_j \sum_{i=1}^{d_n} |a_{ij}|$ *and* $\|A\|_\infty = \max_i \sum_{j=1}^{d_n} |a_{ij}|$, *for any square complex matrix* $A = [a_{ij}]_{i,j=1}^{d_n}$.

Proof. 1. we follow an idea indicated by Tyrtyshnikov in [164] for the case of the Frobenius norm. Let ϵ be a positive arbitrary number and $\gamma_n(\epsilon) = card\{i : |\sigma_i(A_n) - \sigma_i(B_n)| > \epsilon\}$. Or for any pair of matrices $A_n, B_n \in \mathcal{M}_{d_n}(\mathbb{C})$, we have

$$\left[\sum_{j=1}^{d_n}|\sigma_j(A_n)-\sigma_j(B_n)|^2\right]^{\frac{1}{2}} \leq \|A_n-B_n\|_F \tag{8.22}$$

then

$$\left[\sum_{j=1}^{d_n}|\sigma_j(A_n)-\sigma_j(B_n)|^2\right]^{\frac{1}{2}} \leq o(d_n).$$

By definition of $\gamma_n(\epsilon)$ we deduce that

$$o(d_n)=\|A_n-B_n\|_F^2 \geq \sum_{j=1}^{d_n}|\sigma_j(A_n)-\sigma_j(B_n)|^2 \geq \gamma_n(\epsilon)\epsilon^2$$

that is, $\gamma_n(\epsilon)=o(d_n)$. The latter relationship is by definition equivalent to ϵ-EL. In the case of the norm $\|.\|_\infty$ the proof is trivial. Now by the part 5 of Theorem 8.1.12 we deduce the ED property.
2. When $\|A_n-B_n\|_F=O(1)$ or $\|A_n-B_n\|_\infty=O(d_n^{-1})$, then ϵ-SEL property is easily deduced by using the same argument as in the preceding part.
3. Finally, if $\|A_n-B_n\|_1=O(1)$, then the sequences $\{A_n\}_n$ and $\{B_n\}_n$ are strongly equally distributed by the last part of Lemma 8.1.1. □

Theorem 8.1.13. [136, 142, 145]. *Let $\{A_n\}_n$ and $\{B_n\}_n$ be two sequences of $d_n\times d_n$ matrices.*

1. *If $\|A_n-B_n-D_n\|_F^2=o(d_n)$ and $rank(D_n)=o(d_n)$, then the sequences $\{A_n\}_n$ and $\{B_n\}_n$ are equally distributed (ED).*

2. *When $\|A_n-B_n-D_n\|_1=O(1)$, with $rank(D_n)=O(1)$, then $\{A_n\}_n$ and $\{B_n\}_n$ are strongly equally distributed (SED).*

Proof. 1. Let $X_n=B_n+D_n$. Then $\{A_n\}_n$ and $\{X_n\}_n$ fulfill the assumptions of part 1 of Lemma 8.1.2. Therefore $\{A_n\}_n$ and $\{X_n\}_n$ are equally distributed (ED). Moreover, $\{B_n\}_n$ and $\{X_n\}_n$ fulfill the assumptions of part 1 of Lemma 8.1.1 and consequently are equally distributed (ED). Since the ED relation is an equivalence relation, the transitivity yields the claimed result.
2. Let $X_n=B_n+D_n$. Then $\{A_n\}_n$ and $\{X_n\}_n$ are strongly equally distributed (SED) by part 3 of Lemma 8.1.2. Moreover, $\{B_n\}_n$ and $\{X_n\}_n$ fulfill the assumptions of part 2 of Lemma 8.1.1 and consequently are strongly equally distributed (SED). Since the SED relation is an equivalence relation, the proof is concluded by applying the transitivity.
□

We prove the following corollaries with similar tools. In particular, the essentials of the proof of Corollary 8.1.2 can be found in [166].

Corollary 8.1.2. [136, 142, 145, 166]. *Let $\{A_n\}_n$ and $\{B_n\}_n$ be two sequences of $d_n\times d_n$ matrices.*

1. *Suppose that $\|A_n-B_n\|_F^2=o(d_n)$. Then M is a cluster for the sequence $\{A_n\}_n$ if and only if M is a cluster for the sequence $\{B_n\}_n$.*

2. *When $\|A_n-B_n\|_F=O(1)$. Then M is a proper cluster for $\{A_n\}_n$ if and only if M is a proper cluster for $\{B_n\}_n$.*

Proof. 1. Let M be a cluster for $\{A_n\}_n$. Then for any $\epsilon > 0$ we have

$$\gamma_n(A_n, M, \epsilon) = o(d_n), \quad \gamma_n(A_n, M, 2\epsilon) = o(d_n),$$

where the function $\gamma_n(A_n, M, \epsilon)$ measures the cardinality of $I_n(A_n, M, \epsilon)$ being the set of indices j so that $\sigma_j(A_n) \notin M_\epsilon$ (M_ϵ is the ϵ-extension of a set M in the nonnegative real line). Now for any positive $\epsilon^\star$, let $J_n(A_n, B_n, \epsilon^\star)$ be the set of indices j such that $|\sigma_j(A_n) - \sigma_j(B_n)| > \epsilon^\star$. By Lemma 8.1.2, it holds that $\{A_n\}_n$ and $\{B_n\}_n$ are ϵ-EL and consequently $card(J_n(A_n, B_n, \epsilon^\star)) = o(d_n)$ for arbitrary $\epsilon^\star > 0$. For every $i \in U_n(\epsilon, \epsilon^\star) \equiv J_n^c(A_n, B_n, \epsilon^\star) \cap I_n(A_n, M, \epsilon)$ it simultaneously holds that:

$$\sigma_i(A_n) \in M_\epsilon \quad \text{and} \quad |\sigma_i(A_n) - \sigma_i(B_n)| \leq \epsilon^\star.$$

If $\epsilon^\star < \epsilon$ and $i \in U_n(\epsilon, \epsilon^\star)$, by triangle inequality, it follows that $\sigma_i(B_n) \in M_{2\epsilon}$. Finally, recalling that $card(J_n^c(A_n, B_n, \epsilon^\star)) = d_n - o(d_n)$, and $card(I_n^c(A_n, M, \epsilon)) = d_n - o(d_n)$, it is transparent that

$$card(U_n(\epsilon, \epsilon^\star)) = d_n - o(d_n).$$

Since $U_n(\epsilon, \epsilon^\star) \subset \{j : \sigma_j(B_n) \in M_{2\epsilon}\}$ and since ϵ is arbitrary it follows that M is a cluster for $\{B_n\}_n$ and the proof of the first part is concluded.
2. When $\|A_n - B_n\|_F = O(1)$, by following the same argument and by replacing each $o(d_n)$ by $O(1)$, we obtain the desired result. □

Corollary 8.1.3. [136, 142, 145]. *Let $\{A_n\}_n$ and $\{B_n\}_n$ be two sequences of $d_n \times d_n$ matrices and let M be a set of the real line so that for any positive ϵ, the set M_ϵ is made up of a finite union of intervals.*

1. *Suppose that $\|A_n - B_n - D_n\|_F^2 = o(d_n)$, and $rank(D_n) = o(d_n)$. Then M is a cluster for $\{A_n\}_n$ if and only if it is a cluster for $\{B_n\}_n$.*

2. *If $\|A_n - B_n - D_n\|_F = O(1)$ with $rank(D_n) = O(1)$. Then M is a proper cluster for $\{A_n\}_n$ if and only if it is a proper cluster for $\{B_n\}_n$.*

Proof. 1. Let $X_n = B_n + D_n$. Then $\{A_n\}_n$ and $\{X_n\}_n$ have the same clusters by Corollary 8.1.2. But $\{B_n\}_n$ and $\{X_n\}_n$ fulfill the hypotheses of Lemma 8.1.1 so that $\{A_n\}_n$ and $\{B_n\}_n$ are equally localized (EL). Therefore, by definition of equal localization matrix sequences, it follows that for any nontrivial interval $[\alpha, \beta]$ ($\alpha < \beta$), we have

$$card\{i : \sigma_i(A_n) \in [\alpha, \beta]\} = card\{i : \sigma_i(B_n) \in [\alpha, \beta]\} + o(d_n).$$

Since M_ϵ is (for any ϵ) a finite union of nontrivial intervals, the proof is concluded.
2. When $\|A_n - B_n - D_n\|_F = O(1)$ with $rank(D_n) = O(1)$, by following the same argument and by replacing each $o(d_n)$ by $O(1)$, we obtain the desired result. □

Equipped of above results, we can start the study on the preconditioning of collocation matrices approximating elliptic bounded value problems.

8.2 Preconditioning and approximation

Let Ω be an opened domain of model problems (8.23) and (8.24), $\partial\Omega$ its boundary, $\widehat{\Omega} = \partial\Omega \cup \Omega$ an artificial domain greater than real domain Ω and, $\{x_j\}_{j=0}^{n+1}$ and $\{(x_i, y_j)\}_{i,j=0}^{n+1}$ be given or selected points that are chosen out of the real domain Ω and are in the artificial domain. The problems we are interested in, are the best approximations of elliptic boundary value problems defined by:
Uni-dimensional problem

$$\begin{cases} u''(x) = f(x), & x \in \Omega = (0,1) \\ u(0) = a, \;\; u(1) = b \end{cases} \tag{8.23}$$

or **Multi-dimensional problem**

$$\begin{cases} \frac{\partial^2 u}{\partial x^2}(x,y) + \frac{\partial^2 u}{\partial y^2}(x,y) = f(x,y), & (x,y) \in \Omega = (0,1) \times (0,1) \\ u(x,y) = g(x,y), & (x,y) \in \partial\Omega \end{cases} \tag{8.24}$$

by the linear systems of the type: $Ax = y$. The shown method to approximate (8.23) and (8.24) is based on the **Radial Basis Functions.** These types of approximations similar to the others, can be applied for approximating PDEs. In these methods, a radial function is core for approximation space and this space is made by translating a standard radial function with zero as its center (core), to all of the space particles. Here, we present an interesting method using the nodes that most of them are selected out of real domain and the others, in the domain. We study in any case (uni-dimension and multi-dimensions) the preconditioners and the condition numbers of the obtained matrices.
One of the advantages of meshless methods based on the RBFs with respect to another, is high decrease of computational volume that arises when changing multi-dimensions to one dimension. Kansa, [88] is the first researcher that applied an approximation by BRFs (Pseudo interpolation) to the PDEs. The use of the globally supported **RBFs**, reaches to the large linear systems, **poorly condition number,** full matrices as will be shown in the following. A RBFs must be selected experimentally suitable for the model problem. Some of the most commonly used RBFs are:

- Direct Multiquadric: $\phi(t) = (t^2 + c^2)^{\frac{1}{2}}$,
- Inverse Multiquadric: $\phi(t) = (t^2 + c^2)^{-\frac{1}{2}}$,
- Gaussian: $\phi(t) = e^{-\frac{t^2}{c^2}}$,

where c is a shape parameter which determines the "accuracy" and the "stability".

8.2.1 Uni-dimensional problem

$$\begin{cases} u''(x) = f(x), & x \in \Omega = (0,1) \\ u(0) = a, \;\; u(1) = b \end{cases} \tag{8.25}$$

Associated linear system

Let $0 = x_0 < x_1 < ... < x_n < x_{n+1} = 1$ be the subdivision of the interval $(0,1)$. The radial function method consists to set

$$\begin{cases} v(x) = \sum_{j=0}^{n+1} v_j \phi(x - x_j), & x \in \Omega = (0,1) \\ v(0) = a, \;\; v(1) = b \end{cases} \tag{8.26}$$

When differentiae, one obtains from (8.25) − (8.26)

$$\begin{cases} \sum_{j=0}^{n+1} v_j \phi''(x - x_j) = f(x), & x \in \Omega = (0,1) \\ v(0) = a, \;\; v(1) = b \end{cases} \tag{8.27}$$

For all $x = x_0, x_1, ..., x_{n+1}$; system (8.27) becomes

$$\left\{ \begin{array}{l} \sum\limits_{k=0}^{n+1} v_k \phi^{''}(x_j - x_k) = f(x_j), \quad j = 1, ..., n \\ v(0) = a, \;\; v(1) = b \end{array} \right. \tag{8.28}$$

Setting

$$\phi^{''}_{j,k} = \phi^{''}(x_j - x_k), \quad f_j = f(x_j); \tag{8.29}$$

It follows from (8.28) the linear system

$$A_{d_n} v \;\; = \;\; \tilde{f}_{d_n} \tag{8.30}$$

where $d_n = n + 2$ and

$$A_{d_n} = [a_{ij}]_{i,j=0}^{n+1}, \;\; \tilde{f}_{d_n} = [a, f_1, f_2, ..., f_n, b]^T \;\text{ and }\; v = [v_0, v_1, ..., v_{n+1}]^T,$$

with

$$\left\{ \begin{array}{ll} a_{0,j} = \phi_{0,j}, \;\; a_{n+1,j} = \phi_{n+1,j} & j = 0, 1, ..., n+1, \\ a_{i,j} = \phi^{''}_{i,j}, & i = 1, 2, ..., n, \;\; j = 0, 1, ..., n+1. \end{array} \right. \tag{8.31}$$

So, (8.30) is equivalent to

$$\begin{bmatrix} \phi_{0,0} & \phi_{0,1} & \cdots & \phi_{0,n} & \phi_{0,n+1} \\ \phi^{''}_{1,0} & \phi^{''}_{1,1} & \cdots & \phi^{''}_{1,n} & \phi^{''}_{1,n+1} \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ \phi^{''}_{n,0} & \phi^{''}_{n,1} & \cdots & \phi^{''}_{n,n} & \phi^{''}_{n,n+1} \\ \phi_{n+1,0} & \phi_{n+1,1} & & \phi_{n+1.n} & \phi_{n+1,n+1} \end{bmatrix} \begin{bmatrix} v_0 \\ v_1 \\ \vdots \\ v_n \\ v_{n+1} \end{bmatrix} \;\; = \;\; \begin{bmatrix} a \\ f_1 \\ \vdots \\ f_n \\ b \end{bmatrix}. \tag{8.32}$$

Let us notice by T_n the submatrix of A_{d_n} obtained by deleting the first and the last rows and columns, i.e., $T_n = [\phi^{''}_{i,j}]_{i,j=1}^{n}$. Since

$$\phi^{''}(t) \;\; = \;\; \left\{ \begin{array}{ll} \frac{c^2}{(t^2+c^2)^{\frac{3}{2}}} & \text{Multiquadric (MQ)} \\ \frac{2t^2-c^2}{(t^2+c^2)^{\frac{5}{2}}} & \text{Inverse Multiquadric (IMQ)} \\ \frac{2}{c^2}(\frac{2t^2}{c^2} - 1)e^{-\frac{t^2}{c^2}} & \text{Gaussian} \end{array} \right. \tag{8.33}$$

Then $\phi^{''}(t)$ is an even function. Setting $h = (n+1)^{-1}$ and $x_j = jh$ for $j = 0, 1, ..., n+1$, one deduces that T_n is a symmetric Toeplitz matrix. Defining $g = \frac{c}{h}$, it follows from (8.29) and (8.33) that

$$\phi^{''}_{j,k} \;\; = \;\; \left\{ \begin{array}{ll} \frac{1}{h} \frac{g^2}{((j-k)^2+g^2)^{\frac{3}{2}}} & \text{Multiquadric (MQ)} \\ \frac{1}{h^2} \frac{(j-k)^2-2g^2}{((j-k)^2+g^2)^{\frac{5}{2}}} & \text{Inverse Multiquadric (IMQ)} \\ \frac{2}{h^2 g^2} \left(\frac{2(j-k)^2}{g^2} - 1 \right) e^{-\frac{(j-k)^2}{g^2}} & \text{Gaussian} \end{array} \right. \tag{8.34}$$

One deduces from (8.34) that the generating function of the symmetric Toeplitz matrix T_n is given by:

$$s(x) = \sum_{k=-\infty}^{\infty} \phi_k^{''} e^{\hat{i}k\pi x} \qquad x \in (0,1) \tag{8.35}$$

and developing (8.35), one obtains

$$s(x) = \phi_0^{''} + 2\sum_{k=1}^{\infty} \phi_k^{''} \cos(2k\pi x) \qquad x \in (0,1). \tag{8.36}$$

Study of matrices A_{d_n}

1. Determination of the preconditioners of A_{d_n}

When setting $\phi_{j,k}^{''} = \phi_{j-k}^{''}$ and $\phi_{j,k} = \phi_{j-k}$, it follows from (8.30) – (8.32) that the matrix A_{d_n} is given by

$$A_{d_n} = \begin{bmatrix} \phi_0 & \phi_1 & \dots & \phi_n & \phi_{n+1} \\ \phi_1^{''} & \phi_0^{''} & \dots & \phi_{n-1}^{''} & \phi_n^{''} \\ \vdots & \vdots & \dots & \vdots & \vdots \\ \phi_n^{''} & \phi_{n-1}^{''} & \dots & \phi_0^{''} & \phi_1^{''} \\ \phi_{n+1} & \phi_n & & \phi_1 & \phi_0 \end{bmatrix}$$

Let us set: $A_{d_n} = T_{d_n} + \Delta_{d_n}$ where

$$T_{d_n} = \begin{bmatrix} \phi_0^{''} & \phi_1^{''} & \dots & \phi_n^{''} & \phi_{n+1}^{''} \\ \vdots & \vdots & \dots & \vdots & \vdots \\ \phi_n^{''} & \phi_{n-1}^{''} & \dots & \phi_0^{''} & \phi_1^{''} \\ \phi_{n+1}^{''} & \phi_n^{''} & \dots & \phi_1^{''} & \phi_0^{''} \end{bmatrix} \quad \text{and} \quad \Delta_{d_n} = \begin{bmatrix} \phi_0 - \phi_0^{''} & \dots & \phi_{n+1} - \phi_{n+1}^{''} \\ 0 & \dots & 0 \\ \vdots & \dots & \vdots \\ 0 & \dots & 0 \\ \phi_{n+1} - \phi_{n+1}^{''} & \dots & \phi_0 - \phi_0^{''} \end{bmatrix}$$

Since T_{d_n} is a symmetric Toeplitz matrix, then $P_{d_n} = T_{d_n} - H(T_{d_n})$ is a natural preconditioner of T_{d_n} where

$$H(T_{d_n}) = \begin{bmatrix} \phi_2^{''} & \phi_3^{''} & \dots & \phi_{n+1}^{''} & 0 & 0 \\ \phi_3^{''} & & & & 0 & 0 \\ \vdots & & & & & \phi_{n+1}^{''} \\ \phi_{n+1}^{''} & & & & & \vdots \\ 0 & 0 & & & & \phi_3^{''} \\ 0 & 0 & \phi_{n+1}^{''} & \dots & \phi_3^{''} & \phi_2^{''} \end{bmatrix}$$

is the Hankel matrix. Or $rank(A_{d_n} - T_{d_n}) = o(d_n)$, it follows from **Lemma** 8.1.1 that the matrix sequences $\{A_{d_n}\}_n$ and $\{T_{d_n})\}_n$ are equally localized and equally distributed. Then, $P_{d_n} = T_{d_n} - H(T_{d_n})$ are good preconditioners for A_{d_n}.

2. Study of the spectral radius of A_{d_n}

Since $A_{d_n} = T_{d_n} + \Delta_{d_n}$, then

$$\rho(A_{d_n}) = \|A_{d_n}\|_2 \leq \|T_{d_n}\|_2 + \|\Delta_{d_n}\|_2 = \rho(T_{d_n}) + \rho(\Delta_{d_n}) \tag{8.37}$$

In the following we study the ϵ-equal localization and clustering properties of the matrix

sequences $\{A_{d_n}\}_n$ and $\{T_{d_n}\}_n$. First of all, let us determine the eigenvalues of Δ_{d_n}:

$$\begin{aligned}
|\Delta_{d_n} - \lambda I| &= \begin{vmatrix} \phi_0 - \phi_0'' - \lambda & \phi_1 - \phi_1'' & \dots & \phi_{n+1} - \phi_{n+1}'' \\ 0 & -\lambda & \dots & 0 \\ \vdots & \dots & \vdots & \\ 0 & \dots & -\lambda & 0 \\ \phi_{n+1} - \phi_{n+1}'' & \phi_n - \phi_n'' & \dots & \phi_0 - \phi_0'' - \lambda \end{vmatrix} \\
&= (\phi_0 - \phi_0'' - \lambda)^2(-1)^n\lambda^n + (-1)^{n+3}(\phi_{n+1} - \phi_{n+1}'') \\
&\times \begin{vmatrix} \phi_1 - \phi_1'' & \phi_2 - \phi_2'' & \dots & \phi_{n+1} - \phi_{n+1}'' \\ -\lambda & 0 & \dots & 0 \\ \vdots & \dots & \vdots & \\ 0 & \dots & -\lambda & 0 \end{vmatrix} \\
&= (\phi_0 - \phi_0'' - \lambda)^2(-1)^n\lambda^n + (-1)^{n+1}(\phi_{n+1} - \phi_{n+1}'') \\
&\times \begin{bmatrix} \lambda \begin{vmatrix} \phi_2 - \phi_2'' & \phi_3 - \phi_3'' & \dots & \phi_{n+1} - \phi_{n+1}'' \\ -\lambda & 0 & \dots & 0 \\ \vdots & \dots & \vdots & \\ 0 & \dots & -\lambda & 0 \end{vmatrix} \end{bmatrix} \\
&\vdots \\
&= (-1)^n(\phi_0 - \phi_0'' - \lambda)^2\lambda^n + (-1)^{n+1}\lambda^{n-1}(\phi_{n+1} - \phi_{n+1}'') \begin{vmatrix} \phi_n - \phi_n'' & \phi_{n+1} - \phi_{n+1}'' \\ -\lambda & 0 \end{vmatrix} \\
&= (-1)^n(\phi_0 - \phi_0'' - \lambda)^2\lambda^n + (-1)^{n+1}\lambda^n(\phi_{n+1} - \phi_{n+1}'')^2 \\
&= (-1)^n\lambda^n[(\phi_0 - \phi_0'' - \lambda)^2 - (\phi_{n+1} - \phi_{n+1}'')^2] \\
&= (-1)^n\lambda^n(\phi_0 - \phi_0'' - \lambda - \phi_{n+1} + \phi_{n+1}'')(\phi_0 - \phi_0'' - \lambda + \phi_{n+1} - \phi_{n+1}'')
\end{aligned}$$

then

$$|\Delta_{d_n} - \lambda I_{d_n}| = 0 \;\Leftrightarrow\; \begin{cases} \lambda_0 = 0, & \text{mult.=n} \\ \lambda_1 = -(\phi_0'' - \phi_0) + \phi_{n+1}'' - \phi_{n+1}, & \text{mult.=1} \\ \lambda_2 = -(\phi_0'' - \phi_0) - (\phi_{n+1}'' - \phi_{n+1}), & \text{mult.=1} \end{cases}$$

<u>Case 1</u>: Multiquadric

In this case: $\phi_k = h\sqrt{k^2 + g^2}$ and $\phi_k'' = \frac{1}{h}\frac{g^2}{(g^2+k^2)^{\frac{3}{2}}}$, for $k = 0, 1, ..., n+1$. or

$$\phi_0 = hg = c;\ \phi_{n+1} = h\sqrt{(n+1)^2 + g^2} = \sqrt{1 + c^2};$$

$$\phi_0'' = \frac{g^2}{hg^3} = \frac{1}{c};\ \phi_{n+1}'' = \frac{1}{h}\frac{g^2}{(g^2 + (n+1)^2)^{\frac{3}{2}}} = \frac{c^2}{(1+c^2)^{\frac{3}{2}}}$$

So,

$$\begin{cases} \lambda_0 = 0 & \text{mult.=n;} \\ \lambda_1 = \frac{c^2-1}{c} + \frac{c^2-(1+c^2)^2}{(1+c^2)^{\frac{3}{2}}} & \text{mult.=1} \\ \lambda_2 = \frac{c^2-1}{c} - \frac{c^2-(1+c^2)^2}{(1+c^2)^{\frac{3}{2}}} & \text{mult.=1} \end{cases}$$

then

$$\begin{aligned}
\frac{1}{d_n}\|\Delta_{d_n}\|_F^2 &= \frac{1}{d_n}\left[(\frac{c^2-1}{c} - \frac{c^2-(1+c^2)^2}{(1+c^2)^{\frac{3}{2}}})^2 + (\frac{c^2-1}{c} + \frac{c^2-(1+c^2)^2}{(1+c^2)^{\frac{3}{2}}})^2\right] \\
&= \frac{2}{d_n}\left[\frac{(c^2-1)^2}{c^2} + \frac{(1+c^2+c^4)^2}{(1+c^2)^3}\right] \underset{n\to\infty}{\to} 0
\end{aligned}$$

then

$$\frac{1}{\sqrt{d_n}}\|A_{d_n} - T_{d_n}\|_F \underset{n\to\infty}{\to} 0 \tag{8.38}$$

According to (8.38), it follows from **Lemma** 8.1.2 and **Corollary** 8.1.2 that the sequences $\{A_{d_n}\}_n$ and $\{T_{d_n}\}_n$ are ϵ-equally localized and are clustered. So

$$\rho(A_{d_n}) \underset{n\to\infty}{\approx} \rho(T_{d_n}) \tag{8.39}$$

Now, let us study the spectral radius of the Toeplitz matrix T_{d_n}.
Since the coefficients of T_{d_n} are positive, then T_{d_n} is a positive matrix. It follows from **Theorem** 8.1.8 that both the matrices T_{d_n} and $T_{d_n}^T$ possess the strong Perron- Frobenius property, and according to **Theorem** 8.1.10, one deduces that

$$\min_{0\le i\le n+1}\sum_{j=0}^{n+1}\phi''_{i-j} \le \rho(T_{d_n}) \le \max_{0\le i\le n+1}\sum_{j=0}^{n+1}\phi''_{i-j}. \tag{8.40}$$

Or, $\phi''_{i-j} = \frac{1}{h}\frac{g^2}{(g^2+(i-j)^2)^{\frac{3}{2}}}$, then $\min_{0\le i\le n+1}\sum_{j=0}^{n+1}\phi''_{i-j} = \sum_{j=0}^{n+1}\phi''_{n+1-j} = \frac{1}{h}\sum_{j=0}^{n+1}\frac{g^2}{(g^2+j^2)^{\frac{3}{2}}}$. Setting $f_n(x) = \frac{1}{h}\frac{g^2}{(g^2+x^2)^{\frac{3}{2}}}$ $\forall x \in [0, n+1]$, then $f'(x) = \frac{1}{h}\frac{-3xg^2(g^2+x^2)^{\frac{1}{2}}}{(g^2+x^2)^3} \le 0$, then the function f is nonincreasing over the interval $[0, n+1]$. So, $f(n+1) \le f(x)$ $\forall x \in [0, n+1]$, in particular $f(n+1) \le f(j)$ $\forall j = 0, 1, ..., n+1$. Or, $f(n+1) = \frac{c^2}{(1+c^2)^{\frac{3}{2}}}$, then

$$\frac{c^2}{(1+c^2)^{\frac{3}{2}}}(n+2) \le \frac{1}{h}\sum_{j=0}^{n+1}\frac{g^2}{(g^2+j^2)^{\frac{3}{2}}} = \min_{0\le i\le n+1}\sum_{j=0}^{n+1}\phi''_{i-j}. \tag{8.41}$$

Conclusion: From (8.39) − (8.40) − (8.41), one deduces that the spectral radius of A_{d_n} grows as $d_n = n+2$ $\forall c > 0$. So, the matrix A_{d_n} is ill-conditioned.

Case 2: Inverse Multiquadric
In order to simplify the study, we only work in the case where $c \ge \frac{1}{\sqrt{2}}$. Indeed this requirement imposes to the entries of T_{d_n} to be nonpositive. Since,
$\phi_k = \frac{1}{h}\frac{1}{\sqrt{k^2+g^2}}$ and $\phi''_k = \frac{1}{h^2}\frac{k^2-2g^2}{(g^2+k^2)^{\frac{5}{2}}}$, for $k = 0, 1, ..., n+1$, then

$$\phi_0 = \frac{1}{c};\ \phi_{n+1} = \frac{1}{\sqrt{1+c^2}},$$

$$\phi''_0 = \frac{-2}{gc^2};\ \phi''_{n+1} = \frac{(1-2c^2)c}{g(1+c^2)^{\frac{5}{2}}}.$$

So,

$$\begin{cases} \lambda_0 = 0 & \text{mult.=n} \\ \lambda_1 = \frac{1}{c} + \frac{2}{gc^2} + \frac{(1-2c^2)c}{g(1+c^2)^{\frac{5}{2}}} - \frac{1}{(1+c^2)^{\frac{1}{2}}} = \alpha_n(c) & \text{mult.=1} \\ \lambda_2 = \frac{1}{c} + \frac{2}{gc^2} - \frac{(1-2c^2)c}{g(1+c^2)^{\frac{5}{2}}} + \frac{1}{(1+c^2)^{\frac{1}{2}}} = \beta_n(c) & \text{mult.=1} \end{cases}$$

As $\alpha_n(c) = \frac{1}{c} + \frac{2}{c^3(n+1)} + \frac{(1-2c^2)}{(n+1)(1+c^2)^{\frac{5}{2}}} - \frac{1}{(1+c^2)^{\frac{1}{2}}}$ and $\beta_n(c) = \frac{1}{c} + \frac{2}{c^3(n+1)} - \frac{(1-2c^2)}{(n+1)(1+c^2)^{\frac{5}{2}}} + \frac{1}{(1+c^2)^{\frac{1}{2}}}$, then

$$\frac{1}{d_n}\|\Delta_{d_n}\|_F^2 = \frac{1}{d_n}[|\alpha_n(c)|^2 + |\beta_n(c)|^2] \underset{n\to\infty}{\to} 0$$

then

$$\frac{1}{\sqrt{d_n}}\|A_{d_n} - T_{d_n}\|_F \underset{n\to\infty}{\to} 0 \tag{8.42}$$

According to (8.42), it follows from **Lemma** 8.1.2 and **Corollary** 8.1.2 that the sequences $\{A_{d_n}\}_n$ and $\{T_{d_n}\}_n$ are ϵ-equally localized and are clustered. So

$$\rho(A_{d_n}) \underset{n\to\infty}{\approx} \rho(T_{d_n}). \tag{8.43}$$

Now, let us study the spectral radius of the Toeplitz matrix T_{d_n}.
Since $c \geq \frac{1}{\sqrt{2}}$ then the coefficients $\phi_k^{''} = \frac{1}{h^2}\frac{k^2-2g^2}{(g^2+k^2)^{\frac{5}{2}}}$ of T_{d_n} are nonpositive, then $T_{d_n}^2$ is a nonnegative matrix, so $-T_{d_n}$ is an eventually nonnegative matrix. It follows from **Theorem** 8.1.9 that both the matrices $-T_{d_n}$ and $(-T_{d_n})^T$ possess the Perron- Frobenius property, and according to **Theorem** 8.1.10, one deduces that

$$\min_{0\leq i\leq n+1}\sum_{j=0}^{n+1}(-\phi_{i-j}^{''}) \leq \rho(-T_{d_n}) \leq \max_{0\leq i\leq n+1}\sum_{j=0}^{n+1}(-\phi_{i-j}^{''}). \tag{8.44}$$

Because, $-\phi_{i-j}^{''} = \frac{1}{h^2}\frac{2g^2-(i-j)^2}{(g^2+(i-j)^2)^{\frac{5}{2}}}$, setting $g_n(x) = \frac{1}{h^2}\frac{2g^2-x^2}{(g^2+x^2)^{\frac{5}{2}}}$ $\forall x \in [0, n+1]$, then $g_n^{'}(x) = \frac{1}{h^2}\frac{3x(g^2+x^2)^{\frac{3}{2}}(x^2-4g^2)}{(g^2+x^2)^5} \leq 0$, then the function g_n is nonincreasing over the interval $[0, n+1]$, since $g_n^{'}(x) = 0 \Leftrightarrow x = 0$, one deduces that g_n is a decreasing function over $[0, n+1]$. Furthermore, for $j = 1, 2, ..., n$

$$-\phi_{(j-1)-j}^{''} = \frac{1}{h^2}\frac{2g^2-1}{(g^2+1)^{\frac{5}{2}}} = -\phi_{(j+1)-j}^{''} = -\phi_1^{''} \tag{8.45}$$

It follows from (8.45) that

$$\min_{0\leq i\leq n+1}\sum_{j=0}^{n+1}(-\phi_{i-j}^{''}) = \min_{0\leq i\leq n+1}\sum_{j=0}^{n+1}g_n(i-j) = \sum_{j=0}^{n+1}g_n(n+1-j) = \sum_{j=0}^{n+1}g_n(0-j).$$

Since $g_n \searrow$, then $g_n(n+1) \leq g_n(j)$ $\forall j = 0, 1, ..., n+1$ and as $g_n(n+1) = \frac{2c^2-1}{(n+1)(1+c^2)^{\frac{5}{2}}}$, one has

$$\frac{2c^2-1}{(1+c^2)^{\frac{5}{2}}}\frac{n+2}{n+1} \leq \frac{1}{h^2}\sum_{j=0}^{n+1}\frac{2g^2-j^2}{(g^2+j^2)^{\frac{5}{2}}} = \min_{0\leq i\leq n+1}\sum_{j=0}^{n+1}(-\phi_{i-j}^{''}). \tag{8.46}$$

On the other hand, let us set $c_0 = \sum_{j=0}^{n+1}(-\phi_j^{''})$ and $c_k = \sum_{j=0}^{k}(-\phi_j^{''}) + \sum_{j=0}^{n+1-k}(-\phi_j^{''})$, for $k = 1, 2, ..., [\frac{d_n}{2}]$. Because the matrices T_{d_n}, $T_{d_n}^T$, $-T_{d_n}$ and $(-T_{d_n})^T$ possess the Perron-Frobenius property, it follows from **Theorems** 8.1.6 and 8.1.10 that

$$\rho(T_{d_n}) = \rho(-T_{d_n}) \quad \leq \quad \max_{1\leq k\leq [\frac{d_n}{2}]} \{c_0, c_k\} = c_{[\frac{d_n}{2}]} \tag{8.47}$$

Or

$$\begin{aligned} c_{[\frac{d_n}{2}]} &= \sum_{j=1}^{[\frac{d_n}{2}]}(-\phi_j'') + \sum_{j=0}^{n+1-[\frac{d_n}{2}]}(-\phi_j'') \\ &= \begin{cases} -\phi_0'' + 2\sum_{j=1}^{[\frac{d_n}{2}]}(-\phi_j'') & \text{if } d_n \text{ is odd} \\ -\phi_0'' - \phi_{[\frac{d_n}{2}]}'' + 2\sum_{j=1}^{[\frac{d_n}{2}]-1}(-\phi_j'') & \text{otherwise} \end{cases} \end{aligned}$$

and since for $j \geq 1$, $-\phi_j'' = g_n(j) \leq g_n(0) = \frac{2}{c^3(n+1)}$, one deduces that $c_{[\frac{d_n}{2}]} \leq \frac{2}{c^3}\frac{d_n}{n+1}$. From (8.43) – (8.44) – (8.46) – (8.47), one has

$$\frac{2c^2-1}{(1+c^2)^{\frac{5}{2}}}\frac{d_n}{n+1} \lesssim \rho(A_{d_n}) \lesssim \frac{2}{c^3}\frac{d_n}{n+1}$$

Then $\lim_{n\to\infty}\rho(A_{d_n}) \in \left[\frac{2c^2-1}{(1+c^2)^{\frac{5}{2}}}, \frac{2}{c^3}\right]$.

Conclusion: The condition number of A_{d_n} grows as $|\lambda_{\min}(A_{d_n})|^{-1}$.

Case 3: Gaussian

Also in this part, we study the problem in the case where $c \geq \sqrt{2}$. First of all, one has: $\phi_k = e^{-\frac{k^2}{g^2}}$ and $\phi_k'' = \frac{2}{h^2g^2}\left(\frac{2k^2}{g^2} - 1\right)e^{-\frac{k^2}{g^2}}$. Then

$$\phi_0 = 1;\ \phi_{n+1} = e^{-\frac{1}{c^2}};$$

$$\phi_0'' = \frac{-2}{c^2};\ \phi_{n+1}'' = \frac{2}{c^2}\left(\frac{2}{c^2} - 1\right)e^{-\frac{1}{c^2}}$$

So,

$$\begin{cases} \lambda_0 = 0 & \text{mult.=n} \\ \lambda_1 = 1 + \frac{2}{c^2} - \left(1 + \frac{2}{c^2} - \frac{4}{c^4}\right)e^{-\frac{1}{c^2}} = \alpha(c) & \text{mult.=1} \\ \lambda_2 = 1 + \frac{2}{c^2} + \left(1 + \frac{2}{c^2} - \frac{4}{c^4}\right)e^{-\frac{1}{c^2}} = \beta(c) & \text{mult.=1} \end{cases}$$

then

$$\frac{1}{d_n}\|\Delta_{d_n}\|_F^2 = \frac{1}{d_n}[|\alpha(c)|^2 + |\beta(c)|^2] \underset{n\to\infty}{\to} 0$$

then

$$\frac{1}{\sqrt{d_n}}\|A_{d_n} - T_{d_n}\|_F \underset{n\to\infty}{\to} 0 \tag{8.48}$$

According to (8.48), it follows from **Lemma** 8.1.2 and **Corollary** 8.1.2 that the sequences $\{A_{d_n}\}_n$ and $\{T_{d_n}\}_n$ are ϵ-equally localized and are clustered. So

$$\rho(A_{d_n}) \underset{n\gg\gg 1}{\approx} \rho(T_{d_n}) \tag{8.49}$$

Study of the spectral radius of T_{d_n}.

Because $c \geq \sqrt{2}$, the entries $\phi_k^{''} = \frac{2}{h^2 g^2}\left(\frac{2k^2}{g^2} - 1\right) e^{-\frac{k^2}{g^2}}$ of T_{d_n} are nonpositive, then $-T_{d_n}$ is an eventually nonnegative matrix. It follows from **Theorem** 8.1.9 that both the matrices $-T_{d_n}$ and $(-T_{d_n})^T$ possess the Perron-Frobenius property, according to **Theorem** 8.1.10, one deduces that

$$\min_{0\leq i\leq n+1} \sum_{j=0}^{n+1} (-\phi_{i-j}^{''}) \leq \rho(-T_{d_n}) \leq \max_{0\leq i\leq n+1} \sum_{j=0}^{n+1} (-\phi_{i-j}^{''}). \tag{8.50}$$

Setting $g_n(x) = \frac{2}{h^2 g^2}\left(1 - \frac{2x^2}{g^2}\right) e^{-\frac{x^2}{g^2}}$ $\forall x \in [0, n+1]$, then $g_n^{'}(x) = \frac{4x}{h^2 g^4}\left(-3 + \frac{2x^2}{g^2}\right) e^{-\frac{x^2}{g^2}} \leq 0$ $\forall x \in [0, n+1]$, then the function g_n is nonincreasing over the interval $[0, n+1]$, since $g_n^{'}(x) = 0 \Leftrightarrow x = 0$, then g_n is a decreasing function over the interval $[0, n+1]$. Furthermore, for $j = 1, 2, ..., n$

$$-\phi_{(j-1)-j}^{''} = \frac{2}{h^2 g^2}\left(1 - \frac{2}{g^2}\right) e^{-\frac{1}{g^2}} = -\phi_{(j+1)-j}^{''} = -\phi_1^{''}. \tag{8.51}$$

According to (8.51) one has

$$\min_{0\leq i\leq n+1} \sum_{j=0}^{n+1} (-\phi_{i-j}^{''}) = \min_{0\leq i\leq n+1} \sum_{j=0}^{n+1} g_n(i-j) = \sum_{j=0}^{n+1} g_n(n+1-j) = \sum_{j=0}^{n+1} g_n(0-j). \tag{8.52}$$

• Let us suppose that $c > \sqrt{2}$. Since $g_n \searrow$, then $g_n(n+1) < g_n(j)$ $\forall j = 0, 1, ..., n$. Or $g_n(n+1) = \frac{2}{c^2}\left(1 - \frac{2}{c^2}\right) e^{-\frac{1}{c^2}}$, then

$$\frac{2e^{-\frac{1}{c^2}}}{c^2}\left(1 - \frac{2}{c^2}\right) d_n \leq \frac{2}{h^2 g^2} \sum_{j=0}^{n+1} \frac{2}{h^2 g^2}\left(1 - \frac{2j^2}{g^2}\right) e^{-\frac{j^2}{g^2}} = \min_{0\leq i\leq n+1} \sum_{j=0}^{n+1} (-\phi_{i-j}^{''}). \tag{8.53}$$

From (8.50) – (8.53), one deduces that

$$\frac{2e^{-\frac{1}{c^2}}}{c^2}\left(1 - \frac{2}{c^2}\right) d_n \quad \leq \quad \rho(-T_{d_n}) \tag{8.54}$$

• If $c = \sqrt{2}$: then, $g_n(n+1) = 0$, so $g_n(n) < g_n(j)$ $\forall j = 0, 1, ..., n-1$. Or $g_n(n) = \left(1 - \frac{n^2}{(n+1)^2}\right) e^{-\frac{n^2}{(n+1)^2}}$, then

$$\sum_{j=0}^{n} g_n(n) = \frac{2n+1}{n+1} e^{-\frac{n^2}{(n+1)^2}} \underset{\infty}{\sim} \frac{2}{e} \tag{8.55}$$

It follows from (8.50) – (8.52) – (8.55) that

$$\frac{2}{e} \lesssim \rho(-T_{d_n}) \tag{8.56}$$

On the other hand, setting $c_0 = \sum_{j=0}^{n+1}(-\phi_j'')$ and $c_k = \sum_{j=0}^{k}(-\phi_j'') + \sum_{j=0}^{n+1-k}(-\phi_j'')$, for $k = 1, 2, ..., [\frac{d_n}{2}]$. Since both the matrices $-T_{d_n}$ and $(-T_{d_n})^T$ possess the Perron- Frobenius property, according to **Theorem** 8.1.10, one deduces that

$$\rho(-T_{d_n}) \leq \max_{1\leq k\leq [\frac{d_n}{2}]}\{c_0, c_k\} = c_{[\frac{d_n}{2}]} \tag{8.57}$$

Or it follows from (8.51) that

$$\begin{aligned} c_{[\frac{d_n}{2}]} &= \sum_{j=1}^{[\frac{d_n}{2}]}(-\phi_j'') + \sum_{j=0}^{n+1-[\frac{d_n}{2}]}(-\phi_j'') \\ &= \begin{cases} -\phi_0'' + 2\sum_{j=1}^{[\frac{d_n}{2}]}(-\phi_j'') & \text{if } d_n \text{ is odd} \\ -\phi_0'' - \phi_{[\frac{d_n}{2}]}'' + 2\sum_{j=1}^{[\frac{d_n}{2}]-1}(-\phi_j'') & \text{otherwise} \end{cases} \end{aligned}$$

and since for $j \geq 1$, $-\phi_j'' = g_n(j) \leq g_n(0) = 1$, one deduces that $c_{[\frac{d_n}{2}]} \leq d_n$. From (8.57), one has

$$\rho(-T_{d_n}) \leq d_n \tag{8.58}$$

According to $(8.56) - (8.58)$, one has

$$\frac{2}{e} \lesssim \rho(-T_{d_n}) \leq d_n \tag{8.59}$$

One concludes according to $(8.49) - (8.54) - (8.59)$, and Theorem 8.1.6 that the spectral radius of A_{d_n} grows as $d_n = n + 2 \quad \forall c > \sqrt{2}$ and least rapidly grows than $d_n = n + 2$ for $c = \sqrt{2}$. Then, the matrix A_{d_n} is ill-conditioned.

8.2.2 2D-dimensional problem

Let us consider the Poisson equation:

$$\begin{cases} \frac{\partial^2 u(x,y)}{\partial x^2} + \frac{\partial^2 u(x,y)}{\partial y^2} = f(x,y) & \text{for } (x,y) \in \Omega = (0,1)^2 \\ u(x,y) = g(x,y) & \text{if } (x,y) \in \partial\Omega \end{cases} \tag{8.60}$$

2.1 Associated linear system and Preconditioners

Discretizing Ω with grid points $z_{jk} = (x_j, y_k) = (hj, hk)$ with $j, k = 0, 1, ..., n+1$, and $h = \frac{1}{n+1}$, one defines an approximated solution of (8.60) given by

$$v(x,y) = \sum_{j=0}^{n+1}\sum_{k=0}^{n+1} v_{jk}\phi\left((x,y) - (x_j, y_k)\right) \tag{8.61}$$

where

$$\phi(x,y) = \begin{cases} \sqrt{x^2+y^2+c^2} & \text{Multiquadric (MQ)} \\ \frac{1}{\sqrt{x^2+y^2+c^2}} & \text{Inverse multiquadric (IMQ)} \\ e^{-\frac{x^2+y^2}{c^2}} & \text{Gaussian} \end{cases} \tag{8.62}$$

The combination of (8.60) and (8.61) give

$$
(8.63)\qquad \begin{cases} \sum\limits_{j=0}^{n+1}\sum\limits_{k=0}^{n+1} v_{jk}\left[\frac{\partial^2\phi}{\partial x^2}+\frac{\partial^2\phi}{\partial y^2}\right](x-x_j,y-y_k)=f(x,y) & \text{for } (x,y)\in\Omega \\ \sum\limits_{j=0}^{n+1}\sum\limits_{k=0}^{n+1} v_{jk}\phi(-x_j,y-y_k)=g(0,y) & \text{if } y\in[0,1] \\ \sum\limits_{j=0}^{n+1}\sum\limits_{k=0}^{n+1} v_{jk}\phi(1-x_j,y-y_k)=g(1,y) & \text{if } y\in[0,1] \\ \sum\limits_{j=0}^{n+1}\sum\limits_{k=0}^{n+1} v_{jk}\phi(x-x_j,-y_k)=g(x,0) & \text{if } x\in(0,1) \\ \sum\limits_{j=0}^{n+1}\sum\limits_{k=0}^{n+1} v_{jk}\phi(x-x_j,1-y_k)=g(x,1) & \text{if } x\in(0,1) \end{cases}
$$

By direct computations, one has:

$$
\frac{\partial\phi}{\partial x}(x,y)=\begin{cases} x(x^2+y^2+c^2)^{-\frac{1}{2}} \\ -x(x^2+y^2+c^2)^{-\frac{3}{2}} \\ \frac{-2x}{c^2}e^{-\frac{x^2+y^2}{c^2}} \end{cases},\qquad \frac{\partial\phi}{\partial y}(x,y)=\begin{cases} y(x^2+y^2+c^2)^{-\frac{1}{2}} \\ -y(x^2+y^2+c^2)^{-\frac{3}{2}} \\ \frac{-2y}{c^2}e^{-\frac{x^2+y^2}{c^2}} \end{cases}
$$

then

$$
\frac{\partial^2\phi}{\partial x^2}(x,y)=\begin{cases} (x^2+y^2+c^2)^{-\frac{1}{2}}-x^2(x^2+y^2+c^2)^{-\frac{3}{2}} \\ -(x^2+y^2+c^2)^{-\frac{3}{2}}+3x^2(x^2+y^2+c^2)^{-\frac{5}{2}} \\ \frac{-2}{c^2}e^{-\frac{x^2+y^2}{c^2}}+\frac{4x^2}{c^2}e^{-\frac{x^2+y^2}{c^2}} \end{cases}
$$

$$
\frac{\partial^2\phi}{\partial y^2}(x,y)=\begin{cases} (x^2+y^2+c^2)^{-\frac{1}{2}}-y^2(x^2+y^2+c^2)^{-\frac{3}{2}} \\ -(x^2+y^2+c^2)^{-\frac{3}{2}}+3y^2(x^2+y^2+c^2)^{-\frac{5}{2}} \\ \frac{-2}{c^2}e^{-\frac{x^2+y^2}{c^2}}+\frac{4y^2}{c^2}e^{-\frac{x^2+y^2}{c^2}} \end{cases}
$$

then

$$
(8.64)\qquad \frac{\partial^2\phi}{\partial x^2}(x,y)+\frac{\partial^2\phi}{\partial y^2}(x,y)=\begin{cases} (x^2+y^2+2c^2)(x^2+y^2+c^2)^{-\frac{3}{2}} & \text{(MQ)} \\ (x^2+y^2-2c^2)(x^2+y^2+c^2)^{-\frac{5}{2}} & \text{(IMQ)} \\ \frac{4}{c^2}((x^2+y^2)-c^2)e^{-\frac{x^2+y^2}{c^2}} & \text{(Gaussian)} \end{cases}
$$

The linear system associated with (8.63) is defined as follows:

$$
(8.65)\qquad \begin{cases} (a)\ \sum\limits_{l,p=0}^{n+1} v_{lp}\left[\frac{\partial^2\phi}{\partial x^2}+\frac{\partial^2\phi}{\partial y^2}\right](x_j-x_l,y_k-y_p)=f(x_j,y_k): & j,k=1,...,n \\ (b)\ \sum\limits_{l,p=0}^{n+1} v_{lp}\phi(-x_l,y_j-y_p)=g(0,y_j): & j=0,1,...,n+1 \\ (c)\ \sum\limits_{l,p=0}^{n+1} v_{lp}\phi(1-x_l,y_j-y_p)=g(1,y_j) & j=0,1,...,n+1 \\ (d)\ \sum\limits_{l,p=0}^{n+1} v_{lp}\phi(x_j-x_l,-y_p)=g(x_j,0): & j=1,...,n \\ (e)\ \sum\limits_{l,p=0}^{n+1} v_{lp}\phi(x_j-x_l,1-y_p)=g(x_j,1): & j=1,...,n \end{cases}
$$

(a) implies for $j,k=1,2,...,n:\sum\limits_{l,p=0}^{n+1}C_{k-p}^{j-l}v_{lp}=f_{jk}$, i.e.

$$[C_{k-0}^{j-0},...,C_{k-(n+1)}^{j-0}|C_{k-0}^{j-1},...,C_{k-(n+1)}^{j-1}|...|C_{k-0}^{j-(n+1)},...,C_{k-(n+1)}^{j-(n+1)}]\begin{bmatrix}v_{00}\\ \vdots\\ v_{0,n+1}\\ \hline v_{1,0}\\ \vdots\\ v_{1,n+1}\\ \hline \vdots\\ v_{n+1,0}\\ \vdots\\ v_{n+1,n+1}\end{bmatrix}=f_{jk}\tag{8.66}$$

where

$$C_{k-p}^{j-l}=\begin{cases}\frac{1}{\sqrt{(x_j-x_l)^2+(y_k-y_p)^2+c^2}}+\frac{c^2}{(x_j-x_l)^2+(y_k-y_p)^2+c^2)^{\frac{3}{2}}} & \text{(MQ)}\\ \frac{(x_j-x_l)^2+(y_k-y_p)^2-2c^2}{((x_j-x_l)^2+(y_k-y_p)^2+c^2)^{\frac{5}{2}}} & \text{(IMQ)}\\ \frac{4}{c^2}((x_j-x_l)^2+(y_k-y_p)^2-c^2)e^{-\frac{(x_j-x_l)^2+(y_k-y_p)^2}{c^2}} & \text{(Gaussian)}\end{cases}$$

Setting $g=c/h$ and since $x_j=jh$, $y_j=jh$, then

$$C_{k-p}^{j-l}=\begin{cases}\frac{1}{h}\frac{1}{\sqrt{(j-l)^2+(k-p)^2+g^2}}+\frac{1}{h}\frac{g^2}{[(j-l)^2+(k-p)^2+g^2]^{\frac{3}{2}}} & \text{(MQ)}\\ \frac{1}{h^3}\frac{(j-l)^2+(k-p)^2-2g^2}{[(j-l)^2+(k-p)^2+g^2]^{\frac{5}{2}}} & \text{(IMQ)}\\ \frac{4}{h^2g^4}[(j-l)^2+(k-p)^2-g^2]e^{-\frac{(j-l)^2+(k-p)^2}{g^2}} & \text{(Gaussian)}\end{cases}\tag{8.67}$$

(b) implies for $j=0,1,...,n+1:\sum\limits_{l,p=0}^{n+1}v_{lp}\phi(-x_l,y_j-y_p)=g(0,y_j)$, i.e.

(8.68)

$$[\phi_{j-0}^{0-0},...,\phi_{j-(n+1)}^{0-0}|\phi_{j-0}^{0-1},...,\phi_{j-(n+1)}^{0-1}|...|\phi_{j-0}^{0-(n+1)},...,\phi_{j-(n+1)}^{0-(n+1)}]\begin{bmatrix}v_{00}\\ \vdots\\ v_{0,n+1}\\ \hline v_{1,0}\\ \vdots\\ v_{1,n+1}\\ \hline \vdots\\ v_{n+1,0}\\ \vdots\\ v_{n+1,n+1}\end{bmatrix}=g_{0j}=g(0,jh)$$

(c) implies for $j=0,1,...,n+1$: $\sum\limits_{l,p=0}^{n+1}v_{lp}\phi(1-x_l,y_j-y_p)=g(1,y_j)$, i.e.,

$$[\phi_{j-0}^{(n+1)-0},...,\phi_{j-(n+1)}^{(n+1)-0}|...|\phi_{j-0}^{(n+1)-(n+1)},...,\phi_{j-(n+1)}^{(n+1)-(n+1)}]\begin{bmatrix}v_{00}\\ \vdots\\ v_{0,n+1}\\ \hline v_{1,0}\\ \vdots\\ v_{1,n+1}\\ \hline \vdots\\ v_{n+1,0}\\ \vdots\\ v_{n+1,n+1}\end{bmatrix}=g(1,jh)\tag{8.69}$$

(d) implies for $j = 1, ..., n : \sum_{l,p=0}^{n+1} v_{lp}\phi(x_j - x_l, -y_p) = g(x_j, 0)$, i.e.

(8.70)

$$[\phi_{0-0}^{j-0}, ..., \phi_{0-(n+1)}^{j-0} | \phi_{0-0}^{j-1}, ..., \phi_{0-(n+1)}^{j-1} | ... | \phi_{0-0}^{j-(n+1)}, ..., \phi_{0-(n+1)}^{j-(n+1)}] \begin{bmatrix} v_{00} \\ \vdots \\ v_{0,n+1} \\ \hline v_{1,0} \\ \vdots \\ v_{1,n+1} \\ \hline \vdots \\ \hline v_{n+1,0} \\ \vdots \\ v_{n+1,n+1} \end{bmatrix} = g_{j0} = g(jh, 0)$$

(e) implies for $j = 1, ..., n : \sum_{l,p=0}^{n+1} v_{lp}\phi(x_j - x_l, 1 - y_p) = g(x_j, 1)$, i.e.

(8.71)
$$[\phi_{(n+1)-0}^{j-0}, ..., \phi_{(n+1)-(n+1)}^{j-0} | ... | \phi_{(n+1)-0}^{j-(n+1)}, ..., \phi_{(n+1)-(n+1)}^{j-(n+1)}] \begin{bmatrix} v_{00} \\ \vdots \\ v_{0,n+1} \\ \hline v_{1,0} \\ \vdots \\ v_{1,n+1} \\ \hline \vdots \\ \hline v_{n+1,0} \\ \vdots \\ v_{n+1,n+1} \end{bmatrix} = g(jh, 1)$$

where $\phi(x_j - x_l, y_k - y_p) = \phi_{k-p}^{j-l}$ and

(8.72)
$$\phi_{k-p}^{j-l} = \begin{cases} h[(j-l)^2 + (k-p)^2 + g^2]^{\frac{1}{2}} & \text{MQ} \\ ((j-l)^2 + (k-p)^2 + g^2)^{-\frac{1}{2}} & \text{IMQ} \\ e^{-\frac{(j-l)^2+(k-p)^2}{g^2}} & \text{Gaussian} \end{cases}$$

From (8.66) to (8.72), one deduces the following linear system:

(8.73)
$$\begin{bmatrix} A_{0-0}^{(n+2)} & A_{0-1}^{(n+2)} & \dots & A_{0-n}^{(n+2)} & A_{0-(n+1)}^{(n+2)} \\ A_{1-0}^{(n+2)} & A_{1-1}^{(n+2)} & \dots & \dots & A_{1-(n+1)}^{(n+2)} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ A_{n-0}^{(n+2)} & A_{n-1}^{(n+2)} & \dots & A_{n-n}^{(n+2)} & A_{n-(n+1)}^{(n+2)} \\ A_{(n+1)-0}^{(n+2)} & A_{(n+1)-1}^{(n+2)} & \dots & A_{(n+1)-n}^{(n+2)} & A_{(n+1)-(n+1)}^{(n+2)} \end{bmatrix} \begin{bmatrix} v^{(0)} \\ v^{(1)} \\ \vdots \\ v^{(n)} \\ v^{(n+1)} \end{bmatrix} = \begin{bmatrix} f^{(0)} \\ f^{(1)} \\ \vdots \\ f^{(n)} \\ f^{(n+1)} \end{bmatrix}$$

i.e,

$$A_{d_n} v = \tilde{f}$$

where $d_n = (n+2)^2$ and

$$A_{d_n} = \begin{bmatrix} A_{0-0}^{(n+2)} & A_{0-1}^{(n+2)} & \dots & A_{0-n}^{(n+2)} & A_{0-(n+1)}^{(n+2)} \\ A_{1-0}^{(n+2)} & A_{1-1}^{(n+2)} & \dots & \dots & A_{1-(n+1)}^{(n+2)} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ A_{n-0}^{(n+2)} & A_{n-1}^{(n+2)} & \dots & A_{n-n}^{(n+2)} & A_{n-(n+1)}^{(n+2)} \\ A_{(n+1)-0}^{(n+2)} & A_{(n+1)-1}^{(n+2)} & \dots & A_{(n+1)-n}^{(n+2)} & A_{(n+1)-(n+1)}^{(n+2)} \end{bmatrix}; \quad v = \begin{bmatrix} v^{(0)} \\ v^{(1)} \\ \vdots \\ v^{(n)} \\ v^{(n+1)} \end{bmatrix}; \quad \tilde{f} = \begin{bmatrix} f^{(0)} \\ f^{(1)} \\ \vdots \\ f^{(n)} \\ f^{(n+1)} \end{bmatrix}$$

with for $j = 0, 1, ..., n+1$

$$A_{0-j}^{(n+2)} = [\phi_{l-p}^{0-j}]_{l,p=0}^{n+1}; \quad A_{(n+1)-j}^{(n+2)} = [\phi_{l-p}^{(n+1)-j}]_{l,p=0}^{n+1}, \tag{8.74}$$

for $j = 1, 2, ..., n$ and $l = 0, 1, ..., n+1$,

$$A_{j-l}^{(n+2)} = \begin{bmatrix} \phi_{0-0}^{j-l} & \phi_{0-1}^{j-l} & \dots & \phi_{0-n}^{j-l} & \phi_{0-(n+1)}^{j-l} \\ C_{1-0}^{j-l} & C_{1-1}^{j-l} & \dots & C_{1-n}^{j-l} & C_{1-(n+1)}^{j-l} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ C_{n-0}^{j-l} & C_{n-1}^{j-l} & \dots & C_{n-n}^{j-l} & C_{n-(n+1)}^{j-l} \\ \phi_{(n+1)-0}^{j-l} & \phi_{(n+1)-1}^{j-l} & \dots & \phi_{(n+1)-n}^{j-l} & \phi_{(n+1)-(n+1)}^{j-l} \end{bmatrix} \tag{8.75}$$

$$= \begin{bmatrix} \phi_0^{j-l} & \phi_1^{j-l} & \dots & \phi_n^{j-l} & \phi_{(n+1)}^{j-l} \\ C_1^{j-l} & C_0^{j-l} & \dots & C_{n-1}^{j-l} & C_n^{j-l} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ C_n^{j-l} & C_{n-1}^{j-l} & \dots & C_0^{j-l} & C_1^{j-l} \\ \phi_{n+1}^{j-l} & \phi_n^{j-l} & \dots & \phi_1^{j-l} & \phi_0^{j-l} \end{bmatrix},$$

for $k = 0, 1, ..., n+1$,

$$v^{(k)} = (v_{k,0}, v_{k,1}, ..., v_{k,n+1})^T; \quad f^{(0)} = (g_{0,0}, g_{0,1}, ..., g_{0,n+1})^T; \quad f^{(n+1)} = (g_{n+1,0}, ..., g_{n+1,n+1})^T$$

and for $p = 1, 2, ..., n$,

$$f^{(p)} = (g_{p,0}, f_{p,1}, ..., f_{p,n}, g_{p,n+1})^T.$$

For $l = 0, 1, ..., n+1$, it follows from (8.74) that the matrices $A_{0-l}^{(n+2)}$ and $A_{(n+1)-l}^{(n+2)}$ are symmetric Toeplitz matrices. For $l = 0, 1, ..., n+1$, let us set

$$\Delta_{0-l}^{(n+2)} = \left[\phi_{k-p}^{0-l} - C_{k-p}^{0-l}\right]_{k,p=0}^{n+1} \text{ and } \Delta_{(n+1)-l}^{(n+2)} = \left[\phi_{k-p}^{(n+1)-l} - C_{k-p}^{(n+1)-l}\right]_{k,p=0}^{n+1}, \tag{8.76}$$

From (8.76) the matrices $\Delta_{0-l}^{(n+2)}$ and $\Delta_{(n+1)-l}^{(n+2)}$ are symmetric. Hence, it follows from $l = 0, 1, ..., n+1$, the symmetric Toeplitz matrices

$$T_{0-l}^{(n+2)} = A_{0-l}^{(n+2)} - \Delta_{0-l}^{(n+2)} \text{ and } T_{(n+1)-l}^{(n+2)} = A_{(n+1)-l}^{(n+2)} - \Delta_{(n+1)-l}^{(n+2)}.$$

Whence

$$P_{0-l}^{(n+2)} = T_{0-l}^{(n+2)} - H(T_{0-l}^{(n+2)}) \text{ and } P_{(n+1)-l}^{(n+2)} = T_{(n+1)-l}^{(n+2)} - H(T_{(n+1)-l}^{(n+2)}) \tag{8.77}$$

are natural preconditioners for $T_{0-l}^{(n+2)}$ and $T_{(n+1)-l}^{(n+2)}$ respectively.
According to the study done in uni-dimension, for $j = 1, 2, ..., n$ and $l = 0, 1, ..., n+1$, one has:

$$A_{j-l}^{(n+2)} = T_{j-l}^{(n+2)} + \Delta_{j-l}^{(n+2)} \tag{8.78}$$

where

$$\Delta_{j-l}^{(n+2)} = \begin{bmatrix} \phi_0^{j-l} - C_0^{j-l} & \phi_1^{j-l} - C_1^{j-l} & \dots & \phi_{n+1}^{j-l} - C_{n+1}^{j-l} \\ 0 & \dots & & 0 \\ \vdots & & & \vdots \\ 0 & \dots & & 0 \\ \phi_{n+1}^{j-l} - C_{n+1}^{j-l} & \phi_n^{j-l} - C_n^{j-l} & \dots & \phi_0^{j-l} - C_0^{j-l} \end{bmatrix} \text{ and } T_{j-l}^{(n+2)} = [C_{k-p}^{j-l}]_{k,p=0}^{n+1} \tag{8.79}$$

and

$$\begin{cases} \lambda_0(\Delta_{j-l}^{(n+2)}) = 0, & \text{mult.=n} \\ \lambda_1(\Delta_{j-l}^{(n+2)}) = \phi_0^{j-l} - C_0^{j-l} + C_{n+1}^{j-l} - \phi_{n+1}^{j-l} & \text{mult.=1} \\ \lambda_2(\Delta_{j-l}^{(n+2)}) = \phi_0^{j-l} - C_0^{j-l} - C_{n+1}^{j-l} + \phi_{n+1}^{j-l} & \text{mult.=1} \end{cases} \tag{8.80}$$

with

$$\phi_0^{j-l} = \begin{cases} h((j-l)^2+g^2)^{\frac{1}{2}} & \text{(MQ)} \\ \frac{1}{h}((j-l)^2+g^2)^{-\frac{1}{2}} & \text{(IMQ)} \\ e^{-\frac{(j-l)^2}{g^2}} & \text{(Gaussian).} \end{cases} \tag{8.81}$$

$$C_0^{j-l} = \begin{cases} \frac{1}{h}((j-l)^2+g^2)^{-\frac{1}{2}} + \frac{1}{h}g^2((j-l)^2+g^2)^{-\frac{3}{2}} & \text{(MQ)} \\ \frac{1}{h^3}((j-l)^2-2g^2)((j-l)^2+g^2)^{-\frac{5}{2}} & \text{(IMQ)} \\ \frac{4}{h^2g^4}((j-l)^2-g^2)e^{-\frac{(j-l)^2}{g^2}} & \text{(Gaussian).} \end{cases} \tag{8.82}$$

$$\phi_{n+1}^{j-l} = \begin{cases} h((n+1)^2+(j-l)^2+g^2)^{\frac{1}{2}} & \text{(MQ)} \\ \frac{1}{h}((n+1)^2+(j-l)^2+g^2)^{-\frac{1}{2}} & \text{(IMQ)} \\ e^{-\frac{(n+1)^2+(j-l)^2}{g^2}} & \text{(Gaussian).} \end{cases} \tag{8.83}$$

(8.84)

$$C_{n+1}^{j-l} = \begin{cases} \frac{1}{h}((n+1)^2+(j-l)^2+g^2)^{-\frac{1}{2}} + \frac{1}{h}g^2((n+1)^2+(j-l)^2+g^2)^{-\frac{3}{2}} & \text{(MQ)} \\ \frac{1}{h^3}((n+1)^2+(j-l)^2-2g^2)((n+1)^2+(j-l)^2+g^2)^{-\frac{5}{2}} & \text{(IMQ)} \\ \frac{4}{h^2g^4}((n+1)^2+(j-l)^2-g^2)e^{-\frac{(n+1)^2+(j-l)^2}{g^2}} & \text{(Gaussian).} \end{cases}$$

Since for $j = 1, 2, ..., n$ and $l = 0, 1, ..., n+1$; $T_{j-l}^{(n+2)}$ is a symmetric Toeplitz matrix, then

$$P_{j-l}^{(n+2)} = T_{j-l}^{(n+2)} - H(T_{j-l}^{(n+2)}) \tag{8.85}$$

is a natural preconditioner for $T_{j-l}^{(n+2)}$. On the other side, $\Delta_{j-l}^{(n+2)}$ just has two rows non identically null. One easily shows that the non null rows are linearly independent, then $\text{rank}(\Delta_{j-l}^{(n+2)}) = 2$ so, $\text{rank}(A_{j-l}^{(n+2)} - T_{j-l}^{(n+2)}) = o(n+2)$. It follows from **Lemma** 8.1.1 that the matrix sequences $\{A_{j-l}^{(n+2)}\}_n$ and $\{T_{j-l}^{(n+2)}\}_n$ are equally localized (EL) and equally distributed (ED). i.e,

$$\{A_{j-l}^{(n+2)}\}_n \simeq_{L.D} \{T_{j-l}^{(n+2)}\}_n \tag{8.86}$$

One deduces from (8.85) – (8.86) that $P_{j-l}^{(n+2)} = T_{j-l}^{(n+2)} - H(T_{j-l}^{(n+2)})$ is an efficient preconditioner for $A_{j-l}^{(n+2)}$. Setting $T_{d_n} = [T_{j-l}^{(n+2)}]_{j,l=0}^{n+1}$ then, T_{d_n} is a symmetric block Toeplitz matrix with symmetric Toeplitz blocks. Therefore

$$P_{d_n} = [P_{j-l}^{(n+2)}]_{j,l=0}^{n+1} \tag{8.87}$$

is a good preconditioner for T_{d_n} (see, [92, 62, 40]). Furthermore, setting

$$\Delta_{d_n} = A_{d_n} - T_{d_n} = [\Delta_{j-l}^{(n+2)}]_{j,l=0}^{(n+2)} \tag{8.88}$$

According to (8.76) – (8.78) one has

$$\begin{cases} rank(\Delta_{j-l}^{(n+2)}) \leq n+2, & l = 0,1,...,n+1 \text{ and } j \in \{0, n+1\}; \\ \Delta_{j-l}^{(n+2)} = A_{j-l}^{(n+2)} - T_{j-l}^{(n+2)}, & l = 0,1,...,n+1 \text{ and } \text{ j=1,2,...,n.} \end{cases} \tag{8.89}$$

For $j = 1, 2, ..., n$ and $l = 0, 1, ..., n+1$, one deduces from (8.79) that

$$rank(\Delta_{j-l}^{(n+2)}) = 2 \tag{8.90}$$

Exploiting (8.88) – (8.89) – (8.90), one easily shows that rank$(\Delta_{d_n}) \leq 4n + 4$, then

$$rank(A_{d_n} - T_{d_n}) = o(d_n). \tag{8.91}$$

It follows from (8.91) and **Lemma** 8.1.1 that the matrix sequences $\{A_{d_n}\}_n$ and $\{T_{d_n}\}_n$ are equally localized (EL) and equally distributed (ED). i.e,

$$\{A_{d_n}\}_n \simeq_{L.D} \{T_{d_n}\}_n \tag{8.92}$$

From (8.87) and (8.92), one deduces that P_{d_n} is a good preconditioner for A_{d_n}.

2.2. Study of the spectral radius of A_{d_n}

Throughout this section we suppose that $c \geq \sqrt{\frac{1+\sqrt{5}}{2}}$ in Multiquadric case, $c \geq 1$ in the Inverse Multiquadric case and $c \geq \sqrt{2}$ in the Gaussian case. These requirements impose to the Toeplitz matrix T_{d_n} to be nonpositive and to the matrix Δ_{d_n} to be nonnegative. According to (8.88), $A_{d_n} = T_{d_n} + \Delta_{d_n}$ then

$$(A_{d_n})^2 = (T_{d_n})^2 + T_{d_n}\Delta_{d_n} + \Delta_{d_n}T_{d_n} + (\Delta_{d_n})^2 \tag{8.93}$$

or

$$T_{d_n} = [T_{j-l}^{(n+2)}]_{j,l=0}^{n+1} \text{ and } \Delta_{d_n} = [\Delta_{j-l}^{(n+2)}]_{j,l=0}^{n+1}$$

then

$$(T_{d_n})^2 = \left[\sum_{s=0}^{n+1} T_{j-s}^{(n+2)} T_{s-l}^{(n+2)}\right]_{j,l=0}^{n+1}; \ (\Delta_{d_n})^2 = \left[\sum_{s=0}^{n+1} \Delta_{j-s}^{(n+2)} \Delta_{s-l}^{(n+2)}\right]_{j,l=0}^{n+1}$$

$$T_{d_n}\Delta_{d_n} = \left[\sum_{s=0}^{n+1} T_{j-s}^{(n+2)} \Delta_{s-l}^{(n+2)}\right]_{j,l=0}^{n+1}; \ \Delta_{d_n}T_{d_n} = \left[\sum_{s=0}^{n+1} \Delta_{j-s}^{(n+2)} T_{s-l}^{(n+2)}\right]_{j,l=0}^{n+1}$$

then

$$(A_{d_n})^2 = \left[\sum_{s=0}^{n+1} \left\{T_{j-s}^{(n+2)} T_{s-l}^{(n+2)} + T_{j-s}^{(n+2)} \Delta_{s-l}^{(n+2)} + \Delta_{j-s}^{(n+2)} T_{s-l}^{(n+2)} + \Delta_{j-s}^{(n+2)} \Delta_{s-l}^{(n+2)}\right\}\right]_{j,l=0}^{n+1} \tag{8.94}$$

For $k, p = 0, 1, ..., n+1$

$$(T_{j-s}^{(n+2)} T_{s-l}^{(n+2)})_{k,p} = \sum_{q=0}^{n+1} (T_{j-s}^{(n+2)})_{kq} (T_{s-l}^{(n+2)})_{qp} = \sum_{q=0}^{n+1} C_{k-q}^{j-s} C_{q-p}^{s-l}, \tag{8.95}$$

$$(T_{j-s}^{(n+2)}\Delta_{s-l}^{(n+2)})_{k,p} = \sum_{q=0}^{n+1}(T_{j-s}^{(n+2)})_{kq}(\Delta_{s-l}^{(n+2)})_{qp} = \sum_{q=0}^{n+1} C_{k-q}^{j-s}(\phi_{q-p}^{s-l} - C_{q-p}^{s-l}), \tag{8.96}$$

$$(\Delta_{j-s}^{(n+2)}T_{s-l}^{(n+2)})_{k,p} = \sum_{q=0}^{n+1}(\Delta_{j-s}^{(n+2)})_{kq}(T_{s-l}^{(n+2)})_{qp} = \sum_{q=0}^{n+1}(\phi_{k-q}^{j-s} - C_{k-q}^{j-s})C_{q-p}^{s-l}, \tag{8.97}$$

$$(\Delta_{j-s}^{(n+2)}\Delta_{s-l}^{(n+2)})_{k,p} = \sum_{q=0}^{n+1}(\Delta_{j-s}^{(n+2)})_{kq}(\Delta_{s-l}^{(n+2)})_{qp} = \sum_{q=0}^{n+1}(\phi_{k-q}^{j-s} - C_{k-q}^{j-s})(\phi_{q-p}^{s-l} - C_{q-p}^{s-l}). \tag{8.98}$$

From (8.94) − (8.95) − (8.96) − (8.97) − (8.98), one deduces that $(A_{d_n})^2 = \left[[a_{k,p}^{j,l}]_{k,p=0}^{n+1}\right]_{j,l=0}^{n+1}$ where for $j,l = 0,1,...,n+1$ and $k,p = 0,1,...,n+1$

$$\begin{aligned} a_{k,p}^{j,l} &= \sum_{s=0}^{n+1}\sum_{q=0}^{n+1}\{C_{k-q}^{j-s}C_{q-p}^{s-l} + C_{k-q}^{j-s}(\phi_{q-p}^{s-l} - C_{q-p}^{s-l}) + (\phi_{k-q}^{j-s} - C_{k-q}^{j-s})C_{q-p}^{s-l} \\ &+ (\phi_{k-q}^{j-s} - C_{k-q}^{j-s})(\phi_{q-p}^{s-l} - C_{q-p}^{s-l})\} \\ &= \sum_{s=0}^{n+1}\sum_{q=0}^{n+1}\phi_{k-q}^{j-s}\phi_{q-p}^{s-l} > 0 \end{aligned}$$

since $\phi_{n,m}^{i,r} > 0\ \forall i,n,m,r = 0,1,...,n+1$. Then

$$(A_{d_n})^2 > 0. \tag{8.99}$$

On the other side, $-T_{d_n} = \left[[-C_{k-p}^{j-l}]_{k,p=0}^{n+1}\right]_{j,l=0}^{n+1} \geq 0$ since $-C_{k-p}^{j-l} \geq 0$ (for Inverse Multiquadric and Gaussian cases) $\forall i,j,k,p = 0,1,...,n+1$. Then $(T_{d_n})^2 = (-T_{d_n})^2 \geq 0$.
All the coefficients of the matrices $(A_{d_n})^2$ and $(T_{d_n})^2$ are nonnegative, then both $(A_{d_n})^2$ and $(T_{d_n})^2$ are nonnegative matrices, so A_{d_n} and T_{d_n} are eventually nonnegative matrices. According to Theorem 8.1.9, the matrices A_{d_n}, $A_{d_n}^T$, T_{d_n} and $T_{d_n}^T$ possess the Perron-Frobenius property. Since $A_{d_n} - T_{d_n} = \Delta_{d_n} \geq 0$, it follows from Theorems 8.1.6 − 8.1.11 that

$$\rho(-T_{d_n}) = \rho(T_{d_n}) \leq \rho(A_{d_n}) \tag{8.100}$$

and according to Theorem 8.1.10, one has

$$\min_{0\leq j,k\leq n+1}\sum_{l=0}^{n+1}\sum_{p=0}^{n+1}(-C_{k-p}^{j-l}) \leq \rho(-T_{d_n}) \leq \max_{0\leq j,k\leq n+1}\sum_{l=0}^{n+1}\sum_{p=0}^{n+1}(-C_{k-p}^{j-l}) \tag{8.101}$$

<u>Case 4</u>: Multiquadric $\left(c \geq \sqrt{\frac{1+\sqrt{5}}{2}}\right)$

For $j,l,p,k = 0,1,...,n+1$: $C_{k-p}^{j-l} = \frac{1}{h}\frac{1}{\sqrt{(j-l)^2+(k-p)^2+g^2}} + \frac{1}{h}\frac{g^2}{[(j-l)^2+(k-p)^2+g^2]^{\frac{3}{2}}}$. Both T_{d_n} and A_{d_n} are eventually positive matrices. according to Theorem 8.1.8, the matrices A_{d_n}, $A_{d_n}^T$, T_{d_n}

and $T_{d_n}^T$ possess the strong Perron-Frobenius property. Since $A_{d_n} - T_{d_n} = \Delta_{d_n} \geq 0$ (since $c \geq \sqrt{\frac{1+\sqrt{5}}{2}}$), it follows from Theorem 8.1.11 that

$$\rho(T_{d_n}) \leq \rho(A_{d_n}) \tag{8.102}$$

and according to **Theorem** 8.1.10, one has

$$\min_{0\leq j,k\leq n+1} \sum_{l=0}^{n+1}\sum_{p=0}^{n+1} C_{k-p}^{j-l} < \rho(T_{d_n}) < \max_{0\leq j,k\leq n+1} \sum_{l=0}^{n+1}\sum_{p=0}^{n+1} C_{k-p}^{j-l} \tag{8.103}$$

For $j,k = 0,1,...,n+1$,

$$\sum_{l=0}^{n+1}\sum_{p=0}^{n+1} C_{k-p}^{j-l} = \frac{1}{h}\sum_{l=0}^{n+1}\sum_{p=0}^{n+1}\left(\frac{1}{\sqrt{(j-l)^2+(k-p)^2+g^2}} + \frac{g^2}{[(j-l)^2+(k-p)^2+g^2]^{\frac{3}{2}}}\right) \tag{8.104}$$

First of all, let us study the functions: $f_n(x) = \frac{1}{(g^2+x^2+(j-l)^2)^{\frac{1}{2}}}$ and $g_n(x) = \frac{g^2}{(g^2+x^2+(j-l)^2)^{\frac{3}{2}}}$ over the domain $[-n-1, n+1]$. Since f_n and g_n are even functions, the study of these functions reduces over the interval $[0, n+1]$. Since $f_n^{'}(x) = \frac{-x}{(g^2+x^2+(j-l)^2)^{\frac{3}{2}}} < 0$ and $g_n^{'}(x) = \frac{-3g^2x}{(g^2+x^2+(j-l)^2)^{\frac{5}{2}}} < 0$ for $x > 0$, then f_n and g_n are decreasing functions on [0,n+1]. One deduces that,

$$\min_{0\leq j,k\leq n+1} \sum_{l=0}^{n+1}\sum_{p=0}^{n+1} C_{k-p}^{j-l} = \frac{1}{h}\sum_{l=0}^{n+1}\sum_{p=0}^{n+1}\left(\frac{1}{\sqrt{l^2+p^2+g^2}} + \frac{g^2}{[l^2+p^2+g^2]^{\frac{3}{2}}}\right) \tag{8.105}$$

since for $l,p = 0,1,...,n+1$,

$$\frac{1}{n+1}\frac{1}{\sqrt{2+c^2}} \leq \frac{1}{\sqrt{l^2+p^2+g^2}} \text{ and } \frac{1}{n+1}\frac{c^2}{(2+c^2)^{\frac{3}{2}}} \leq \frac{g^2}{(l^2+p^2+g^2)^{\frac{3}{2}}}$$

then

$$\left(\frac{1}{\sqrt{2+c^2}} + \frac{c^2}{(2+c^2)^{\frac{3}{2}}}\right)(n+2)^2 \leq \frac{1}{h}\sum_{l=0}^{n+1}\sum_{p=0}^{n+1}\left(\frac{1}{\sqrt{l^2+p^2+g^2}} + \frac{g^2}{[l^2+p^2+g^2]^{\frac{3}{2}}}\right) \tag{8.106}$$

It follows from (8.102) – (8.103) – (8.105) – (8.106) that the spectral radius of the matrix A_{d_n} grows as $(n+2)^2$. Then the matrix A_{d_n} is ill-conditioned.

Case 5: Inverse Multiquadric $(c \geq 1)$
For $j,l,p,k = 0,1,...,n+1$:

$$C_{k-p}^{j-l} = \frac{1}{h^3}\frac{(j-l)^2+(k-p)^2-2g^2}{[(j-l)^2+(k-p)^2+g^2]^{\frac{5}{2}}} \leq 0 \quad \text{since } c \geq 1, \tag{8.107}$$

it follows from (8.99) and (8.107) that, both T_{d_n} and A_{d_n} are eventually nonnegative matrices. according to Theorem 8.1.8, the matrices A_{d_n}, $A_{d_n}^T$, T_{d_n} and $T_{d_n}^T$ possess the Perron-Frobenius property. For $j,l = 0,1,...,n+1$ fixed, and for $k = 0,1,...n+1$, $\phi_k^{j-l} - C_k^{j-l} \geq 0$; Then $\Delta_{j-l}^{(n+2)} \geq 0$, so, $A_{d_n} - T_{d_n} = \Delta_{d_n} \geq 0$, it follows from Theorem 8.1.11 that

$$\rho(T_{d_n}) \leq \rho(A_{d_n}) \tag{8.108}$$

Furthermore, $(-T_{d_n})^T$ possesses the Perron-Frobenius property. According to Theorem 8.1.10, one has

$$\min_{0\leq j,k\leq n+1} \sum_{l=0}^{n+1}\sum_{p=0}^{n+1}(-C_{k-p}^{j-l}) \leq \rho(-T_{d_n}) \leq \max_{0\leq j,k\leq n+1} \sum_{l=0}^{n+1}\sum_{p=0}^{n+1}(-C_{k-p}^{j-l}) \tag{8.109}$$

For $j,k = 0,1,...,n+1$,

$$\sum_{l=0}^{n+1}\sum_{p=0}^{n+1}(-C_{k-p}^{j-l}) = \frac{1}{h^3}\sum_{l=0}^{n+1}\sum_{p=0}^{n+1}\left(\frac{2g^2-(j-l)^2-(k-p)^2}{[(j-l)^2+(k-p)^2+g^2]^{\frac{5}{2}}}\right) \tag{8.110}$$

Case: $c > 1$. Let us study the function: $f_{n,j-l}(x) = \frac{-x^2-(j-l)^2+2g^2}{[x^2+(j-l)^2+g^2]^{\frac{5}{2}}}$ over the domain $[-n-1, n+1]$. Since $f_{n,j-l}$ is an even function, the study of $f_{n,j-l}$ reduces on the interval $[0, n+1]$. Because $f'_{n,j-l}(x) = \frac{-x(g^2+x^2+(j-l)^2)^{\frac{3}{2}}\left(12g^2-3x^2-3(j-l)^2\right)}{(g^2+x^2+(j-l)^2)^5} < 0$ for $x \neq 0$ (since $c \geq 1$), then $f_{n,j-l}$ is a decreasing function over [0,n+1]. So,

$$\min_{0\leq j,k\leq n+1}\sum_{l=0}^{n+1}\sum_{p=0}^{n+1}(-C_{k-p}^{j-l}) = \min_{0\leq j\leq n+1}\sum_{l=0}^{n+1}\sum_{p=0}^{n+1}(-C_{p}^{j-l}) \tag{8.111}$$

Also, the function $g_{n,p}(x) = \frac{-x^2-p^2+2g^2}{[x^2+p^2+g^2]^{\frac{5}{2}}}$ is a decreasing function over the interval [0,n+1]. One deduces that

$$\min_{0\leq j\leq n+1}\sum_{l=0}^{n+1}\sum_{p=0}^{n+1}(-C_{p}^{j-l}) = \sum_{l=0}^{n+1}\sum_{p=0}^{n+1}(-C_{p}^{l}) \tag{8.112}$$

According to (8.111) − 8.112), it follows that

$$\min_{0\leq j,k\leq n+1}\sum_{l=0}^{n+1}\sum_{p=0}^{n+1}(-C_{k-p}^{j-l}) = \sum_{l=0}^{n+1}\sum_{p=0}^{n+1}(-C_{p}^{l}) \tag{8.113}$$

Since, $\min_l\{\min_p(-C_p^l)\} = \frac{1}{h^3}\frac{2g^2-2(n+1)^2}{(g^2+2(n+1)^2)^{\frac{5}{2}}} = \frac{1}{h^3}\frac{2}{(n+1)^3}\frac{c^2-1}{(c^2+2)^{\frac{5}{2}}}$, then for $l,p = 0,1,...,n+1$,

$$\frac{1}{h^3}\frac{2}{(n+1)^3}\frac{c^2-1}{(c^2+2)^{\frac{5}{2}}} \leq \frac{1}{h^3}\frac{2g^2-l^2-p^2}{(g^2+l^2+p^2)^{\frac{5}{2}}} = -C_p^l$$

so,

$$2\left(\frac{c^2-1}{(c^2+2)^{\frac{5}{2}}}\right)(n+2)^2 \leq \frac{1}{h^3}\sum_{l=0}^{n+1}\sum_{p=0}^{n+1}\frac{2g^2-l^2-p^2}{(g^2+l^2+p^2)^{\frac{5}{2}}} = \sum_{l=0}^{n+1}\sum_{p=0}^{n+1}(-C_p^l) \tag{8.114}$$

It follows from (8.100) − (8.101) − (8.113) − (8.114) that the spectral radius of the matrix A_{d_n} grows as $(n+2)^2$ (if $c > 1$).

Case: $c = 1$. One has:

$$\sum_{l=0}^{n+1}\sum_{p=0}^{n+1}\left(\frac{2(n+1)^2-(j-l)^2-(k-p)^2}{[(n+1)^2+(j-l)^2+(k-p)^2]^{\frac{5}{2}}}\right)= \tag{8.115}$$

$$\sum_{p=0}^{n+1}\frac{2(n+1)^2-(j-n-1)^2-(k-p)^2}{[(n+1)^2+(j-n-1)^2+(k-p)^2]^{\frac{5}{2}}}+\sum_{l=0}^{n}\sum_{p=0}^{n+1}\frac{2(n+1)^2-(j-l)^2-(k-p)^2}{[(n+1)^2+(j-l)^2+(k-p)^2]^{\frac{5}{2}}}$$

One shows as in the **case** $c > 1$ that

$$\min_{0\leq j,k\leq n+1}\sum_{p=0}^{n+1}\left(\frac{2(n+1)^2-(j-n-1)^2-(k-p)^2}{[(n+1)^2+(j-n-1)^2+(k-p)^2]^{\frac{5}{2}}}\right)=\sum_{p=0}^{n+1}\frac{(n+1)^2-p^2}{[2(n+1)^2+p^2]^{\frac{5}{2}}} \tag{8.116}$$

and

$$\min_{0\leq j,k\leq n+1}\sum_{l=0}^{n}\sum_{p=0}^{n+1}\left(\frac{2(n+1)^2-(j-l)^2-(k-p)^2}{[(n+1)^2+(j-l)^2+(k-p)^2]^{\frac{5}{2}}}\right)= \tag{8.117}$$

$$\sum_{l=0}^{n}\sum_{p=0}^{n+1}\left(\frac{2(n+1)^2-l^2-p^2}{[(n+1)^2+l^2+p^2]^{\frac{5}{2}}}\right)$$

Also here, as in the studies done in **case** $c > 1$, one has

$$\frac{(n+1)^2-n^2}{[2(n+1)^2+n^2]^{\frac{5}{2}}}\leq\frac{(n+1)^2-p^2}{[2(n+1)^2+p^2]^{\frac{5}{2}}}\ \ \forall p=0,1,...,n \tag{8.118}$$

and

$$\frac{(n+1)^2-n^2}{[2(n+1)^2+n^2]^{\frac{5}{2}}}\leq\frac{2(n+1)^2-l^2-p^2}{[(n+1)^2+l^2+p^2]^{\frac{5}{2}}}\ \ \forall p=0,1,...,n+1\ \text{ and }\ \forall l=0,1,...,n \tag{8.119}$$

From (8.116) − (8.118),

$$\frac{(2n+1)(n+1)^3}{[2(n+1)^2+n^2]^{\frac{5}{2}}}(n+2)\leq\min_{0\leq j,k\leq n+1}\frac{1}{h^3}\sum_{p=0}^{n+1}\frac{2(n+1)^2-(j-n-1)^2-(k-p)^2}{[(n+1)^2+(j-n-1)^2+(k-p)^2]^{\frac{5}{2}}} \tag{8.120}$$

and from (8.117) − (8.119),

$$\frac{(2n+1)(n+1)^4}{[2(n+1)^2+n^2]^{\frac{5}{2}}}(n+2)\leq\min_{0\leq j,k\leq n+1}\frac{1}{h^3}\sum_{l=0}^{n}\sum_{p=0}^{n+1}\frac{2(n+1)^2-(j-l)^2-(k-p)^2}{[(n+1)^2+(j-l)^2+(k-p)^2]^{\frac{5}{2}}} \tag{8.121}$$

since

$$\frac{(2n+1)(n+1)^3}{[2(n+1)^2+n^2]^{\frac{5}{2}}}(n+2)\underset{n>>1}{\sim}\frac{2}{9\sqrt{3}}\ \text{ and }\ \frac{(2n+1)(n+1)^4}{[2(n+1)^2+n^2]^{\frac{5}{2}}}(n+2)\underset{n>>1}{\sim}\frac{2}{9\sqrt{3}}(n+2) \tag{8.122}$$

one deduces from (8.115) − (8.120) − (8.121) − (8.122) that

$$\frac{2}{9\sqrt{3}}(n+3) \lesssim \min_{j,k} \sum_{l=0}^{n+1}\sum_{p=0}^{n+1}(-C_{k-p}^{j-l}). \tag{8.123}$$

It follows from (8.100) – (8.101) – (8.123) that the spectral radius of A_{d_n} more rapidly grows that $n+3$.

Conclusion: one deduces from the above studies that the spectral radius of A_{d_n} grows as $(n+2)^2$ if $c > 1$ (respectively more rapidly than $n+3$ if $c = 1$). Then, the condition number of the matrix A_{d_n} more rapidly grows than $(n+2)^2$ if $c > 1$ (respectively $n+3$ if $c = 1$). So, the matrix A_{d_n} is ill-conditioned.

Case 5: Gaussian ($c \geq \sqrt{2}$)
For $j, l, p, k = 0, 1, ..., n+1$:

$$C_{k-p}^{j-l} = \frac{4}{h^2 g^4}[(j-l)^2 + (k-p)^2 - g^2]e^{-\frac{(j-l)^2+(k-p)^2}{g^2}} \leq 0 \quad \text{since } c \geq \sqrt{2}, \tag{8.124}$$

hence one shows as in the case of **Inverse Multiquadric** that

$$\rho(T_{d_n}) = \rho(-T_{d_n}) \leq \rho(A_{d_n}) \tag{8.125}$$

and

$$\min_{0\leq j,k\leq n+1} \sum_{l=0}^{n+1}\sum_{p=0}^{n+1}(-C_{k-p}^{j-l}) \leq \rho(-T_{d_n}) \leq \max_{0\leq j,k\leq n+1} \sum_{l=0}^{n+1}\sum_{p=0}^{n+1}(-C_{k-p}^{j-l}). \tag{8.126}$$

For $j, k = 0, 1, ..., n+1$,

$$\sum_{l=0}^{n+1}\sum_{p=0}^{n+1}(-C_{k-p}^{j-l}) = \frac{4}{h^2 g^4}\sum_{l=0}^{n+1}\sum_{p=0}^{n+1}\left([g^2 - (j-l)^2 - (k-p)^2]e^{-\frac{(j-l)^2+(k-p)^2}{g^2}}\right) \tag{8.127}$$

Case: $c > \sqrt{2}$. Let us study the function: $f_{n,j-l}(x) = [g^2 - (j-l)^2 - x^2]e^{-\frac{(j-l)^2+x^2}{g^2}}$ over the domain $[-n-1, n+1]$. It is obvious that $f_{n,j-l}$ is an even function, whence study of $f_{n,j-l}$ over $[0, n+1]$. Because $f'_{n,j-l}(x) = -2x[1 + \frac{1}{g^2}(g^2 - (j-l)^2 - x^2)]e^{-\frac{(j-l)^2+x^2}{g^2}} < 0$ for $x \neq 0$ (since $c > \sqrt{2}$), then $f_{n,j-l}$ is decreasing over $[0, n+1]$. One deduces that,

$$\min_{0\leq j,k\leq n+1} \sum_{l=0}^{n+1}\sum_{p=0}^{n+1}(-C_{k-p}^{j-l}) = \min_{0\leq j\leq n+1} \sum_{l=0}^{n+1}\sum_{p=0}^{n+1}(-C_{p}^{j-l}) \tag{8.128}$$

Also here, the function $g_{n,p}(x) = [g^2 - p^2 - x^2]e^{-\frac{p^2+x^2}{g^2}}$ is decreasing over $[0, n+1]$, then

$$\min_{0\leq j\leq n+1} \sum_{l=0}^{n+1}\sum_{p=0}^{n+1}(-C_{p}^{j-l}) = \sum_{l=0}^{n+1}\sum_{p=0}^{n+1}(-C_{p}^{l}) \tag{8.129}$$

According to (8.128) – (8.129), it follows that

$$\min_{0\leq j,k\leq n+1}\sum_{l=0}^{n+1}\sum_{p=0}^{n+1}(-C_{k-p}^{j-l})=\sum_{l=0}^{n+1}\sum_{p=0}^{n+1}(-C_p^l) \tag{8.130}$$

On the other side, $\min_l\{\min_p(-C_p^l)\}=\frac{1}{h^2g^4}[g^2-2(n+1)^2]e^{-\frac{2(n+1)^2}{g^2}}=\frac{1}{c^4}(c^2-2)e^{-\frac{2}{c^2}}$, then for $l,p=0,1,...,n+1$,

$$\frac{(c^2-2)e^{-\frac{2}{c^2}}}{c^4}\leq\frac{1}{h^2g^4}[g^2-l^2-p^2]e^{-\frac{l^2+p^2}{g^2}}=-C_p^l$$

so

$$\left(\frac{(c^2-2)e^{-\frac{2}{c^2}}}{c^4}\right)(n+2)^2\leq\frac{1}{h^2g^4}\sum_{l=0}^{n+1}\sum_{p=0}^{n+1}[g^2-l^2-p^2]e^{-\frac{l^2+p^2}{g^2}}=\sum_{l=0}^{n+1}\sum_{p=0}^{n+1}(-C_p^l) \tag{8.131}$$

It follows from (8.100) – (8.101) – (8.130) – (8.131) that the spectral radius of the matrix A_{d_n} grows as $(n+2)^2$ (if $c>\sqrt{2}$).

Case: $c=\sqrt{2}$. First of all, one has

$$\sum_{l=0}^{n+1}\sum_{p=0}^{n+1}\left([g^2-(j-l)^2-(k-p)^2]e^{-\frac{(j-l)^2+(k-p)^2}{g^2}}\right)=\sum_{p=0}^{n+1}[g^2-(j-n-1)^2 \tag{8.132}$$

$$-(k-p)^2]e^{-\frac{(j-n-1)^2+(k-p)^2}{g^2}}+\sum_{l=0}^{n}\sum_{p=0}^{n+1}\left([g^2-(j-l)^2-(k-p)^2]e^{-\frac{(j-l)^2+(k-p)^2}{g^2}}\right)$$

Furthermore, one shows as in **Inverse Multiquadric case** (**case** $c>1$) that

$$\min_{0\leq j,k\leq n+1}\sum_{p=0}^{n+1}\left(2(n+1)^2-(j-n-1)^2-(k-p)^2]e^{-\frac{(j-n-1)^2+(k-p)^2}{g^2}}\right)= \tag{8.133}$$

$$\sum_{p=0}^{n+1}\left([(n+1)^2-p^2]e^{-\frac{(n+1)^2+p^2}{g^2}}\right)$$

and

$$\min_{0\leq j,k\leq n+1}\sum_{l=0}^{n}\sum_{p=0}^{n+1}\left([g^2-(j-l)^2-(k-p)^2]e^{-\frac{(j-l)^2+(k-p)^2}{g^2}}\right)= \tag{8.134}$$

$$\sum_{l=0}^{n}\sum_{p=0}^{n+1}\left([g^2-l^2-p^2]e^{-\frac{l^2+p^2}{g^2}}\right)$$

where $g=\sqrt{2}(n+1)$. Also here, as in the studies done in **case** $c>\sqrt{2}$, one has

$$[(n+1)^2-n^2]e^{-\frac{(n+1)^2+n^2}{g^2}}\leq[(n+1)^2-p^2]e^{-\frac{(n+1)^2+p^2}{g^2}}\ \ \forall p=0,1,...,n \tag{8.135}$$

and

$$(8.136)\ \ [(n+1)^2-n^2]e^{-\frac{(n+1)^2+n^2}{g^2}} \leq [g^2-l^2-p^2]e^{-\frac{l^2+p^2}{g^2}}\ \ \forall p=0,1,...,n+1 \text{ and } \forall l=0,1,...,n.$$

From $(8.133)-(8.135)$,

$$(8.137)\qquad [(n+1)^2-n^2]\frac{n+2}{4(n+1)^2}e^{-\frac{(n+1)^2+n^2}{g^2}} \leq \min_{j,k}\frac{1}{h^2g^4}\sum_{p=0}^{n+1}[g^2-(j-n-1)^2$$

$$-(k-p)^2]e^{-\frac{(j-n-1)^2+(k-p)^2}{g^2}}$$

and from $(8.134)-(8.136)$,

$$(8.138)\qquad [(n+1)^2-n^2]\frac{n+2}{4(n+1)}e^{-\frac{(n+1)^2+n^2}{g^2}} \leq \min_{j,k}\frac{1}{h^2g^4}\sum_{l=0}^{n}\sum_{p=0}^{n+1}[g^2-(j-l)^2$$

$$-(k-p)^2]e^{-\frac{(j-l)^2+(k-p)^2}{g^2}}$$

Because

$$(8.139)\qquad [(n+1)^2-n^2]\frac{n+2}{4(n+1)^2}e^{-\frac{(n+1)^2+n^2}{g^2}} \underset{n>>1}{\gtrsim} \frac{1}{2e}$$

and

$$(8.140)\qquad [(n+1)^2-n^2]\frac{n+2}{4(n+1)}e^{-\frac{(n+1)^2+n^2}{g^2}} \underset{n>>1}{\gtrsim} \frac{1}{2e}(n+2)$$

One deduces from $(8.132)-(8.137)-(8.138)-(8.139)-(8.140)$ that

$$(8.141)\qquad \frac{1}{2e}(n+3) \lesssim \min_{j,k}\sum_{l=0}^{n+1}\sum_{p=0}^{n+1}(-C_{k-p}^{j-l}).$$

It follows from $(8.125)-(8.126)-(8.141)$ that the spectral radius of A_{d_n} more rapidly grows than $n+3$.

Conclusion: According to the studies done in cases $c>\sqrt{2}$ and $c=\sqrt{2}$, the spectral radius of A_{d_n} grows as $(n+2)^2$ if $c>\sqrt{2}$ (respectively $n+3$ if $c=\sqrt{2}$). Then, the condition number of the matrix A_{d_n} more rapidly grows than $(n+2)^2$ if $c>\sqrt{2}$ (respectively $n+3$ if $c=\sqrt{2}$). So, the matrix A_{d_n} is ill-conditioned.

Remark 8.2.1. *It follows from the studies done in this section that the spectral radius of the collocation matrices grows as the size of the matrices if $c>\sqrt{2}$.*

Remark 8.2.2. *The following property was fundamental in **Inverse Multiquadric** and **Gaussian** cases: For two real functions f and g defined on the same interval $I\subset\mathbb{R}$,*

$$\inf_{x\in I}f(x)+\inf_{x\in I}g(x) \leq \inf_{x\in I}(f(x)+g(x)).$$

Conclusion

We have studied in detail the preconditioners and spectral radii of collocation matrices A_{d_n} approximating the Elliptic Boundary Value Problems (8.23) – (8.24) by imposing some constraints on the shape parameter "c" figuring in the radial basis functions $\phi(t)$. Our future researches will consist to delete these requirements and to look for another theory in order to study the spectral radii of these collocation matrices. Furthermore, we exploit in chapter 9 the sequence of symmetric bolck Toeplitz matrices with symmetric Toeplitz blocks (SBTMSTB) $\{T_{d_n}\}_n$ which is equally distributed and equally localized as the sequence of collocation matrices $\{A_{d_n}\}_n$ in order to present an application (with some numerical results) of the PCG method for the SBTMSTB with unbounded generating functions.

Chapter Nine

Application of the PCG Method to SBTMSTB with Unbounded Generating Functions

In this chapter, we introduce and discuss the PCG method for the solution of linear systems associated with ill-conditioned symmetric block Toeplitz matrices with symmetric Toeplitz blocks (SBTMSTB) generated by unbounded functions. Furthermore, we perform some numerical experiments which confirm the theoretical results.

9.1 Introduction

The systems of linear equations associated with Toeplitz (block Toeplitz) matrices arise in many one-dimensional (two-dimensional) digital signal processing applications, such as linear prediction and estimation [86], [97], and [98], image restoration [46], and the discretization of constant-coefficient partial differential equations. In order to solve the system of linear equations $A_{d_n}x = b$, where $\{A_{d_n}\}_n$ is the sequence of collocation matrices approximating elliptic boundary value problems, it useful to find a sequence $\{T_{d_n}(s)\}_n$ of $d_n \times d_n$ symmetric Toeplitz (block Toeplitz) matrices which is equally distrituted and equally localized as the sequence $\{A_{d_n}\}_n$. Hence, solve the Toeplitz (block Toeplitz) system $T_{d_n}x = b$, by direct methods, such as Levinson-type algorithms, requires $O(d_n^3 d_n^2)$ operations [8], [113], [169]. The spectral properties of these matrices, which are related to the behavior of the generating function s, have been well understood and deeply studied in this century (see for instance [77, 110, 2, 170, 162, 167, 153]). More Recently, there has been active research on the application of iterative methods such as the preconditioned conjugate gradient (PCG) method to the solution of Toeplitz systems. The most successful preconditioners for the case where s is strictly positive have been devised in the algebras of circulant, Hartley, and τ matrices [37, 12, 11]. For the nonnegative case, under the assumption of zeros of even orders, the only "optimal" preconditioners [7] are those chosen in the τ algebra [52] and in the band Toeplitz class [31, 56, 42, 120]. In particular, this last preconditioning strategy has been the most flexible and versatile in allowing one to treat also nondefinite [118], the block [120], and the non-Hermitian [35] cases and the case of zeros of any order [140]. This is the reason for which we focus our attention on the Toeplitz preconditioning: therefore we consider the positive definite matrix $T_{d_n}(g)$ generated by a nonnegative, nonzeros function g, and then, on following a known strategy [56, 31], the preconditioned matrix takes the form $T_{d_n}(g)^{-1}T_{d_n}(s)$. The proposed preconditioning techniques can be easily generalized to block Toeplitz matrices. Since both $T_{d_n}(g)^{-1}w$ and $T_{d_n}(s)w$, where w denotes an arbitrary vector of length d_n, can be performed with $O(d_n \log d_n)$ operations via fast Fourier transform, the computational complexity per PCG iteration is $O(d_n \log d_n)$ only. The PCG method can be much more attractive than direct methods for solving Toeplitz systems if it converges fast. Furthermore, it is important to recall that the PCG method is more appropriated for this special type of preconditioned matrix (often when the generating function s of the Toeplitz (block Toeplitz) matrix is unbounded) and it does not contradict the preconditioned ma-

trix $P_{d_n}^{-1}T_{d_n}(s)$ defined in chapter 8 which can be exploited by the symmetric quasi-minimal residual (QMR) algorithm (see [26, 65, 107]) or the Bi-conjugate gradient (BCG) algorithm (a direct generalization of the classical conjugate gradient method of Hestenes and Stiefel). Indeed, a very simple no-look-ahead version of the coupled two-term QMR algorithm was derived by Freud and Szeto in [66]. They have shown that the no-look-ahead QMR iterates can be obtained from the classical biconjugate gradient (BCG) algorithm (see [94]) by performing one additional vector update and a few scalar updates at each BCG iteration. The resulting algorithm is called "QMR-from-BCG".

The convergence rate of the PCG method depends on the eigenvalue distribution of the preconditioned matrix $T_{d_n}(g)^{-1}T_{d_n}(s)$ [6]. Generally speaking, the PCG method converges faster if $T_{d_n}(g)^{-1}T_{d_n}(s)$ has eigenvalues clustered to 1 and/or small condition number. Chan and Strang have proved that, for a Toeplitz matrix with a positive generating function in the Wiener class, the spectrum of the preconditioned matrix has eigenvalues clustered around unity except for a finite number of outliers.

In this chapter, by collecting known results, we show that the function s/g describes very precisely the spectrum of the family of matrices $\{T_{d_n}(g)^{-1}T_{d_n}(s)\}_n$ and consequently we introduce the concept of the generating function for such kind of preconditioned Toeplitz (block Toeplitz) matrices.

Our main results can be summarized as follows. Let $\{A_{d_n}\}_n$ be the sequence of $d_n \times d_n$ collocation matrices, find a sequence $\{T_{d_n}(s)\}_n$ of $d_n \times d_n$ symmetric Toeplitz (block Toeplitz) matrices such that the sequences $\{A_{d_n}\}_n$ and $\{T_{d_n}(s)\}_n$ are equally distributed and equally localized. Furthermore, find a preconditioner $T_{d_n}(g)$ generated by a nonnegative, nonzeros function g of $T_{d_n}(s)$ and study the asymptotic growth of the spectral radius of $T_{d_n}(g)^{-1}T_{d_n}(s)$ by PCG method for positive definite Toeplitz (block Toeplitz) matrices. Finally, perform some numerical experiments in order to give numerical evidence for the theoretical results.

9.2 Preliminary

In this section, we consider the Poisson equation (uni-dimension and two-dimensions cases) defined in the chapter 8 by

$$\begin{cases} u''(x) = f(x), & x \in \Omega = (0,1) \\ u(0) = a, \;\; u(1) = b \end{cases} \tag{9.1}$$

and

$$\begin{cases} \frac{\partial^2 u(x,y)}{\partial x^2} + \frac{\partial^2 u(x,y)}{\partial y^2} = f(x,y) & \text{for } (x,y) \in \Omega = (0,1)^2 \\ u(x,y) = g(x,y) & \text{if } (x,y) \in \partial\Omega \end{cases} \tag{9.2}$$

As it was shown in chapter 8, the associated linear system obtained by the method of radial basis functions is defined as follows

$$A_{d_n} v = \tilde{f} \tag{9.3}$$

where in **uni-dimension case**,

$$A_{n+2} = [a_{ij}]_{i,j=0}^{n+1}, \;\; \tilde{f}_{n+2} = [a, f_1, f_2, ..., f_n, b]^T \;\text{ and }\; v = [v_0, v_1, ..., v_{n+1}]^T,$$

with

$$\begin{cases} a_{0,j} = \phi_{0,j}, \;\; a_{n+1,j} = \phi_{n+1,j} & j = 0, 1, ..., n+1 \\ a_{i,j} = \phi''_{i,j}, & i = 1, 2, ..., n; \; j = 0, 1, ..., n+1. \end{cases}$$

Furthermore, setting $T_{n+2}(s) = [c_{j-k}]_{j,k=0}^{n+1}$ where

$$(9.4)\qquad c_{j-k}=\phi''_{jk} \;=\; \begin{cases} \frac{1}{h}\frac{g^2}{((j-k)^2+g^2)^{\frac{3}{2}}} & \text{Multiquadric (MQ)} \\ \frac{1}{h^2}\frac{(j-k)^2-2g^2}{((j-k)^2+g^2)^{\frac{5}{2}}} & \text{Inverse Multiquadric (IMQ)} \\ \frac{2}{h^2g^2}\left(\frac{2(j-k)^2}{g^2}-1\right)e^{-\frac{(j-k)^2}{g^2}} & \text{Gaussian} \end{cases}$$

with $g=c/h$, it follows from (9.4) that the generating function of the symmetric Toeplitz matrix $T_{n+2}(s)$ is given by:

$$(9.5)\qquad s(x) \;=\; c_0+2\sum_{k=1}^{\infty}c_k\cos(2kx) \qquad x\in(0,\pi).$$

The Fourier coefficients related to $s(x)$ are given by

$$c_k=\frac{1}{\pi}\int_Q s(x)e^{-\hat{i}(kx)}dx, \quad I=[0,\pi].$$

Remark 9.2.1. *The matrix sequences $\{A_{n+2}\}_n$ and $\{T_{n+2}\}_n$ are equally distributed and equally localized. (see chapter 8).*

In two-dimensions case, $d_n=(n+2)^2$ and one has

$$A_{d_n}=\begin{bmatrix} A_{0-0}^{(n+2)} & A_{0-1}^{(n+2)} & \dots & A_{0-n}^{(n+2)} & A_{0-(n+1)}^{(n+2)} \\ A_{1-0}^{(n+2)} & A_{1-1}^{(n+2)} & \dots & \dots & A_{1-(n+1)}^{(n+2)} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ A_{n-0}^{(n+2)} & A_{n-1}^{(n+2)} & \dots & A_{n-n}^{(n+2)} & A_{n-(n+1)}^{(n+2)} \\ A_{(n+1)-0}^{(n+2)} & A_{(n+1)-1}^{(n+2)} & \dots & A_{(n+1)-n}^{(n+2)} & A_{(n+1)-(n+1)}^{(n+2)} \end{bmatrix};\quad v=\begin{bmatrix} v^{(0)} \\ v^{(1)} \\ \vdots \\ v^{(n)} \\ v^{(n+1)} \end{bmatrix};\quad \tilde{f}=\begin{bmatrix} f^{(0)} \\ f^{(1)} \\ \vdots \\ f^{(n)} \\ f^{(n+1)} \end{bmatrix}$$

for $j=0,1,...,n+1$,

$$(9.6)\qquad A_{0-j}^{(n+2)}=[\phi_{l-p}^{0-j}]_{l,p=0}^{n+1};\quad A_{(n+1)-j}^{(n+2)}=[\phi_{l-p}^{(n+1)-j}]_{l,p=0}^{n+1}$$

for $j=1,2,...,n$ and $l=0,1,...,n+1$,

$$(9.7)\qquad A_{j-l}^{(n+2)}=\begin{bmatrix} \phi_{0-0}^{j-l} & \phi_{0-1}^{j-l} & \dots & \phi_{0-n}^{j-l} & \phi_{0-(n+1)}^{j-l} \\ C_{1-0}^{j-l} & C_{1-1}^{j-l} & \dots & C_{1-n}^{j-l} & C_{1-(n+1)}^{j-l} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ C_{n-0}^{j-l} & C_{n-1}^{j-l} & \dots & C_{n-n}^{j-l} & C_{n-(n+1)}^{j-l} \\ \phi_{(n+1)-0}^{j-l} & \phi_{(n+1)-1}^{j-l} & \dots & \phi_{(n+1)-n}^{j-l} & \phi_{(n+1)-(n+1)}^{j-l} \end{bmatrix}=$$

$$\begin{bmatrix} \phi_0^{j-l} & \phi_1^{j-l} & \dots & \phi_n^{j-l} & \phi_{(n+1)}^{j-l} \\ C_1^{j-l} & C_0^{j-l} & \dots & C_{n-1}^{j-l} & C_n^{j-l} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ C_n^{j-l} & C_{n-1}^{j-l} & \dots & C_0^{j-l} & C_1^{j-l} \\ \phi_{n+1}^{j-l} & \phi_n^{j-l} & \dots & \phi_1^{j-l} & \phi_0^{j-l} \end{bmatrix}$$

with for every $k = 0, 1, ..., n+1$

$$v^{(k)} = (v_{k,0}, v_{k,1}, ..., v_{k,n+1})^T; \quad f^{(0)} = (g_{0,0}, g_{0,1}, ..., g_{0,n+1})^T; \quad f^{(n+1)} = (g_{n+1,0}, ..., g_{n+1,n+1})^T$$

for $p = 1, 2, ..., n$,

$$f^{(p)} = (g_{p,0}, f_{p,1}, ..., f_{p,n}, g_{p,n+1})^T$$

for $j = 1, 2, ..., n$ and $l = 0, 1, ..., n+1$,

$$A_{j-l}^{(n+2)} = T_{j-l}^{(n+2)} + \Delta_{j-l}^{(n+2)} \tag{9.8}$$

where

$$\Delta_{j-l}^{(n+2)} = \begin{bmatrix} \phi_0^{j-l} - C_0^{j-l} & \phi_1^{j-l} - C_1^{j-l} & \dots & \phi_{n+1}^{j-l} - C_{n+1}^{j-l} \\ 0 & \dots & & 0 \\ \vdots & & & \vdots \\ 0 & \dots & & 0 \\ \phi_{n+1}^{j-l} - C_{n+1}^{j-l} & \phi_n^{j-l} - C_n^{j-l} & \dots & \phi_0^{j-l} - C_0^{j-l} \end{bmatrix}, \quad T_{j-l}^{(n+2)} = [C_{k-p}^{j-l}]_{k,p=0}^{n+1} \tag{9.9}$$

On the other side, for $j, k, l, p = 0, 1, ..., n+1$, let us define the symmetric Toeplitz matrices $T_{j-l}^{(n+2)}$ and whose Fourier coefficients are given by

$$C_{k-p}^{j-l} = \begin{cases} \frac{1}{h}\frac{1}{\sqrt{(j-l)^2+(k-p)^2+g^2}} + \frac{1}{h}\frac{g^2}{[(j-l)^2+(k-p)^2+g^2]^{\frac{3}{2}}} & \text{(MQ)} \\ \frac{1}{h^3}\frac{(j-l)^2+(k-p)^2-2g^2}{[(j-l)^2+(k-p)^2+g^2]^{\frac{5}{2}}} & \text{(IMQ)} \\ \frac{4}{h^2 g^4}[(j-l)^2+(k-p)^2-g^2]e^{-\frac{(j-l)^2+(k-p)^2}{g^2}} & \text{(Gaussian)} \end{cases} \tag{9.10}$$

Setting $T_{d_n} = [T_{j-l}^{(n+2)}]_{j,l=0}^{n+1}$, then T_{d_n} is a symmetric block Toeplitz matrix with symmetric Toeplitz blocks and whose the generating function is defined as

$$s(x,y) = c_{0,0} + 2\sum_{k=1}^{\infty} c_{0,k}\left(\cos(2kx) + \cos(2ky)\right) + 4\sum_{k=1}^{\infty}\sum_{j=1}^{\infty} c_{k,j}\cos(2kx)\cos(2jy) \tag{9.11}$$

for all $(x, y) \in (0, \pi) \times (0, \pi)$. The Fourier coefficients related to $s(x, y)$ are given by

$$c_{j,k} = \frac{1}{\pi^2}\int_Q s(x,y)e^{-\hat{i}(jx+ky)}dxdy, \quad Q = [0,\pi] \times [0,\pi]$$

and setting $g = c/h$, it follows from (9.10) that, for $j, k = 0, 1, 2, ..., n+1$,

$$c_{j,k} = \begin{cases} \frac{1}{h}\frac{1}{\sqrt{j^2+k^2+g^2}} + \frac{1}{h}\frac{g^2}{[j^2+k^2+g^2]^{\frac{3}{2}}} & \text{(MQ)} \\ \frac{1}{h^3}\frac{j^2+k^2-2g^2}{[j^2+k^2+g^2]^{\frac{5}{2}}} & \text{(IMQ)} \\ \frac{4}{h^2 g^4}[j^2+k^2-g^2]e^{-\frac{j^2+k^2}{g^2}} & \text{(Gaussian)} \end{cases} \tag{9.12}$$

Remark 9.2.2. *When choosing the form parameter "$c > \sqrt{2}$", it follows from relations (9.4) − (9.12) that the Fourier coefficients of the integrable function s are all positive (in Multiquadric case) and all negative (for Inverse Multiquadric and Gaussian cases). Because in the literature, most of the globally defined RBFs are only conditionally positive definite,*

without loss of this generality, we suppose that the matrices T_{d_n} are positive definite (in Multiquadric case) and negative definite (in Inverse Multiquadric and Gaussian cases: indeed, the smallest eigenvalue of these symmetric matrices is negative) since the matrix sequences $\{A_{d_n}\}_n$ and $\{T_{d_n}\}_n$ are equally distributed and equally localized. Otherwise, consider a positive number α (respectively a negative number β) such that $s+\alpha$ is essentially positive in Multiquadric case (respectively $s+\beta$ is essentially negative in Inverse Multiquadric and Gaussian cases). Therefore the Toeplitz matrices $T_{d_n}(s+\alpha)$ are positive definite (in Multiquadric case) and $T_{d_n}(s+\beta)$ are negative definite (in Inverse Multiquadric and Gaussian cases). Furthermore, the eigenvalues of $T_{d_n}(s+\alpha)$ (respectively $T_{d_n}(s+\beta)$) are the eigenvalues of $T_{d_n}(s)$ shifted according to the constant α (respectively β).

In the following we suppose that the shape parameter "c" appearing in the radial basis function $\phi(t)$ is strictly greater than $\sqrt{2}$.

Before starting the study of the asymptotical behavior of generating functions $s(x)$ and $s(x,y)$ first, let us recall the following results due to Riemann (or Riemann-Lebesgue) for the integrable functions.

Theorem 9.2.1. *(Riemann or Riemann-Lebesgue). Let $[a,b]$ be a bounded closed interval of $\mathbb{R}$. Then, any continuous function $f:[a,b]\to\mathbb{R}$ is Riemann-Lebesgue integrable on $[a,b]$.*

Theorem 9.2.2. *(Riemann or Riemann-Lebesgue). One suppose that the real-valued function $f:[a,b]\to\mathbb{R}$ is integrable over $[a,b]$. Let us consider the regular grid points $a=x_0<x_1<\cdots<x_n=b$ $(n>1)$ of step $h=\frac{b-a}{n}=x_i-x_{i-1}$ $(1\leq i\leq n)$ and let us set $I_n=\sum_{i=1}^{n} f(a+x_i)(x_i-x_{i-1})$. Then the real sequence of general term I_n converges in $\mathbb{R}$ and its limit, denoted $\int_a^b f(x)dx$ is called definite integral of f over $[a,b]$.*

9.3 Asymptotical behavior of generating functions $s(x)$ and $s(x,y)$

The purpose of this section is to study the behavior of $s(x)$ and $s(x,y)$ over the domains I and Q respectively.

9.3.1 Toeplitz case

In this subsection, we study the behavior of $s(x)$ over the interval $I=[-\pi,\pi]$.

Multiquadric and Gaussian

Lemma 9.3.1. *The real-valued integrable function $s(x)$ is even and unbounded over the compact domain $[-\pi,\pi]$.*

Proof. First of all, for $k=0,1,...,n+1$, let us recall that: $x_k=\frac{k}{n+1}$, $h=\frac{1}{n+1}$ and $g=c/h$.

- **Case 1: Multiquadric (MQ).**

$$c_k=\frac{1}{h}\frac{g^2}{(k^2+g^2)^{\frac{3}{2}}}=\frac{g^3}{c(n+1)^3}\frac{1}{\left(\left(\frac{k}{n+1}\right)^2+c^2\right)^{\frac{3}{2}}}=\frac{c^2}{(c^2+x_k^2)^{\frac{3}{2}}}.$$

Since the function $x\mapsto\frac{c^2}{(c^2+x^2)^{\frac{3}{2}}}$ is positive and continuous over the domain $[0,1]$, it follows from Theorem 9.2.1 that it is Riemann-Lebesgue integrable, so

$$0<\int_0^1\frac{c^2}{(c^2+x^2)^{\frac{3}{2}}}dx=\alpha_0<\infty,$$

and according to Theorem 9.2.2, one has

$$\lim_{n\to\infty}\frac{1}{n+1}\sum_{k=1}^{n+1}\frac{c^2}{(c^2+x_k^2)^{\frac{3}{2}}}=\int_0^1\frac{c^2}{(c^2+x^2)^{\frac{3}{2}}}dx=\alpha_0.$$

Then, for $\epsilon=\alpha_0/2$, $\exists N_\epsilon\in\mathbb{N}$ such that

$$\begin{aligned} n>N_\epsilon &\Rightarrow \left|\frac{1}{n+1}\sum_{k=1}^{n+1}\frac{c^2}{(c^2+x_k^2)^{\frac{3}{2}}}-\alpha_0\right|<\frac{\alpha_0}{2}\\ &\Rightarrow \frac{\alpha_0}{2}(n+1)<\sum_{k=1}^{n+1}\frac{c^2}{(c^2+x_k^2)^{\frac{3}{2}}}<\frac{3\alpha_0}{2}(n+1)\\ &\Rightarrow \frac{\alpha_0}{2}(n+1)<\sum_{k=1}^{n+1}c_k<\frac{3\alpha_0}{2}(n+1). \end{aligned}$$

Then,

$$\sum_{k=1}^{n+1}c_k\sim(n+1). \tag{9.13}$$

From (9.13) and (9.5), one obtains

$$\lim_{x\to\pm\pi,0}s(x)=\infty.$$

Hence, $s(x)$ is unbounded over I. Since the Toeplitz matrix $T_{d_n}(s)$ is symmetric, one deduces from Lemma 1.3.2 that the function $s(x)$ is even.

• **Case 2: Gaussian**.

$$c_k=\frac{2}{h^2g^2}\left(\frac{2k^2}{g^2}-1\right)e^{-\frac{k^2}{g^2}}=\frac{-4}{c^4}\left(\frac{c^2}{2}-\left(\frac{k}{n+1}\right)^2\right)e^{-\frac{1}{c^2}\left(\frac{k}{n+1}\right)^2}=\frac{-4}{c^4}\left(\frac{c^2}{2}-x_k^2\right)e^{-\frac{1}{c^2}x_k^2}.$$

Since the function $x\mapsto\frac{-4}{c^4}\left(\frac{c^2}{2}-x^2\right)e^{-\frac{1}{c^2}x^2}$ is negative and continuous on the compact subset $[0,1]$, according to Theorem 9.2.1 it is Riemann integrable so,

$$\int_0^1\frac{-4}{c^4}\left(\frac{c^2}{2}-x^2\right)e^{-\frac{1}{c^2}x^2}dx=\alpha_1<0,$$

with $|\alpha_1|<\infty$. Furthermore, it follows from Theorem 9.2.2 that

$$\lim_{n\to\infty}\frac{1}{n+1}\sum_{k=1}^{n+1}\left\{\frac{-4}{c^4}\left(\frac{c^2}{2}-x_k^2\right)e^{-\frac{1}{c^2}x_k^2}\right\}=\int_0^1\frac{-4}{c^4}\left(\frac{c^2}{2}-x^2\right)e^{-\frac{1}{c^2}x^2}dx=\alpha_1.$$

Then, for $\epsilon=|\alpha_1|/2$, $\exists N_\epsilon\in\mathbb{N}$ such that

$$\begin{aligned} n>N_\epsilon &\Rightarrow \left|\frac{1}{n+1}\sum_{k=1}^{n+1}c_k-\alpha_1\right|<\frac{|\alpha_1|}{2}\\ &\Rightarrow \frac{3\alpha_1}{2}(n+1)<\sum_{k=1}^{n+1}c_k<\underbrace{\frac{\alpha_1}{2}}_{<0}(n+1) \end{aligned}$$

then

$$\sum_{k=1}^{n+1} c_k \sim -(n+1). \tag{9.14}$$

From (9.14) and (9.5), one obtains

$$\lim_{x \to \pm\pi, 0} s(x) = -\infty.$$

Hence, $s(x)$ is even and unbounded over I.

□

Case 3: Inverse Multiquadric

Lemma 9.3.2. *The real-valued integrable function $s(x)$ is even and bounded over the compact domain $[-\pi, \pi]$.*

Proof.

$$c_k = \frac{1}{h^2} \frac{k^2 - 2g^2}{(k^2+g^2)^{\frac{5}{2}}} = \frac{-(n+1)^4}{(n+1)^5} \frac{2c^2 - \left(\frac{k}{n+1}\right)^2}{\left(c^2 + \left(\frac{k}{n+1}\right)^2\right)^{\frac{5}{2}}} = \frac{1}{(n+1)} \frac{-2c^2 + x_k^2}{(c^2+x_k^2)^{\frac{5}{2}}}.$$

Then, the function $x \mapsto \frac{-2c^2+x^2}{(c^2+x^2)^{\frac{5}{2}}}$ is negative and continuous on the compact subset $[0,1]$, according to Theorem 9.2.1 it is Riemann-Lebesgue integrable then,

$$\int_0^1 \frac{-2c^2 + x^2}{(c^2+x^2)^{\frac{5}{2}}} dx = \alpha_2 < 0,$$

with $|\alpha_2| < \infty$. Furthermore, one deduces from Theorem 9.2.2 that

$$\lim_{n\to\infty} \frac{1}{n+1} \sum_{k=1}^{n+1} \frac{-2c^2 + x_k^2}{(c^2+x_k^2)^{\frac{5}{2}}} = \int_0^1 \frac{-2c^2 + x^2}{(c^2+x^2)^{\frac{5}{2}}} dx = \alpha_2.$$

that is

$$\lim_{n\to\infty} \sum_{k=1}^{n+1} c_k = \alpha_2.$$

Because the Fourier coefficients c_k of $s(x)$ are all negative, one deduces that

$$\sum_{k=0}^{\infty} |c_k| = \left| \sum_{k=0}^{\infty} c_k \right| = |c_0 + \alpha_2|.$$

So, $s(x)$ belongs to the Wiener class of functions hence, $s(x) \in L^\infty[0,\pi]$. Since the Toeplitz matrix $T_{d_n}(s)$ is symmetric, according to Lemma 1.3.2, it follows that $s(x)$ is even over I □

9.3.2 Block Toeplitz case

Throughout this subsection, we study the behavior of the function $s(x,y)$ over the domain $Q = [-\pi,\pi] \times [-\pi,\pi]$.

Lemma 9.3.3. *The real-valued integrable function $s(x,y)$ is even and unbounded over the domain Q.*

Proof. In this proof, we treat separately the cases: Multiquadric, Inverse Multiquadric and Gaussian.

Case 4: Multiquadric

$$c_{j,k} = \frac{1}{h}\frac{1}{\sqrt{j^2+k^2+g^2}} + \frac{1}{h}\frac{g^2}{(j^2+k^2+g^2)^{\frac{3}{2}}} = c_{j,k}^{(1)} + c_{j,k}^{(2)}.$$

Here, one deduces from **case** 1 above that

$$\sum_{k=1}^{n+1} c_{0,k}^{(2)} \sim n+1,$$

then

$$\lim_{n\to\infty} \sum_{k=1}^{n+1} c_{0,k}^{(2)} = \infty. \tag{9.15}$$

Because the Fourier coefficients $c_{j,k}$ of $s(x,y)$ are all positive, then

$$\sum_{k=1}^{n+1} c_{0,k}^{(2)} \leq c_{0,0} + 4\left(\sum_{k=1}^{n+1} c_{0,k}^{(2)} + \sum_{k=1}^{n+1}\sum_{j=1}^{n+1} c_{j,k}\right). \tag{9.16}$$

From (9.11) – (9.15) – (9.16), it follows that

$$\lim_{(x,y)\to(\pm\pi,\pm\pi)} s(x,y) = \infty, \quad \lim_{(x,y)\to(0,\pm\pi)} s(x,y) = \infty$$

and

$$\lim_{(x,y)\to(\pm\pi,0)} s(x,y) = \infty, \quad \lim_{(x,y)\to(0,0)} s(x,y) = \infty.$$

Hence, $s(x,y)$ is unbounded on the domain Q. On the other hand, because $T_{d_n}(s)$ is symmetric, it follows from Lemma 1.3.2 (a generalization) that $s(x,y)$ when defined over the domain Q is even, in the sense that for every $(x,y) \in Q$ such that s is defined in (x,y),

$$s(-x,-y) = s(-x,y) = s(x,-y) = s(x,y).$$

Case 5: Inverse Multiquadric

$$c_{j,k} = \frac{1}{h^3}\frac{j^2+k^2-2g^2}{(j^2+k^2+g^2)^{\frac{5}{2}}},$$

then

$$c_{0,k} = \frac{1}{h^3}\frac{k^2-2g^2}{(k^2+g^2)^{\frac{5}{2}}} = \frac{(n+1)^5}{(n+1)^5}\frac{-2c^2+\left(\frac{k}{n+1}\right)^2}{(c^2+\left(\frac{k}{n+1}\right)^2)^{\frac{5}{2}}} = \frac{-2c^2+x_k^2}{(c^2+x_k^2)^{\frac{5}{2}}}.$$

One shows as in **case** 2 that

$$\sum_{k=1}^{n+1} c_{0,k} \sim -(n+1),$$

then

$$\lim_{n\to\infty} \sum_{k=1}^{n+1} c_{0,k} = -\infty. \tag{9.17}$$

Since the Fourier coefficients $c_{j,k}$ of $s(x,y)$ are all negative, one has

$$c_{0,0} + 4\left(\sum_{k=1}^{n+1} c_{0,k} + \sum_{k=1}^{n+1}\sum_{j=1}^{n+1} c_{j,k}\right) \leq \sum_{k=1}^{n+1} c_{0,k}. \tag{9.18}$$

From (9.11) – (9.17) – (9.18), it follows that

$$\lim_{(x,y)\to(\pm\pi,\pm\pi)} s(x,y) = -\infty, \quad \lim_{(x,y)\to(0,\pm\pi)} s(x,y) = -\infty$$

and

$$\lim_{(x,y)\to(\pm\pi,0)} s(x,y) = -\infty, \quad \lim_{(x,y)\to(0,0)} s(x,y) = -\infty$$

whence, $s(x,y)$ is unbounded over the domain Q, and because $T_{d_n}(s)$ is symmetric, it follows from Lemma 1.3.2 (a generalization) that $s(x,y)$ when defined over the domain Q is even.

Case 6: Gaussian

$$c_{j,k} = \frac{4}{h^2 g^4}(j^2 + k^2 - g^2)e^{-\frac{j^2+k^2}{g^2}}.$$

Whence,

$$c_{0,k} = \frac{4}{h^2 g^4}(k^2 - g^2)e^{-\frac{k^2}{g^2}} = \frac{4(n+1)^4}{c^4(n+1)^4}\left(-c^2 + \left(\frac{k}{n+1}\right)^2\right)e^{-\frac{1}{c^2}\left(\frac{k}{n+1}\right)^2} = \frac{4}{c^4}(-c^2 + x_k^2)e^{-\frac{1}{c^2}x_k^2}.$$

The expression of $c_{0,k}$ is similar to that of c_k obtained in **case** 2 so, one deduces from **case** 2 that

$$\lim_{n\to\infty} \sum_{k=1}^{n+1} c_{0,k} = -\infty. \tag{9.19}$$

Since the Fourier coefficients $c_{j,k}$ of $s(x,y)$ are all negative, it follows that

$$c_{0,0} + 4\left(\sum_{k=1}^{n+1} c_{0,k} + \sum_{k=1}^{n+1}\sum_{j=1}^{n+1} c_{j,k}\right) \leq \sum_{k=1}^{n+1} c_{0,k}. \tag{9.20}$$

According to (9.11) – (9.19) – (9.20), we deduce that

$$\lim_{(x,y)\to(\pm\pi,\pm\pi)} s(x,y) = -\infty, \quad \lim_{(x,y)\to(0,\pm\pi)} s(x,y) = -\infty$$

and

$$\lim_{(x,y)\to(\pm\pi,0)} s(x,y) = -\infty, \quad \lim_{(x,y)\to(0,0)} s(x,y) = -\infty.$$

Therefore, $s(x,y)$ is unbounded over the domain Q, and since $T_{d_n}(s)$ is symmetric, it follows from Lemma 1.3.2 (a generalization) that $s(x,y)$ is even. □

Remark 9.3.1. *The matrix sequences $\{A_{d_n}\}_n$ and $\{T_{d_n}(s)\}_n$ are equally distributed and equally localized (see chapter 8). The generating function s of the Toeplitz (respectively block Toeplitz) matrix $T_{d_n}(s)$ extended over the domain $[-\pi,\pi]$ (respectively $[-\pi,\pi]\times[-\pi,\pi]$) is real-valued and even.*

Remark 9.3.2. *The solution of the system of linear equations $T_{d_n}(s)v = \tilde{f}$ provides an approximate solution of the initial system $A_{d_n}v = \tilde{f}$. This solution is also an approximate solution of the Poisson problems* (9.1) − (9.2).

As it was shown in chapter 8, *the spectral radius of the matrices $T_{d_n}(s)$ is bounded independently of n in the Inverse Multiquadric case (for the Toeplitz case) and grows than $(n+2)^2$ for Block Toeplitz case) so, the matrices $T_{d_n}(s)$ are ill-conditioned for any value of n. More precisely, the Euclidean condition number of $T_{d_n}(s)$, as a function of the dimensions, is unbounded:*

$$\lim_{n\to\infty} k_2(T_{d_n}(s)) = \infty. \tag{9.21}$$

Hence, unless some preconditioning are used, all classic iterative methods are very slow.

We end this section by recalling the following fundamental result

Proposition 9.3.1. *Let $f, g \in L^1(-\pi,\pi)$ be two functions of constant signs such that $sign(f) = sign(g)$. Let $\lambda_j\left(T_n(g)^{-1}T_n(f)\right)$ $(1 \le j \le n)$ be the eigenvalues of $T_n(g)^{-1}T_n(f)$. If*

$$0 < m \le \frac{f(\theta)}{g(\theta)} \le M \qquad \forall\theta \in [-\pi,\pi] \tag{9.22}$$

then for $j = 1, 2, ..., n$,

$$\lambda_j\left(T_n(g)^{-1}T_n(g)\right) \in (m, M). \tag{9.23}$$

Proof. Because f and g are of constant signs and $sign(f) = sign(g)$ then, the Toeplitz matrices $T_n(g)$ and $T_n(f)$ are positive (or negative) definite. By looking the contradiction, let us suppose that there exists an eigenvalue λ_{j_0} of $T_n(g)^{-1}T_n(f)$ such that $\lambda_{j_0} \le m$ or $\lambda_{j_0} \ge M$. Without loss of generality, one can suppose that $\lambda_{j_0} \le m$. Then the matrix $B_n = T_n(f) - \lambda_{j_0}T_n(g) = T_n(f - \lambda_{j_0}g)$ is singular. **Indeed**: since $T_n(g)^{-1}B_n = T_n(g)^{-1}T_n(f) - \lambda_{j_0}I_n$ and λ_{j_0} is an eigenvalue of $T_n(g)^{-1}T_n(f)$ then 0 is an eigenvalue of $T_n(g)^{-1}B_n$ whence, $T_n(g)^{-1}B_n$ is singular. Because $T_n(g)$ is nonsingular (since $T_n(g)$ is positive (or negative) definite) one deduce that B_n is singular.

On the other side, $\forall\theta \in [-\pi,\pi]$,

$$f(\theta) - \lambda_{j_0}g(\theta) = g(\theta)\left[\frac{f(\theta)}{g(\theta)} - \lambda_{j_0}\right] \ge 0 \text{ for } g(\theta) \ge 0 \text{ (or } \le 0 \text{ if } g(\theta) \le 0).$$

Then $B_n = T_n(f - \lambda_{j_0}g)$ is positive (or negative) definite, contradiction.Therefore relation (9.23) holds true. □

In the following, we restrict our study on the Toeplitz matrices whose the generating functions $f(x)$ are Riemann-Lebesgue integrable over $I = [-\pi,\pi]$ and on the block Toeplitz matrices whose the generating functions $f(x,y)$ are just integrable over $Q = [-\pi,\pi]^2$. Furthermore, we define the numbers d_n as the positive integer numbers such that $d_n < d_{n+1}$ $\forall n \in \mathbb{N}$.

9.4 The Classical Szegö theory

This section deals with the main results of the Szegö theory on the spectral behavior of symmetric or Hermitian Toeplitz matrices generated by a Lebesgue integrable function f. Here and in the following we denote by $ess\inf f$ and $ess\sup f$ the essential infimum and the essential supremum of f [77], that is, $\inf f$ and $\sup f$ up to zero-measure sets; moreover, the symbol $m\{A\}$ denotes the Lebesgue measure of the set A, where A is a subset of $\mathbb{N}^N$, $N > 0$. We denote $I = [-\pi, \pi]$. Certain of these results were stated in chapters 1 and 2.

Theorem 9.4.1. [77, 167]. *Let $\lambda_i^{(d_n)}$ be the eigenvalues of $T_{d_n}(f)$ (which are real, since f is (even) real-valued and the matrix $T_{d_n}(f)$ is (symmetric) Hermitian) ordered in a non-decreasing way then, for any continuous function $F \in \mathbf{C}[m_f, M_f]$ with $m_f = ess\inf f$, $M_f = ess\sup f$, the asymptotic formula*

$$\lim_{n\to\infty} \frac{1}{d_n} \sum_{i=1}^{d_n} F(\lambda_i^{(d_n)}) = \frac{1}{2\pi} \int_{-\pi}^{\pi} F(f(x))dx \tag{9.24}$$

holds true.

As a direct consequence of the definition of the Fourier coefficients a_k of the function f and in the light of the Theorem 9.24, it is possible to prove a localization result on the spectrum $\sum_{d_n}(f)$ of the matrices $T_{d_n}(f)$ and a first estimate of the asymptotic distribution of $\sum_{d_n}(f)$.

Definition 9.4.1. *We define $\mathcal{ER}(f)$, the essential range of f, as the closed set of the real values y such that, $\forall \epsilon > 0$, the Lebesgue measure, $m\{x \in I : f(x) \in (y-\epsilon, y+\epsilon)\}$ is always positive.*

Theorem 9.4.2. [121]. *Let f be a Lebesgue integrable (even) real-valued function not essentially zero on I ($m_f < M_f$). Then we have*

1. *$\lambda_i^{(d_n)} \in (m_f, M_f)$ for any $i \le d_n$ and $d_n \in \mathbb{N}^\star$,*
2. *The topological closure of $\bigcup_{n\in\mathbb{N}^\star} \{\lambda_i^{(d_n)} : i = 1, ..., d_n\}$ contains $\mathcal{ER}(f)$,*
3. *If we fix an index i independently of n, then*

$$\lim_{n\to\infty} \lambda_i^{(d_n)} = m_f \quad and \quad \lim_{n\to\infty} \lambda_{d_n-i}^{(d_n)} = M_f.$$

Of course, if $m_f = M_f$ then $T_{d_n}(f) = m_f \cdot I$.

More recently, in [170] finer relationships have been proved.

Theorem 9.4.3. [121]. *If $m\{x \in I :\ f(x) = a\} = m\{x \in I :\ f(x) = b\} = 0$, then*

$$\lim_{n\to\infty} \frac{\sharp\{\lambda_i^{(d_n)} \in [a,b]\}}{d_n} = \frac{m\{x \in I :\ f(x) \in [a,b]\}}{2\pi}.$$

Therefore, if we denote by X_ϵ the set

$$[m_f, M_f]/B(\mathcal{ER}(f), \epsilon),$$

where $B(A, \epsilon)$ is the union of the open balls centered in $a \in A$ with radius equal to ϵ then, roughly speaking, few eigenvalues of $T_{d_n}(f)$ belongs to X_ϵ. More precisely, for any positive ϵ, $\sharp\left\{\sum_{d_n}(f) \cap X_\epsilon\right\} = o(d_n)$, so that most of the spectrum of $T_{d_n}(f)$ must be contained in the essential range of f. However, surprisingly enough, in [170] the following theorem is proved.

Theorem 9.4.4. $\bigcup_{n\in\mathbb{N}^\star}\left\{\lambda_i^{(d_n)} : i = 1, ..., d_n\right\}$ *is dense in* $[m_f, M_f]$.

In the case where f is complex-valued, it is evident that the related matrices $T_{d_n}(f)$ are, in general, not Hermitian. When considering the singular values in place of the eigenvalues, an elegant version of the Szegö theory results.

Theorem 9.4.5. [121]. *Let f be the complex-valued generating function of the matrix $T_n(f)$ and let F be a continuous function defined on the closed set $[m_f, M_f]$, where $M_f = ess\sup|f|$ and m_f is the Euclidean distance of zero from the convex hull of the essential range of f. Then the relation*

$$\lim_{n\to\infty}\frac{1}{d_n}\sum_{i=1}^{d_n}F(\sigma_i^{(d_n)}) = \frac{1}{2\pi}\int_{-\pi}^{\pi}F(|f(x)|)dx \tag{9.25}$$

is verified, where the quantities $\sigma_i^{(d_n)}$ denote the singular values of the matrix $T_{d_n}(f)$.

This result, under some restriction assumptions, has been proved by Parter [110] by exploiting linear-algebra tools, more general statements can be found in [2, 167]: there the argument is mainly based on functional analysis and operator theory.

In addition, in [172], under some hypotheses about f, the author proves a second-order result which can be used in order to give an asymptotic estimate about the number of the few eigenvalues not belonging to $\mathcal{ER}(f)$. This refinement is a little simpler to state in terms of the squares of $\sigma_i^{(d_n)}$ than in terms of the singular values themselves. The restriction about f is due to the assumption that it belongs to the Krein algebra K [91] of all the functions f such that f is essentially bounded and

$$\|f\|_K^2 = \sum_{k=-\infty}^{\infty}|k||a_k|^2 < \infty.$$

We remark that the Krein algebra contains the linear subspace of the C^2 periodic functions. More precisely, by means of classical estimates on the Fourier coefficients of continuous functions (see the chapter on the trigonometric interpolation in [149] and the Bible [176] on trigonometric series), we can say that

1. K contains $C_\star^1 = \{f \in C^1 : f'$ is absolutely continuous$\}$,

2. K contains $T_\alpha = \{f \in C^1 : f' \in Lip_\alpha\},\ \alpha \in (0, 1]$,

3. K contains $T_\star = \{f \in C^1 : f'$ is in the Dini-Lipchitz *class*$\}$.

More precisely, for the class $C_\star^1$, we can conclude that the Fourier coefficients a_k are $O(k^{-2})$ [149], while for the classes T_α and $T_\star$ we have $a_k = O(k^{-(1+\alpha)})$ and $a_k = o\left((k\log k)^{-1}\right)$ respectively. Notice that all these spaces are subspaces of the C^1 and, with the exception of $T_\star$, are also subspaces of the well-known Wiener class [37, 48]. Concerning the functions in the Krein algebra the second-order result of Widom is stated in the following theorem [172].

Theorem 9.4.6. *Let $t_i^{(d_n)} = \left(\sigma_i^{(d_n)}\right)^2$, let $f \in K$, and let G be a function belonging to $C^3[m_f^2, M_f^2]$. Then*

$$\lim_{n\to\infty}\left(\sum_{i=1}^{d_n}G(t_i^{(d_n)}) - \frac{d_n}{2\pi}\int_{-\pi}^{\pi}G(|f(x)|^2)dx\right) = c, \tag{9.26}$$

where c is known constant characterized in [172].

9.5 Fundamental results on the distribution of Toeplitz spectra

In this section, by using the preceding limit relations (9.24) – (9.25) – (9.26), we want to analyze the asymptotic distribution of the eigenvalues and/or the singular values of the Toeplitz matrix $T_{d_n}(f)$. We will show that the essential range of f or $|f|$ plays a fundamental role in understanding where "most" of the eigenvalues and/or the singular values of $T_{d_n}(f)$ tend to concentrate.

In order to explain the technique used in the following, we start by proving in the following proposition, a relation which is also a consequence of Theorem 9.4.3.

Proposition 9.5.1. *For any positive number ϵ, let X_ϵ be the set*

$$[m_f, M_f]/B(\mathcal{ER}(f), \epsilon),$$

with f an (even) real-valued integrable function, $m_f = ess\inf f$, $M_f = ess\sup f$. Then $o(d_n)$ eigenvalues of $T_{d_n}(f)$ belong to X_ϵ.

Proof. For any $\epsilon > 0$, let us consider a continuous function F_ϵ constructed in the following way:

1. F_ϵ is defined over the domain $[m_f, M_f]$, and $0 \le F_\epsilon \le 1$,

2. $\mathrm{Supp}(F_\epsilon) \cap \mathcal{ER}(f)$ is empty, and $F_\epsilon \equiv 1$ on X_ϵ.

Since F_ϵ is chosen continuous, we can apply Equation (9.24), obtaining

$$\lim_{n\to\infty} \frac{1}{d_n} \sum_{i=1}^{d_n} F_\epsilon(\lambda_i^{(d_n)}) = 0 \tag{9.27}$$

since F_ϵ is nonzero only where $f(x)$ does not belong to $\mathcal{ER}(f)$. Therefore, as $\sum_{i=1}^{d_n} F_\epsilon(\lambda_i^{(d_n)}) \ge \sharp\{\lambda_i^{(d_n)} \in X_\epsilon\}$ we have proved $\sharp\{\lambda_i^{(d_n)} \in X_\epsilon\} = o(d_n)$. □

Analogously it is possible to prove a similar result for the singular values.

Proposition 9.5.2. *For any positive number ϵ, let X_ϵ be the set*

$$[m_f, M_f]/B(\mathcal{ER}(f), \epsilon),$$

with f a complex-valued integrable function, $M_f = ess\sup f$, and m_f the distance of 0 from the convex hull of the essential range of f (which is clearly less or equal to $ess\inf |f|$). Then $o(d_n)$ singular values of $T_{d_n}(f)$ belong to X_ϵ.

Proof. It is enough to follow the same steps as in Proposition 9.5.1. □

Now, with the help of the second-order result obtained by Widom (see Theorem 9.4.6), we can obtain a finer result about the asymptotic distribution of the singular values of $T_{d_n}(f)$.

Theorem 9.5.1. *Let $f \in K$. Then, for any $\epsilon > 0$, only $O(1)$ singular values of $T_{d_n}(f)$ lie in the set*

$$X_\epsilon = [m_f, M_f]/B(\mathcal{ER}(|f|), \epsilon).$$

Here m_f is the Euclidean distance of the convex hull of the essential range of f from zero, while M_f is the supremum of $|f|$.

Proof. For any $\epsilon > 0$, we define a three times continuously differentiable function G_ϵ characterized by the following properties:

1. G_ϵ is defined on $[m_f^2, M_f^2]$ and $0 \leq G_\epsilon \leq 1$,

2. $G_\epsilon \equiv 0$ on $\mathcal{ER}(|f|^2)$ and $G_\epsilon \equiv 1$ on

$$[m_f^2, M_f^2]/B(\mathcal{ER}(|f|^2), \epsilon),$$

3. $G_\epsilon \in C^3[m_f^2, M_f^2]$.

From the regularity features of G_ϵ we may apply Theorem 9.4.6, obtaining that

$$\frac{1}{d_n}\sum_{i=1}^{d_n} G_\epsilon\left((\sigma_i^{(d_n)})^2\right) = \frac{1}{2\pi}\int_{-\pi}^{\pi} G_\epsilon(|f(x)|^2)dx + \frac{c_\epsilon}{d_n} + o\left(\frac{1}{d_n}\right), \tag{9.28}$$

where c_ϵ is the constant c in Theorem 9.4.6. In the light of the definition of G_ϵ we have

$$\frac{1}{d_n}\sum_{i=1}^{d_n} G_\epsilon\left((\sigma_i^{(d_n)})^2\right) = \frac{1}{d_n}\sharp\{\sigma_i^{(d_n)} \in X_\epsilon\} + \frac{\theta_\epsilon}{d_n} + o\left(\frac{1}{d_n}\right),$$

with $\theta_\epsilon > 0$ and going to zero when $\epsilon \to 0$. Since $G_\epsilon(|f(x)|^2) = 0$ almost everywhere (a.e.) by definition, it follows that $\sharp\{\sigma_i^{(d_n)} \in X_\epsilon\}$ is bounded by a constant depending only on ϵ. □

Proposition 9.5.3. *Let $f \in K$ and f be an (even) real-valued function. Then, for any $\epsilon > 0$, only $O(1)$ eigenvalues of $T_{d_n}(f)$ lie in the set*

$$[m_f, M_f]/B(\mathcal{ER}(f), \epsilon).$$

Proof. It is enough to consider a positive number α such that $f + \alpha$ is essentially positive. In this case the eigenvalues of $T_{d_n}(f + \alpha)$ coincide with the singular values of $T_{d_n}(f + \alpha)$. Therefore we apply the former theorem by obtaining that $O(1)$ eigenvalues of $T_{d_n}(f + \alpha)$ lie in the set

$$[m_f + \alpha, M_f + \alpha]/B(\mathcal{ER}(f + \alpha), \epsilon).$$

Since the eigenvalues of $T_{d_n}(f + \alpha)$ are the eigenvalues of $T_{d_n}(f)$ shifted according to the constant α, the claim holds true. □

9.6 Preconditioned Toeplitz matrices $\mathcal{P}_{d_n}(f; g)$

The aim of this section is a brief exposition of recent results [56, 137, 138, 139, 140, 118, 117] about the spectra of preconditioned Toeplitz matrices. In particular, we want to point out the formal connection with the classical spectral theory of symmetric or Hermitian (non-preconditioned) Toeplitz matrices. Here and in the following, by preconditioned Toeplitz matrix we mean a matrix of the form $T_{d_n}(g)^{-1}T_{d_n}(f)$ which is also indicated by the shorter symbol $\mathcal{P}_{d_n}(f; g)$, where f and g belong to $L^1(I)$ are two (even) real-valued functions with g is essentially nonnegative and nonzero. We observe that, from the assumptions, the matrices $T_{d_n}(f)$ and $T_{d_n}(g)$ are well defined and $T_{d_n}(g)$ is (symmetric) positive definite (see the first part of Theorem 9.4.2), therefore the preconditioned matrix $T_{d_n}(g)^{-1}T_{d_n}(f)$ exists and is well defined.

The first localization result is the version for the preconditioned Toeplitz matrices of the first point of the classical Theorem 9.4.2.

Theorem 9.6.1. *Let $\{T_{d_n}(g)\}_n$ and $\{T_{d_n}(f)\}_n$ be two sequences of (symmetric) Hermitian Toeplitz matrices generated by two (even) real-valued integrable functions g and f respectively, where g is essentially nonnegative and nonzero. Then, for any positive integer n, the preconditioned matrix $\mathcal{P}_{d_n}(f;g)$ has eigenvalues in the open set (r,R), where $r = ess\inf(f/g)$, $R = ess\sup(f/g)$ with $r < R$. Otherwise, if $r = R$ then the preconditioned matrix has the form $\mathcal{P}_{d_n}(f;g) = r \cdot I_{d_n}$.*

The next theorem gives indications about the asymptotic distribution of the spectrum of $\mathcal{P}_{d_n}(f;g)$ in the closed set $[r,R]$ that is in the convex hull of the essential range of f/g.

Theorem 9.6.2. [137, 140]. *Let $\lambda_i^{(d_n)}$ be the eigenvalues of $\mathcal{P}_{d_n}(f;g)$ ordered in a nondecreasing way. Then the following relations hold true*

(a) *The topological closure of $\bigcup_{n\in\mathbb{N}^\star}\{\lambda_i^{(d_n)} : i = 1,...,d_n\}$ contains $\mathcal{ER}(f/g)$,*

(b) *$\mathcal{P}_{d_n}(f;g)$ has eigenvalues in (r,R), and $\lim_{n\to\infty}\lambda_1^{(d_n)} = r$, $\lim_{n\to\infty}\lambda_{d_n}^{(d_n)} = R$,*

(c) *If $f/g \sim |x-x_0|^\rho$ for some $x_0 \in I$, then $\lambda_1^{(d_n)} - r$ is asymptotic to $d_n^{-\rho}$.*

Now, we consider the PCG method proposed in [118] for the solution of indefinite symmetric or Hermitian Toeplitz systems. By using Proposition 9.5.3, we refine the convergence analysis previously performed [118].

Brief description of the method

In this subsection we give a concise description of the PCG method introduced in [118] in order to deal with the nondefinite case.

Let us assume that the entries of the Toeplitz matrix $T_{d_n}(f)$ are given and that the function (with nondefinite sign) $f(x)$ is known, say, by means of its formal expression. Here and in hereafter we assume that $f(x)$ has zeros $x_1, x_2, ..., x_m \in I$.

The method is outlined by the following stages:

• **Stage** 1. Find g such that $g(x_i) = 0$, g is positive elsewhere, and the closed set $\mathcal{ER}(f/g)$ is contained in $[\alpha^-,\beta^-]\cup[\alpha^+,\beta^+]$, where $\alpha^- \le \beta^- < 0 < \alpha^+ \le \beta^+$, for instance, set $g = |f|$.

Remark 9.6.1. *The hypotheses made in stage* 1 *are a generalization of Proposition* 9.3.1.

• **Stage** 2. Compute the symmetric or Hermitian positive definite Toeplitz matric $T_{d_n}(g)$ (see Theorems 1.3.2 and 1.3.3), and consider the equivalent non-Hermitian system

$$\mathcal{P}_{d_n}(f;g)\mathbf{x} = \hat{\mathbf{b}}$$

where $\hat{\mathbf{b}} = T_{d_n}^{-1}(g)\mathbf{b}$.

• **Stage** 3. Consider the new equivalent system

$$[\mathcal{P}_{d_n}(f;g)]^2\,\mathbf{x} = \tilde{\mathbf{b}}$$

where the new coefficient matrix is associated with a symmetrizable positive definite form and $\tilde{\mathbf{b}} = \mathcal{P}_{d_n}(f;g)\hat{\mathbf{b}}$. Solve it by means of the PCG method.

Actually, since $[\mathcal{P}_{d_n}(f;g)]^2 = T_{d_n}^{-1}(g)T_{d_n}(f)T_{d_n}^{-1}(g)T_{d_n}(f)$, we can look at the coefficient matrix as the product of $T_{d_n}^{-1}(g)$, which is a symmetric or Hermitian positive definite matrix, and $T_{d_n} = T_{d_n}(f)T_{d_n}^{-1}(g)T_{d_n}(f)$, which is a (symmetric) Hermitian positive definite matrix if $T_{d_n}(g)$ is (symmetric) positive definite and $T_{d_n}(f)$ is nonsingular and is nonnegative (or

nonpositive) definite otherwise.

Therefore, in stage 3, we may apply a PCG method in which T_{d_n} is the coefficients matrix of a new equivalent system and $T_{d_n}(g)$ is a preconditioner. Of course, the convergence features of the PCG method, that is, the number of iterations needed to reach the solution within a fixed accuracy, are determined by the spectral properties of the matrix $[\mathcal{P}_{d_n}(f;g)]^2$, or equivalently by the spectral behavior of $\mathcal{P}_{d_n}(f;g)$.

Now, the main goal is to analyze the convergence speed of this procedure. Actually, by applying the result [6] and the following theorems, we expect that the conjugate-gradient method, applied to the system $[\mathcal{P}_{d_n}(f;g)]^2\,\mathbf{x} = \tilde{\mathbf{b}}$, converges to the solution with a preassigned accuracy ϵ in at most $k+q$ iterations, where

$$k = \left\lceil \frac{log(2/\epsilon) + q\, log(c^+/\lambda_1^{(d_n)})}{log(1/\delta)} \right\rceil, \quad \delta = \frac{\sqrt{c^+} - \sqrt{c^-}}{\sqrt{c^+} + \sqrt{c^-}}, \tag{9.29}$$

$$c^+ = (\max\{|\alpha^-|, |\beta^+|\})^2, \quad c^- = (\max\{|\alpha^+|, |\beta^-|\})^2,$$

and q is a constant if f and g belong to the Krein algebra (see Theorem 9.6.4). In addition, if ϵ is fixed and $\lambda_1^{(d_n)} \geq \theta > 0$ or goes to zero slowly, then the desired precision is practically reached through a constant number of iterations.

From a practical point of view, in order to perform stage 1, we choose $g(x)$ such that $g(x_i) = 0$, g is positive elsewhere and

$$0 < \liminf_{x \to x_i} \left|\frac{f}{g}\right| \leq \limsup_{x \to x_i} \left|\frac{f}{g}\right| < \infty. \tag{9.30}$$

In this case we have

$$\mathcal{ER}\left(\frac{f}{g}\right) \subset [\alpha^-, \beta^-] \cup [\alpha^+, \beta^+] \tag{9.31}$$

where $\alpha^- \leq \beta^- = \sup_{x \in I} \left\{ \frac{f(x)}{g(x)} < 0 \right\} < 0 < \alpha^+ = \inf_{x \in I} \left\{ \frac{f(x)}{g(x)} > 0 \right\} \leq \beta^+$.

The following result holds true.

Theorem 9.6.3. [118]. *Let f and g be two continuous functions satisfying the conditions* (9.30) *and* (9.31). *Then the following statements hold true*

(a) *For any $\eta > 0$, $\sharp\{\lambda(\mathcal{P}_{d_n}(f;g)) \cap ([\alpha^-, \beta^- + \eta] \cup [\alpha^+ - \eta, \beta^+])\} = d_n - o(d_n)$,*

(b) *If f and g are even rational functions then, for any $\eta > 0$,*

$$\sharp\{\lambda(\mathcal{P}_{d_n}(f;g)) \cap ([\alpha^-, \beta^- + \eta] \cup [\alpha^+ - \eta, \beta^+])\} = d_n - o(1).$$

Remark 9.6.2. *It is worth noting that in that theorem we have demonstrated that the number of "outliers" is constant only in the special case of f and g being trigonometric rational symmetric functions. The next result is a strong generalization of the cited result, because we extend this property to the whole Krein algebra, which contains all the spaces $C^1_\star[I]$ functions.*

Theorem 9.6.4. *Let f and g be two functions belonging to the Krein algebra and satisfying the conditions* (9.30) *and* (9.31). *Then, for any $\eta > 0$,*

$$\sharp\{\lambda(\mathcal{P}_{d_n}(f;g)) \cap ([\alpha^-, \beta^- + \eta] \cup [\alpha^+ - \eta, \beta^+])\} = d_n - O(1).$$

Proof. Let us take $\alpha = (\beta^+ + \alpha^+)/2$, and let us define $f_t(x)$ as $f(x) - tg(x)$, it is evident that the function $f_t(x)$ vanishes if and only if $x \in \{x_1, x_2, ..., x_m\}$, for all $t \in (\beta^-, \alpha^+)$. Moreover, when setting $\epsilon = (\alpha^+ - \beta^-)/2 - \eta$, $f_t(x)$ keeps the same sign for all $t \in [\alpha - \epsilon, \alpha + \epsilon]$ (see Lemma 2.1 in [118]), and it follows that

$$m\{x \in I : f_{\alpha-\epsilon}(x) < 0\} = m\{x \in I : f_{\alpha+\epsilon}(x) < 0\}. \tag{9.32}$$

It is worth pointing out that, by the assumptions of Theorem 9.4.3, we can choose $\eta = 0$ if $m\{x \in I : f_{\alpha-\epsilon}(x) = 0\} = 0$ and $m\{x \in I : f_{\alpha+\epsilon}(x) = 0\} = 0$, that is, if the sets Z_1 and Z_2 for which f/g coincides with β^- and α^+ are zero-measure sets, otherwise, η has to be positive but, it can be chosen as small as we like.

Therefore, by the results of Theorem 9.4.3, we obtain that the inertia of $T_{d_n}(f) - (\alpha - \epsilon)T_{d_n}(g)$ is almost the same as that of $T_{d_n}(f) - (\alpha + \epsilon)T_{d_n}(g)$, that is,

$$\begin{aligned} &\sharp\{\lambda_i\left(T_{d_n}(f) - (\alpha - \epsilon)T_{d_n}(g)\right) < 0,\ i = 1, 2, ..., d_n\} \\ &= \sharp\{\lambda_i\left(T_{d_n}(f) - (\alpha + \epsilon)T_{d_n}(g)\right) < 0,\ i = 1, 2, ..., d_n\} + w, \end{aligned} \tag{9.33}$$

where $w = O(1)$ if f and g are symmetric rational functions (see Theorem 2.4 of [118]) or if f and g are in the Krein algebra (see Proposition 9.5.3). But $T_{d_n}^{-1/2}(g)$ is (symmetric) positive definite, and therefore the inertia of $T_{d_n}(f) - tT_{d_n}(g)$ coincides with that of $T_{d_n}^{-1/2}(g)T_{d_n}(f)T_{d_n}^{-1/2}(g) - t \cdot I$ for any $t \in \mathbb{R}$, the latter matrix is similar to $T_{d_n}^{-1}(g)T_{d_n}(f) - t \cdot I$, and so, using (9.33), we have

$$sign\{\lambda_i\left(T_{d_n}(f) - (\alpha - \epsilon)T_{d_n}(g)\right)\} = sign\sharp\{\lambda_i\left(T_{d_n}(f) - (\alpha + \epsilon)T_{d_n}(g)\right)\}$$

for all $i \in J$ where $\sharp J = d_n - w$. Moreover,

$$\lambda_i\left((T_{d_n}^{-1}(g))[T_{d_n}(f) - (\alpha + \epsilon)T_{d_n}(g)]\right) = \lambda_i\left((T_{d_n}^{-1}(g))[T_{d_n}(f) - (\alpha - \epsilon)T_{d_n}(g)]\right) - 2\epsilon,$$

and consequently, $T_{d_n}^{-1}(g)T_{d_n}(f) - \alpha \cdot I$ has at most w eigenvalues in $(-\epsilon, \epsilon)$, that is, $T_{d_n}^{-1}(g)T_{d_n}(f)$ has at most w eigenvalues in $(\alpha - \epsilon, \alpha + \epsilon)$. Since $(\alpha - \epsilon, \alpha + \epsilon)$ coincides with $(\beta^- + \eta, \alpha^+ - \eta)$, the theorem is proved, with $\eta = 0$ if Z_1 and Z_2 are zero-measure sets, and with $\eta > 0$ but as small as we like otherwise. □

Therefore, we can conclude that the PCG method devised in [118] has a good convergence rate. Since the number of outliers is independent of n (under the assumption that f and g belong to the Krein algebra K), the total number of iterations required to reach the solution within a preassigned accuracy ϵ can grow, at most, logarithmically with the condition number of the preconditioned matrix $[\mathcal{P}_{d_n}(f; g)]^2$. Finally, we recall that, in the light of the numerical experiments [118], it seems that the decrease to zero of the minimal eigenvalue of $[\mathcal{P}_{d_n}(f; g)]^2$ is also very slow, and, consequently, the convergence speed of the related PCG method is substantially constant and independent of the dimension d_n of the problem.

In the following we extend our works to the Hermitian block Toelitz matrices with Hermitian Toeplitz bloks, Indeed: this generalizes the case of symmetric block Toeplitz matrices with symmetric Toeplitz blocks.

9.7 Hermitian block Toeplitz matrices with Hermitian Toeplitz blocks

We start by recalling that if $f(x, y)$ is a Lebesgue integrable function defined over the domain $Q = [-\pi, \pi] \times [-\pi, \pi]$ and extended periodically to $\mathbb{R}^2$ then, the Fourier coefficients of f are given by

$$a_{k,q} = \frac{1}{4\pi^2}\int_Q f(x,y)e^{-\hat{i}(kx+qy)}dxdy, \qquad k,q \in \mathbb{Z}. \tag{9.34}$$

By $T_{d_n,d_m}(f)$ we denote the $d_n \times d_n$ block Toeplitz matrix with $d_m \times d_m$ Toeplitz blocks defined according to relation (9.34). If the pair (j,k) indicates the position in the matrix $T_{d_n,d_m}(f)$ and the pair (p,q) the position of the entry in the block, then we have

$$(T_{d_n,d_m}(f))_{(j,k),(p,q)} = a_{k-j,p-q}$$

for $j,k = 0,1,...,d_n-1$ and $p,q = 0,1,...,d_m-1$. Of course, if f is (even) real-valued then the matrix $T_{d_n,d_m}(f)$ is (symmetric) Hermitian so that its spectrum is real. In addition we have this characterizing result which relates the eigenvalues of $T_{d_n,d_m}(f)$ and the analytic behavior of the function f.

Theorem 9.7.1. [117, 154]. *Let $f : Q \to \mathbb{R}$ be a Lebesgue integrable function. Then all the eigenvalues of $T_{d_n,d_m}(f)$ lie in the open set (m_f, M_f) in the case where $m_f = \underset{(x,y)\in Q}{ess\inf} f < M_f = \underset{(x,y)\in Q}{ess\sup} f$. Moreover, if we order the eigenvalues $\lambda_j^{(d_n,d_m)}(T_{d_n,d_m}(f))$ in a nondecreasing way, then for any fixed k independent of n and m we find*

$$\lim_{n,m\to\infty} \lambda_k^{(d_n,d_m)}(T_{d_n,d_m}(f)) = m_f, \quad \lim_{n\to\infty} \lambda_{d_nd_m-k}^{(d_n,d_m)}(T_{d_n,d_m}(f)) = M_f.$$

Finally, if $m_f = M_f$ then $T_{d_n,d_m}(f) = m_f \cdot I_{d_n\cdot d_m}$

In addition we have [154] (for the whole ergodic theorem analogous to Theorem 9.4.1 holding in $d_m > 1$ dimensions and for L^1 functions see [167]).

Theorem 9.7.2. *If $m\{(x,y) \in Q : f(x,y) = a\} = m\{(x,y) \in Q : f(x,y) = b\} = 0$, then*

$$\lim_{n,m\to\infty} \frac{\sharp\left\{\lambda_i^{(d_n,d_m)} \in [a,b]\right\}}{d_n \cdot d_m} = \frac{m\left\{(x,y) \in Q : f(x,y) \in [a,b]\right\}}{4\pi^2}.$$

Now we recall a theorem about the localization of the spectrum of the considered block Toeplitz matrices which is a direct generalization of the analogous Theorem 9.6.1 in the scalar case.

Theorem 9.7.3. [137, 117]. *Let $\{T_{d_n,d_m}(f)\}_{n,m}$ and $\{T_{d_n,d_m}(g)\}_{n,m}$ be two sequences of (symmetric) Hermitian block Toeplitz matrices generated by two (even) real-valued integrable functions f and g respectively, with g essentially nonnegative and nonzero is shown such that f/g is bounded over Q. Then the preconditioned matrix*

$$\mathcal{P}_{d_n,d_m}(f;g) \equiv T_{d_n,d_m}(g)^{-1}T_{d_n,d_m}(f)$$

has eigenvalues in the open set (r,R) with

$$r = \underset{(x,y)\in Q}{ess\inf}\frac{f(x,y)}{g(x,y)} < R = \underset{(x,y)\in Q}{ess\sup}\frac{f(x,y)}{g(x,y)}.$$

If $r = R$, then the precoditioned matrix simply coincides with $r \cdot I_{d_n}$

9.7.1 Some consequences of Tyrtyshnikov and Zamarashkin's result

In this subsection we obtain a result of asymptotical distribution about the spectra of the preconditioned matrices. As in the scalar case, the main tools are the properties of the eigenvalues of the nonpreconditioned matrices, more specifically, the thesis of Theorem 9.7.2 will play a crucial role in the proof of the next result.

Theorem 9.7.4. *Let $\lambda_i^{(d_n,d_m)}$ be the eigenvalues of $\mathcal{P}_{d_n,d_m}(f;g)$ ordered in a nondecreasing way. Then the following relations hold true*

(a) *The topological closure of* $\bigcup_{1\le n,m\le\infty}\left\{\lambda_i^{(d_n,d_m)}:\ i=1,...,d_n\cdot d_m\right\}$ *contains the essential range $\mathcal{ER}(f/g)$,*

(b) *$\mathcal{P}_{d_n,d_m}(f;g)$ has eigenvalues in (r,R) and*

$$\lim_{n,m\to\infty}\lambda_1^{(d_n,d_m)}=r,\quad \lim_{n,m\to\infty}\lambda_{d_n\cdot d_m}^{(d_n,d_m)}=R.$$

Proof. We prove the general result under the sole hypothesis that g is essentially nonnegative and $m\{(x,y)\in Q:\ g(x,y)=0\}=0$. Indeed, the thesis is equivalent to the following statement

$$\forall\alpha\in\mathcal{ER}(f/g),\ \forall\epsilon>0,\ \exists n,m\in\mathbb{N},\ \exists\lambda\in\Sigma_{d_n,d_m}\text{ such that }|\lambda-\alpha|<\epsilon.$$

Here, Σ_{d_n,d_m} denotes the eigenvalue sets of $\mathcal{P}_{d_n,d_m}(f;g)$.

Let $H_{d_n,d_m,\alpha}=T_{d_n,d_m}(f)-\alpha T_{d_n,d_m}(g)$. If $H_{d_n,d_m,\alpha}$ is singular for some values of d_n,d_m, then there exists $\lambda\in\Sigma_{d_n,d_m}$ such that $\lambda=\alpha$. Otherwise $H_{d_n,d_m,\alpha}$ is nonsingular for any pair of positive integers (d_n,d_m). Moreover, $H_{d_n,d_m,\alpha}=T_{d_n,d_m}\left(c_\alpha(x,y)\right)$, where $c_\alpha(x,y)$ is defined as $f(x,y)-\gamma g(x,y)$, with γ a real parameter.

Now we consider $m_\epsilon^\alpha=m\{(x,y)\in Q:\ f(x,y)-(\alpha+\epsilon)g(x,y)<0\}$ and $m_{-\epsilon}^\alpha=m\{(x,y)\in Q:\ f(x,y)-(\alpha-\epsilon)g(x,y)<0\}$. Since $g>0$ a.e, we have that $\forall(x,y)\in Q$ $f(x,y)-(\alpha+\epsilon)g(x,y)<f(x,y)-(\alpha-\epsilon)g(x,y)$ a.e, that is, $\frac{f}{g}-(\alpha+\epsilon)<\frac{f}{g}-(\alpha-\epsilon)$ a.e. But $\alpha\in\mathcal{ER}(f/g)$ and therefore $m_\epsilon^\alpha>m_{-\epsilon}^\alpha$.

At this point we recall an asymptotic result about the distribution of the eigenvalues of Toeplitz matrices due to Tilli [154] (it is also a simple consequence of the powerful result of Tyrtyshnikov and Zamarashkin [167]): for all $[a,b]$ and for any $f\in L^1(Q)$ such that $m\{(x,y)\in Q:\ f(x,y)=a\ \text{ or }f(x,y)=b\}=0$, we have

$$\lim_{n,m\to\infty}\frac{\sharp\left\{i:\ \lambda_i^{(d_n,d_m)}\left(T_{d_n,d_m}(f)\right)\in(a,b)\right\}}{d_n\cdot d_m}=\frac{m\left\{(x,y)\in Q:\ f(x,y)\in(a,b)\right\}}{4\pi^2}.$$

Consequently

$$\sharp\left\{i:\ \lambda_i^{(d_n,d_m)}\left(T_{d_n,d_m}(c_{\alpha+\epsilon})\right)<0\right\}=d_n\cdot d_m\left[\frac{m_\epsilon^\alpha}{4\pi^2}+o(1)\right],\tag{9.35}$$

$$\sharp\left\{i:\ \lambda_i^{(d_n,d_m)}\left(T_{d_n,d_m}(c_{\alpha-\epsilon})\right)<0\right\}=d_n\cdot d_m\left[\frac{m_{-\epsilon}^\alpha}{4\pi^2}+o(1)\right].\tag{9.36}$$

When using the relation $m_\epsilon^\alpha>m_{-\epsilon}^\alpha$, it follows that, for d_n,d_m large enough, "many" eigenvalues of $T_{d_n,d_m}(c_\gamma)$ move from positive values to negative ones when the parameter γ continuously moves from $\alpha-\epsilon$ to $\alpha+\epsilon$. As a consequence, by using a continuity argument, we find $\lambda(d_n,d_m)\in(\alpha-\epsilon,\alpha+\epsilon)$ such that the matrix $T_{d_n,d_m}(c_{\lambda(d_n,d_m)}(x,y))$ is singular, that is, $\lambda(d_n,d_m)\in\Sigma_{d_n,d_m}$. Therefore the theorem is proved. □

Remark 9.7.1. *In the proof of the Theorem* 9.7.4*, in the equations* (9.35), (9.36)*, we have supposed that* $m\{(x,y) \in Q : f(x,y) - (\alpha+\epsilon)g(x,y) = 0\} + m\{(x,y) \in Q : f(x,y) - (\alpha-\epsilon)g(x,y) = 0\} = 0$*. In the case where this assumption is not verified we can obviously choose* $\epsilon^\star$*,* $0 < \epsilon^\star < \epsilon$*, such that*

$$m\{(x,y) \in Q : f(x,y)-(\alpha+\epsilon^\star)g(x,y) = 0\} = m\{(x,y) \in Q : f(x,y)-(\alpha-\epsilon^\star)g(x,y) = 0\} = 0.$$

Moreover, if the thesis of the theorem is proved for $\epsilon^\star$ *such that* $0 < \epsilon^\star < \epsilon$*, then it holds for* ϵ*.*

9.7.2 Applications to the solution of block Toeplitz linear systems

In this subsection we want to stress the consequences of the last results on the convergence analysis of the CG and PCG methods applied to this kind of symmetric block Toeplitz with symmetric Toeplitz blocks systems. In [117], Serra Capizzano has proved that the minimal eigenvalue of $T_{d_n,d_m}(f)$ tends to m_f as n, m tends to ∞, and according to Theorem 9.7.1, the maximal eigenvalue of $T_{d_n,d_m}(f)$ tends to M_f as n, m tends to ∞. This means that, if $m_f = 0$, or $m_f = -\infty$, or $M_f = 0$ or $M_f = \infty$, then the associated linear system becomes ill-conditioned (for estimates on the magnitude of this ill-conditioning see [138]), but we are not able to give precise results on the convergence rate of the CG method, because we have no information about the distribution of the rest of the spectrum. On the other hand, by virtue of Theorems 9.7.1 and 9.7.2, we know very well the asymptotical behavior of the whole spectrum of $T_{d_n,d_m}(f)$ and therefore, we can give a tight result on the convergence speed of the CG method. More precisely, by means of Theorems 9.7.3 and 9.7.4, we can extend the same remark to the case of preconditioned matrices. In particular we can conclude that the asymptotic behavior of the CG and PCG methods is decided by the numbers M_f/m_f and R/r and by the sets $\mathcal{ER}(f)$ and $\mathcal{ER}(f/g)$.

For instance, let us consider the symmetric block Toeplitz with symmetric Toeplitz blocks linear system $T_{d_n,d_n}(s)v = \tilde{f}$ with $s(x,y) = \frac{-(x^2+y^2)}{2\pi^2-x^2-y^2}$. $s(x,y)$ is nonpositive, even, unbounded and Lebesgue integrable over the domain Q. **Indeed**: it is obvious that $s(x,y)$ is nonpositive and even over Q. On the other side, $\lim_{(x,y)\to(\pm\pi,\pm\pi)} s(x,y) = -\infty$, then $s(x,y)$ is unbounded over Q. Now, let us show that $s(x,y)$ is Lebesgue integrable on Q. First of all, the function $s(x,y)$ is even, then $\int_Q s(x,y)dxdy = 2\int_0^\pi \int_0^\pi s(x,y)dxdy$. Furthermore:

$$\begin{aligned}
\int_0^\pi \int_0^\pi s(x,y)dxdy &= \int_0^\pi \int_0^\pi \frac{-x^2-y^2}{2\pi^2-x^2-y^2}dxdy \\
&= \int_0^\pi \int_0^\pi (1 - \frac{2\pi^2}{2\pi^2-x^2-y^2})dxdy \\
&= \pi^2 - 2\pi^2 \int_0^\pi \int_0^\pi \frac{1}{2\pi^2-x^2-y^2}dxdy \\
&= \pi^2 - 2\pi^2 \int_0^\pi \frac{1}{\sqrt{2\pi^2-y^2}} \int_0^\pi \frac{d(\frac{x}{\sqrt{2\pi^2-y^2}})}{1-(\frac{x}{\sqrt{2\pi^2-y^2}})^2}dy \\
&= \pi^2 - 2\pi^2 \int_0^\pi \frac{1}{\sqrt{2\pi^2-y^2}} \int_0^{\frac{\pi}{\sqrt{2\pi^2-y^2}}} \frac{1}{1-t^2}dtdy \\
&= \pi^2 - 2\pi \int_0^\pi \frac{\pi}{\sqrt{2\pi^2-y^2}} \arg th(\frac{\pi}{\sqrt{2\pi^2-y^2}})dy.
\end{aligned}$$

Since the function $x \mapsto \arg th(x)$ is increasing on $[0,1)$ and $0 = \arg th(0) < \arg th(\frac{1}{\sqrt{2}}) \leq \arg th\left(\frac{\pi}{\sqrt{2\pi^2-y^2}}\right)$ $\forall y \in [0,\pi)$, then the function $y \mapsto \arg th\left(\frac{\pi}{\sqrt{2\pi^2-y^2}}\right)$ is positive over $[0,\pi)$.

Furthermore the function $y \mapsto \frac{\pi}{\sqrt{2\pi^2-y^2}}$ is continuous on $[0,\pi]$, then it is bounded on this subset. Now, let us show that the function $y \mapsto \arg th\left(\frac{\pi}{\sqrt{2\pi^2-y^2}}\right)$ is Lebesgue integrable on $[0,\pi]$. First of all, for $|x|<1$, $\arg th(x) = \frac{1}{2}\ln\left(\frac{1+x}{1-x}\right)$. **Indeed**: setting $f(x) = \arg th(x)$ and $g(x) = \ln\left(\frac{1+x}{1-x}\right)$, one has $g^{'}(x) = 2f^{'}(x) = \frac{2}{1-x^2}$, so $g(x) = 2f(x) + c$ where c is a constant number. Because $f(0) = g(0) = 0$, it follows that $c = 0$, hence $g(x) = 2f(x)$. Hence, $\arg th\left(\frac{\pi}{\sqrt{2\pi^2-y^2}}\right) = \frac{1}{2}\ln\left(\frac{(\pi+\sqrt{2\pi^2-y^2})^2}{\pi^2-y^2}\right) = \frac{1}{2}\ln\left(\frac{(\pi+\sqrt{2\pi^2-y^2})^2}{\pi+y}\right) - \frac{1}{2}\ln(\pi-y)$. Because the function $y \mapsto \frac{1}{2}\ln\left(\frac{(\pi+\sqrt{2\pi^2-y^2})^2}{\pi+y}\right)$ is continuous on $[0,\pi]$, it is Lebesgue integrable. On the other side, $\int_0^\pi \ln(\pi-y)dy = \int_0^\pi \ln(t)dt = \lim_{\epsilon\to 0}\int_\epsilon^\pi \ln(t)dt = \lim_{\epsilon\to 0}[t\ln(t)-t]_\epsilon^\pi = \pi\ln\pi - \pi$, with the change of variable $t = \pi - y$. So, the function $y \mapsto \ln(\pi-y)$ is Lebesgue integrable on $[0,\pi]$. Whence, the function $y \mapsto \arg th\left(\frac{\pi}{\sqrt{2\pi^2-y^2}}\right)$ is Lebesgue integrable on $[0,\pi]$. Therefore the function $y \mapsto \frac{\pi}{\sqrt{2\pi^2-y^2}}\arg th\left(\frac{\pi}{\sqrt{2\pi^2-y^2}}\right)$ is Lebesgue integrable on $[0,\pi]$, so $s(x,y)$ is Lebesgue integrable over Q. Considering the function $g(x,y) = \frac{(|x|+|y|)^2}{2\pi^2-x^2-y^2}$, then $g(x,y)$ is real-valued, even, essentially nonnegative and nonzero, and has the same zero which is $(0,0)$ of two-order as the function $s(x,y)$ in the domain Q. Also, $g(x,y)$ is Lebesgue integrable over Q. Indeed: $\lim_{(x,y)\to(\pm\pi,\pm\pi)}\frac{s(x,y)}{g(x,y)} = \frac{-1}{2}$, then $s(x,y) \underset{(\pm\pi,\pm\pi)}{\sim} \frac{-1}{2}g(x,y)$. Because $s(x,y)$ is Lebesgue integrable at the neighborhood of $(\pm\pi,\pm\pi)$, it follows that $g(x,y)$ is also Lebesgue integrable in this neighborhood. Hence, $g(x,y)$ is Lebesgue integrable over Q. Furthermore, $\frac{s(x,y)}{g(x,y)}$ is bounded in Q. In fact,

$$\frac{s(x,y)}{g(x,y)} = \frac{-(x^2+y^2)}{(|x|+|y|)^2} = \frac{-(|x|+|y|)^2+2|xy|}{(|x|+|y|)^2} = -1 + \frac{2|xy|}{(|x|+|y|)^2}.$$

Since $\forall(x,y) \in (0,\pi]\times(0,\pi]$, $2xy \le x^2+y^2$ then, $\forall(x,y)\in(0,\pi]\times(0,\pi]$, $\frac{2xy}{(x+y)^2} \le 1/2$. So $0 \le \lim_{(x,y)\to(0,0)}\frac{2xy}{(x+y)^2} \le 1/2$. Because $\lim_{(x,y)\to(0,0)}\frac{2xy}{(x+y)^2} = \begin{cases} 1/2 & \text{if } x=y \\ \alpha < 1/2 & \text{otherwise} \end{cases}$, then $\lim_{(x,y)\to(0,0)}\frac{s(x,y)}{g(x,y)} = \beta \in [-1+\alpha, -1/2]$. So $\frac{s(x,y)}{g(x,y)}$ is bounded on Q. Whence, $T_{d_n,d_n}(g)$ is a good preconditioner of $T_{d_n,d_n}(s)$. From Theorems 9.7.3 (see also [117]) we can conclude that the number of iterations of this PCG method [5], in order to reach the solution within a fixed accuracy $\epsilon > 0$, is

$$N_\epsilon < \frac{R}{2r} log\left(\frac{1}{\epsilon}\right) + 1, \tag{9.37}$$

where $R = \sup_{(x,y)\in Q}(s/g) = -1/2$ and $r = \inf_{(x,y)\in Q}(s/g) = -1$.

When using the new Theorems 9.7.4, we observe that the eigenvalues of $\mathcal{P}_{d_n,d_n}(s;g)$ are quite uniformly distributed on $[r,R] = \mathcal{ER}(s/g)$. This implies that the estimate of (9.37) is tight in the sense that we cannot expect a number N_ϵ of PCG iterations much less that $(R/2r)log(1/\epsilon)+1$. More precisely, see Figure 9.3 in section 9.8 for a graphic illustration.

In sections 9.8 and 9.9 we present some numerical results for preconditioning of ill-conditioned block Toeplitz matrices and we give some numerical evidences of ill-conditioned g-Toeplitz matrices and related g-circulant preconditioning.

NB: The references are ordered as follows: $[n]$, $[n^{(1)}]$, $[n^{(2)}]$, $[n^{(3)}]$, $[n^{(4)}]$, $[n^{(5)}]$, $[n+1]$, and so on, for $n \in \mathbb{N}^{\star}$.

9.8 Numerical results for ill-conditioned block Toeplitz matrices

In this chapter we have proved some properties about the generating functions $s(x)$ (respectively $s(x,y)$) of the symmetric Toplitz matrices $T_{d_n}(s)$ (respectively symmetric block Toeplitz matrices with symmetric Toeplitz blocks $T_{d_n,d_n}(s)$) which are equally distributed and equally localized as the collocation matrices A_{d_n} (respectively A_{d_n,d_n}) approximating elliptic boundary value problems and we have recalled many fundamental results (due to Serra Capizzano [121]) on the asymptotic distribution of the eigenvalues of preconditioned matrices of the form $\mathcal{P}_{d_n}(f;g)$. These results allow one to improve the convergence analysis of the PCG method defined in [118] for the iterative solution of symmetric nondefinite Toeplitz (respectively block Toeplitz with symmetric Toeplitz blocks) linear systems. Moreover, we have given a more precise description of the behavior of the PCG method devised in [117] for the solution of positive (negative) definite block Toeplitz linear systems. Finally, by combining this new spectral analysis with known properties [56, 137, 140] of the localization and the extremes of the spectrum of $\mathcal{P}_{d_n}(f;g)$, we can conclude that the function f/g describes the eigenvalues of $\mathcal{P}_{d_n}(f;g)$ as well as the the function f describes the eigenvalues of $T_{d_n}(f)$ [77]. In this way a concept of generating function is introduced even for this class of preconditioned matrices. The reason is to avoid the zeros and/or the unboundedness.

In order to illustrate, we consider in the symmetric Toeplitz case, the generating function having the same properties as the generating functions of Toeplitz matrices (obtained in the Inverse Multiquadric case (see subsection 9.3.2)) and any unbounded integrable function for the case of symmetric block Toeplitz matrices with symmetric Toeplitz blocks. In fact, let us consider the following three examples: $f(x) = \frac{-x^2}{\pi^2+x^2}$, $g(x) = \frac{2(1-\cos x)}{\pi^2+x^2}$; $f(x) = \frac{sign(x)\cdot x^2}{1+|x|}$, $g(x) = \frac{x^2}{1+|x|}$ and $f(x,y) = \frac{-(x^2+y^2)}{2\pi^2-x^2-y^2}$, $g(x,y) = \frac{(|x|+|y|)^2}{2\pi^2-x^2-y^2}$.

Case: $f(x) = \frac{-x^2}{\pi^2+x^2}$ **and** $g(x) = \frac{2(1-\cos x)}{\pi^2+x^2}$

First of all, f has one zero "0" of order 2. Furthermore, $\frac{f}{g}(x) = \frac{-x^2}{2(1-\cos x)}$. Here, "0" is not a zero of $\frac{f}{g}$, indeed: $\lim_{x\to 0}\frac{f}{g}(x) = -1$. If we denote again by $\frac{f}{g}$ the continuation by continuity of $\frac{f}{g}$ over $[-\pi,\pi]$ then, the study of the function $\frac{f}{g}$ over $I = [-\pi,\pi]$ provides the following properties

1. $\frac{f}{g}$ is nonpositive, symmetric and bounded on I;

2. $\min_{x\in[0,\pi]} \frac{f(x)}{g(x)} = -\frac{\pi^2}{4}$;

3. $\max_{x\in[0,\pi]} \frac{f(x)}{g(x)} = -1$;

4. $\frac{f}{g}$ is a decreasing function on $[0,\pi]$;

5. $\lim_{x\to 0^+} \left(\frac{f(x)}{g(x)}\right)' = 0$ (horizontal tangent at $x = 0^+$);

6. $\lim_{x\to \pi^-} \left(\frac{f(x)}{g(x)}\right)' = \frac{-\pi}{2} < 0$ (tangent at $x = \pi^-$ directed upward).

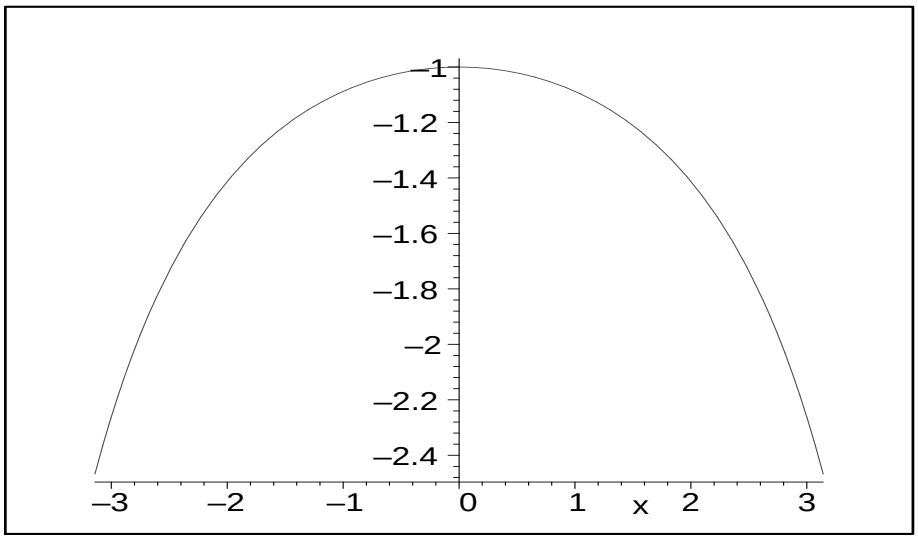

Figure 9.1: Eigenvalues of $\mathcal{P}_{d_n}(f;g)$, $f(x) = \frac{-x^2}{\pi^2+x^2}$ and $g(x) = \frac{2(1-\cos x)}{\pi^2+x^2}$, $d_n = 999$

Conclusion: Figure 9.1 shows that the eigenvalues of $\mathcal{P}_{d_n}(f;g)$, for $d_n = 999$, plotted with respect to a uniform grid points $x_k = \frac{k\pi}{1000}$, $k = 0, 1, 2, ..., 999$, form a curve which has the expected shape of f/g.

Case: $f(x) = \frac{sign(x)\cdot x^2}{1+|x|}$ and $g(x) = \frac{x^2}{1+|x|}$

Also in this case, "0" is a zero of order 2 of f. Next, $\forall x \in [-\pi, 0[\cup]0, \pi]$, $\frac{f}{g}(x) = sign(x)$. Let us denote again by $\frac{f}{g}$ the continuation by continuity of $\frac{f}{g}$ over the domain $[-\pi, \pi]$ then, Figure 9.2 shows perfect argument of the spectrum of the preconditioned matrix with the behavior of the function f/g. More precisely, notice that all the eigenvalues belong to the interval $(-0.999; 0.999)$ (because $\mathcal{P}_{d_n}(f;g) = T_{d_n}(f/g) + \Delta_{d_n}$ where $\{\Delta_n\}_n \sim_\lambda 0$) with the only exception of a few outliers, and the ones closest to zero (± 0.4) are not very close to zero.

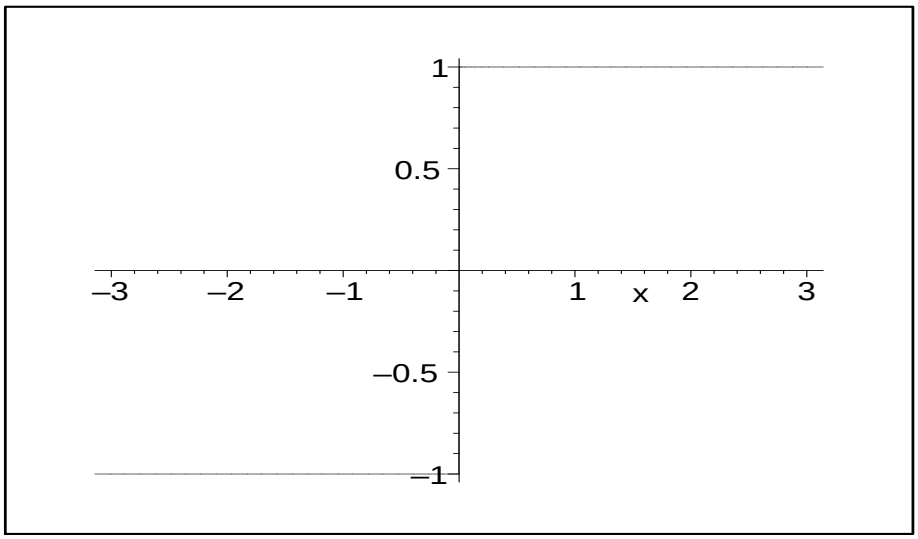

Figure 9.2: Eigenvalues of $\mathcal{P}_{d_n}(f;g)$, $f(x)=\frac{sign(x)\cdot x^2}{1+|x|}$, $g(x)=\frac{x^2}{1+|x|}$ and $d_n=350$.

<u>Case</u>: $f(x,y)=\frac{-(x^2+y^2)}{2\pi^2-x^2-y^2}$ and $g(x,y)=g(x,y)=\frac{(|x|+|y|)^2}{2\pi^2-x^2-y^2}$.

First of all, $(0,0)$ is a zero of order 2 of $f(x,y)$. Furthermore, as it was proved in subsection 9.7.2, the functions $f(x,y)$ and $g(x,y)$ are unbounded over the domain Q but, are Lebesgue integrable.

- **Brief study of the function $\frac{f}{g}$ over Q**

$\forall (x,y)\in[-\pi,\pi]\times[-\pi,\pi]$ and $(x,y)\neq(\pm\pi,\pm\pi)$, $\frac{f(x,y)}{g(x,y)}=\frac{-(x^2+y^2)}{(|x|+|y|)^2}$. Still by indicating by $\frac{f}{g}$ the continuation by continuity of $\frac{f}{g}$ over the domain $[-\pi,\pi]\times[-\pi,\pi]$ then, $\frac{f(x,y)}{g(x,y)}$ is a symmetric function, whence reduction of the study on $[0,\pi]\times[0,\pi]$. Hence $\frac{f(x,y)}{g(x,y)}=\frac{-(x^2+y^2)}{(x+y)^2}$. Next, $(0,0)$ is not a zero of $\frac{f}{g}$, indeed: $\lim\limits_{(x,y)\to(0,0)}\frac{f}{g}(x,y)=-\frac{1}{2}$. Setting $h(x,y)=\frac{f(x,y)}{g(x,y)}$, a direct computation gives

$$\frac{\partial h}{\partial x}(x,y)=-\frac{2y(x-y)}{(x+y)^3},\quad \frac{\partial h}{\partial y}(x,y)=-\frac{2x(y-x)}{(x+y)^3},\quad \frac{\partial^2 h}{\partial x^2}(x,y)=\frac{4y(x-2y)}{(x+y)^4},$$

$$\frac{\partial^2 h}{\partial y^2}(x,y)=\frac{4x(y-2x)}{(x+y)^4},\quad \frac{\partial^2 h}{\partial x\partial y}(x,y)=\frac{\partial^2 h}{\partial y\partial x}(x,y)=-\frac{2(x^2+y^2-4xy)}{(x+y)^4}$$

then the derivative function and the Hessian matrix of h are given by

$$D(h)(x,y)|_{(x_0,y_0)}=\begin{pmatrix}-\frac{2y_0(x_0-y_0)}{(x_0+y_0)^3}\\ -\frac{2x_0(y_0-x_0)}{(x_0+y_0)^3}\end{pmatrix},\quad \mathbf{H}(h)(x_0,y_0)=\begin{bmatrix}\frac{4y_0(x_0-2y_0)}{(x_0+y_0)^4} & -\frac{2(x_0^2+y_0^2-4x_0y_0)}{(x_0+y_0)^4}\\ -\frac{2(x_0^2+y_0^2-4x_0y_0)}{(x_0+y_0)^4} & \frac{4x_0(y_0-2x_0)}{(x_0+y_0)^4}\end{bmatrix}$$

Furthermore,

$$\begin{aligned} det\left(\mathbf{H}(h)(x_0,y_0)-\lambda I_2\right) &= \left(\frac{4y_0(x_0-2y_0)}{(x_0+y_0)^4}-\lambda\right)\left(\frac{4x_0(y_0-2x_0)}{(x_0+y_0)^4}-\lambda\right) \\ &- \frac{4(x_0^2+y_0^2-4x_0y_0)^2}{(x_0+y_0)^8} \\ &= \lambda^2-\lambda\left[\frac{4y_0(x_0-2y_0)+4x_0(y_0-2x_0)}{(x_0+y_0)^4}\right] \\ &+ \frac{16x_0y_0(x_0-2y_0)(y_0-2x_0)}{(x_0+y_0)^8}-\frac{4(x_0^2+y_0^2-4x_0y_0)^2}{(x_0+y_0)^8} \\ &= \lambda^2+8\frac{x_0^2-x_0y_0+y_0^2}{(x_0+y_0)^4}\lambda-4\frac{x_0^4-2x_0^2y_0^2+y_0^4}{(x_0+y_0)^8} \\ &= \lambda^2+8\frac{x_0^2-x_0y_0+y_0^2}{(x_0+y_0)^4}\lambda-4\frac{(x_0^2-y_0^2)^2}{(x_0+y_0)^8}. \end{aligned}$$

If we denote by $S=-8\frac{x_0^2-x_0y_0+y_0^2}{(x_0+y_0)^4}$ and $P=-4\frac{(x_0^2-y_0^2)^2}{(x_0+y_0)^8}$ the sum and the product of roots of the polynomial $det\left(\mathbf{H}(h)(x_0,y_0)-\lambda I_2\right)$, since $P\leq 0$, then these roots are of different signs if $x_0\neq y_0$, one deduces that the eigenvalues of the Hessian matrix $\mathbf{H}(h)(x_0,y_0)$ are of different signs if $x_0\neq y_0$. If $x_0\approx y_0$, then $P\approx 0$ and at least one of the eigenvalues is close to zero. More precisely, one can summarize the properties of the function f/g as follows

1. $\frac{f}{g}$ is nonpositive, symmetric and bounded on Q;

2. $\inf\limits_{(x,y)\in[0,\pi]^2}\frac{f(x,y)}{g(x,y)}=-1$;

3. $\sup\limits_{(x,y)\in[0,\pi]^2}\frac{f(x,y)}{g(x,y)}=-\frac{1}{2}$;

4. If $x_0\neq y_0$ then, the Hessian matrix $(\mathbf{H}(f/g)(x_0,y_0))$ of $\frac{f}{g}$ is nonpositive definite and nonnegative definite so, the function f/g admits a **saddle point** (x_1,y_1) in $[0,\pi]\times[0,\pi]$, that is: $\frac{f}{g}(x_1,y_1)=\min\limits_{x\in[0,\pi],y\in[0,\pi]}\max\frac{f}{g}(x,y)=\max\limits_{y\in[0,\pi],x\in[0,\pi]}\min\frac{f}{g}(x,y)$. If $x_0\approx y_0$, then $P\approx 0$ and at least one eigenvalue of $(\mathbf{H}(f/g)(x_0,y_0))$ is close to zero;

5. $\lim\limits_{(x_0,y_0)\to(0^+,0^+)}D\left(\frac{f}{g}\right)(x,y)|_{(x_0,y_0)}=(\infty,\infty)^T$;

6. $\lim\limits_{(x_0,y_0)\to(\pi^-,\pi^-)}D\left(\frac{f}{g}\right)(x,y)|_{(x_0,y_0)}=(0,0)^T$.

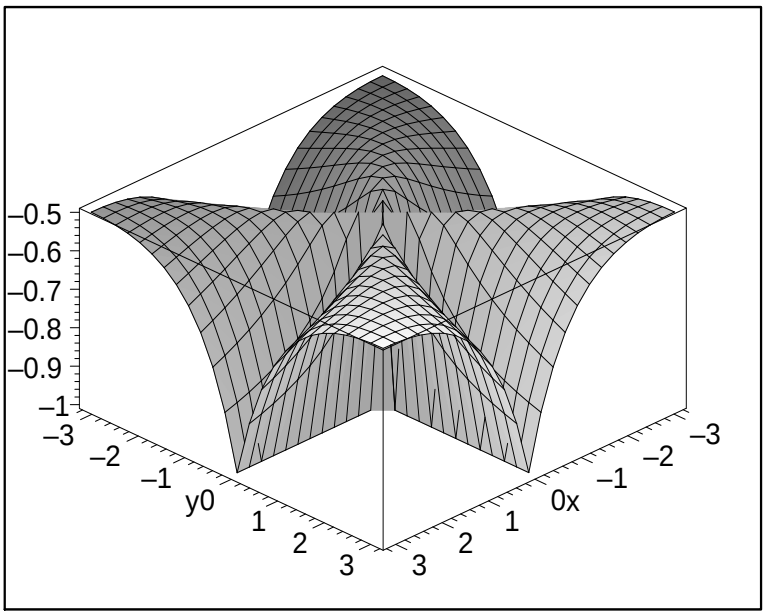

Figure 9.3: Eigenvalues of $\mathcal{P}_{d_n,d_n}(f;g)$, $f(x,y) = \frac{-(x^2+y^2)}{2\pi^2-x^2-y^2}$, $g(x,y) = \frac{(|x|+|y|)^2}{2\pi^2-x^2-y^2}$ and $d_n = 99$.

Conclusion: Figure 9.3 shows that the eigenvalues of the preconditioned matrix $\mathcal{P}_{d_n,d_n}(f;g)$ for $d_n = 99$, plotted with respect to a uniform grid points $x_{jk} = (\frac{j\pi}{100}, \frac{k\pi}{100})$, $j,k = 0,1,2,...,99$, are concentrated in the interval $[-1,-\frac{1}{2}]$ except few of them which are outliers. Indeed, one can write $\mathcal{P}_{d_n,d_n}(f;g) = T_{d_n,d_n}(f/g) + \Delta_{d_n,d_n}$ where $\{\Delta_{d_n,d_n}\}_n \sim_\lambda 0$. Because the spectrum of $T_{d_n,d_n}(f/g)$ forms a curve which has the expected shape of f/g. It follows that Figure 9.3 shows perfect argument of the spectrum of the preconditioned matrix $\mathcal{P}_{d_n,d_n}(f;g)$ with the behavior of the function $\frac{f}{g}$.

9.9 Numerical evidences of g-Toeplitz matrices and related g-Circulant preconditioning

In chapter 7 we have studied by regularizing technique the singular values of matrix sequences obtained by preconditioning g-Toeplitz sequences associated with a given integrable function via g-circulant sequences. In this section, we present some numerical evidences of ill-conditioned g-Toeplitz matrices and related g-circulant preconditioning. In fact, we consider an $n \times n$ linear system $Af = g$, where A is a g-Toeplitz matrix. Aimed of providing numerical evidences to the theoretical results, we analyze

(i) the distribution of the singular values of g-Toeplitz matrices and related g-circulant preconditioned matrices (subsection 9.9.1),

(ii) the effectiveness of the g-circulant preconditioning for the solution of the corresponding g-Toeplitz linear system $Ax = b$ (subsection 9.9.2), and

(iii) a possible real application related to a 2D inverse problem in imaging (subsection 9.9.3).

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 $A_{k,1}$ 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 [$150^{(1)}$, 38]
- Test 2 $\quad A_{k,1} = k^{-2}$
  Strictly positive Wiener generating function, Well-conditioned [$150^{(1)}$, 38]
- Test 3 $\quad A_{:,1} = (2, -1, 0, \ldots, 0)^t$
  Sparsely vanishing trigonometric polynomial generating function $f(x) = 2 - 2\cos x$, Ill-conditioned, Zero valued (order 2) at the origin [163]
- Test 4 $\quad A_{:,1} = (20, -15, 6, -1, 0, \ldots, 0)^t$
  Sparsely vanishing trigonometric polynomial generating function $f(x) = (2 - 2\cos x)^3$, Ill-conditioned, Zero valued (order 6) at the origin
- Test 5 $\quad A_{:,1} = (\pi^2/2, -2, 0, -2/9, 0, -2/25, 0, \ldots, 0, -2/(k-1)^2, 0, \ldots)^t$
  Sparsely vanishing generating function $f(x) = \pi|x|$, Ill-conditioned, Zero valued (order 1) at the origin [$58^{(4)}$]
- Test 6 $\quad A_{:,1} = (2, 0, 2\frac{1}{3}, 0, -2\frac{1}{15}, 0, 2\frac{1}{35}, 0, \ldots, 0, (-1)^{(k+1)/2}\, 2/((k-1)^2 - 1), 0, \ldots)^t$
  Sparsely vanishing generating function $f(x) = \pi|\cos x|$, Ill-conditioned, Zero valued (order 1) at $\pi/2$ [$58^{(4)}$]

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 section 7.5.2, for any g-Toeplitz test matrix we consider both (i) the Natural g-circulant preconditioner and (ii) the Optimal g-circulant preconditioner (see [$150^{(1)}$, 38] for the classical Toeplitz case $g = 1$). The numerical tests have been developed with MatLab, and the singular value decomposition has been computed by the built-in Matlab function svd().

9.9.1 The distribution of the singular values

First, we plot the distribution of the singular values of the $n \times n$ g-Toeplitz matrix A, 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 co-prime for $g = 3$ and $g = 7$, and are not co-prime for $g = 2$). In particular, we have:

I) Fig. 9.5 and Fig. 9.6 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. 9.7 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 7.5-7.6-7.7 and 7.8-7.9-7.10. Indeed, for $g > 1$ and sparsely vanishing generating functions, we have:

(i) regarding the g-Toeplitz matrix A, the first n/g singular values are positive, and equals to $\sqrt{\widehat{|f|^{(2)}}(x)}$, while the remaining $n - n/g$ are null, as stated by 7.6 (see the blue line in Fig. 9.5, 9.6 and 9.7);

(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 Fig. 9.5 and 9.6), and, if $\gamma > 1$, the first n/γ 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 $\sqrt{\widehat{|h|^{(3)}}(x)}$, while the remaining $n - n/\gamma$ are null, as stated by 7.9 (see the green line in Figg. 9.7).

In particular, since $n = 1000$, then $\gamma = 1$ for $g = 3, 7$, and $\gamma = 2$ for $g = 2$: in Fig. 9.5 and Fig. 9.6 the singular values of both the natural and optimal g-circulant preconditioners are bounded away from zero in the well-conditioned test cases 1 and 2, and sparsely vanishing in the ill- conditioned test cases 3, 4,5 and 6, while one half of the singular values are always null in Fig. 9.7 (green lines).

It is now interesting to analyze the distribution of the preconditioned matrix. Any coprime case (Figg. 9.5 and 9.6, red line) gives rise to a good clustering at unity, in the first n/g singular values, while the remaining ones are null. This is a result which was expected in the light of Theorem 7.5.1: 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 n/g, so that the rank of $P^{-1}A$ is just no larger than n/g). This good clustering at unity of the preconditioned matrix $P^{-1}A$ occurs in both the well-conditioned case (see, in Figg. 9.5 and 9.6. the cases Test1 and Test 2, where the preconditioners have no vanishing singular values) and the ill-conditioned case (see, again in Figg. 9.5 and 9.6, 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 w.r.t. 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. 9.5 and 9.6.

The situation is different in the non-coprime case, as Fig. 9.7 shows. Before going on, according to subsection 7.4.1, we mention that in this case instead of the inverse P^{-1} we have to consider the Moore-Penrose generalized inverse $P^{\dagger}$, P being 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 informations which were contained in the related g-Toeplitz matrix A, which means that P becomes less correlate to A, and a good clustering is no more possible. This is well explained by Fig. 9.7, 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.

9.9.2 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 $A^\star Ax = A^\star b$.
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, and 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 $i = 0, \ldots, n-1$, is given by $\cos(g\pi i/n)$, so that the first 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 $RelRes = \|A^\star Ax_k - A^\star b\|/\|A^\star b\|$ and the relative error on the restored signal $RelErr = \|x_k - x^\dagger\|/\|x^\dagger\|$, where $x^\dagger$ is the projection on $N(A)^\perp$ of the true data (which is obviously the best possible restoration). Since n/g is the rank of A, to obtain $x^\dagger$ we compute $x^\dagger = \tilde{V}\tilde{V}^\star x$, where $\tilde{V}$ is the $n \times (n/g)$ matrix given by the first n/g columns of V, being V the orthogonal matrix of the singular value decomposition $A = U\Sigma V^\star$.

By using the built-in Matlab function pcg() within the first 100 iterations, in Table 9.1 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 $\|b_\eta - b\|/\|b\| =$ is $10^{-4}\%$; in the central columns the relative noise level $\|b_\eta - b\|/\|b\| =$ is $10^{-1}\%$.

As we can observe, the optimal g-circulant preconditioned conjugate gradient method does not allow to obtain better results than 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. 9.8 we plot the first 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 [$58^{(3)}$], 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.

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	1.2376e-007	1.3158e-007
Relative Error	4.29e-006	1.20e-008	1.89e-005	1.88e-005	0.0183	0.0183
Test 2						
Iter. Number	47	13	47	36	78	79
Relative Residual	2.87e-010	2.89e-010	3.48e-010	8.94e-010	9.0800e-009	7.4287e-009
Relative Error	4.64e-010	4.72e-010	8.75e-006	8.75e-006	0.0088	0.0088
Test 3						
Iter. Number	100*	2	100*	2	100*	2
Relative Residual	0.0214	1.10e-016	0.0213	1.12e-016	0.0213	1.1900e-016
Relative Error	0.0181	1.19e-016	0.0181	2.98e-006	0.0183	0.0031
Test 4						
Iter. Number	100*	14	100*	20	100*	20
Relative Residual	1.93e-005	3.78e-010	1.93e-005	8.39e-010	2.3375e-005	9.5010e-010
Relative Error	7.48e-006	2.40e-010	7.62e-006	2.20e-006	0.0021	0.0021
Test 5						
Iter. Number	100*	8	100*	37	98	99
Relative Residual	6.08e-005	1.04e-010	5.97e-005	9.93e-010	8.9519e-005	1.3370e-008
Relative Error	3.22e-005	8.93e-011	3.17e-005	2.76e-006	0.0027	0.0027
Test 6						
Iter. Number	77	4	75	10	76	72
Relative Residual	1.92e-004	2.64e-013	1.93e-004	6.14e-010	2.0942e-004	7.8879e-010
Relative Error	1.04e-004	2.65e-013	1.05e-004	7.57e-006	0.0076	0.0076

Table 9.1: $g = 3$: Best relative residual $\|A^\star A x_k - A^\star b_\eta\|/\|A^\star b_\eta\|$, with corresponding iteration number k and relative restoration error $\|x_k - x^\dagger\|/\|x^\dagger\|$, with respect to different noise levels $\delta = \|b - b_\eta\|/\|b\|$ of the CGNR and PCGNR with optimal g-circulant preconditioner.

9.9.3 Two dimensional g-Toeplitz matrices for structured shift-variant image deblurring

We conclude the numerical section by introducing a real problem of image deblurring [$9^{(1)}$] 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

$$g(x) = \int_{\mathbb{R}^2} h(x,u)\, f(u)\, du \tag{9.38}$$

where the integral kernel $h \in L^2(\mathbb{R}^2 \times \mathbb{R}^2)$ 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 of 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 to the weights $h(x,\cdot)$. Among the proposed mathematical models, the simplest and most common blurring operator (9.38) 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

$$h(x,u) = h_I(x-u)\,. \tag{9.39}$$

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 (9.38) 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

$$h(x, u) = \tilde{h}_I(b(x), c(u))$$

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

$$g(x) = \int_U h(x, u) f(u) du = \int_U h_I(b(x) - c(u)) f(u) du. \tag{9.40}$$

that is

$$\tilde{g}(\tilde{x}) = \int_{c(U)} h_I(\tilde{x} - \tilde{u}) \tilde{f}(\tilde{u}) d\tilde{u} \tag{9.41}$$

where $\tilde{x} = b(x)$, $\tilde{u} = c(u)$, $\tilde{g} = g \circ b^{-1}$, $\tilde{f} = (c^{-1})' \cdot f \circ c^{-1}$. 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 (9.41). 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 to 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 $h_I(\rho, \theta) = 1/\sigma$ 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

$$A_{i,j} = h(id, jd) = h_I(b(id) - c(jd)) = h_I(ivd - jgd)\,. \tag{9.42}$$

Figure 9.4: Non-perpendicular imaging system geometry.

If $b(x) = x$, then the PSF matrix A is a g-Toeplitz matrix. 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. 9.4). We remark that this is the classical scenario of high-way traffic flow control systems.

A numerical simulation is shown in Fig. 9.9, 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 linear transformation w.r.t. 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. Toepliz) case. In Fig. 9.9, 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 $O(n^2 \log n)$ computations.

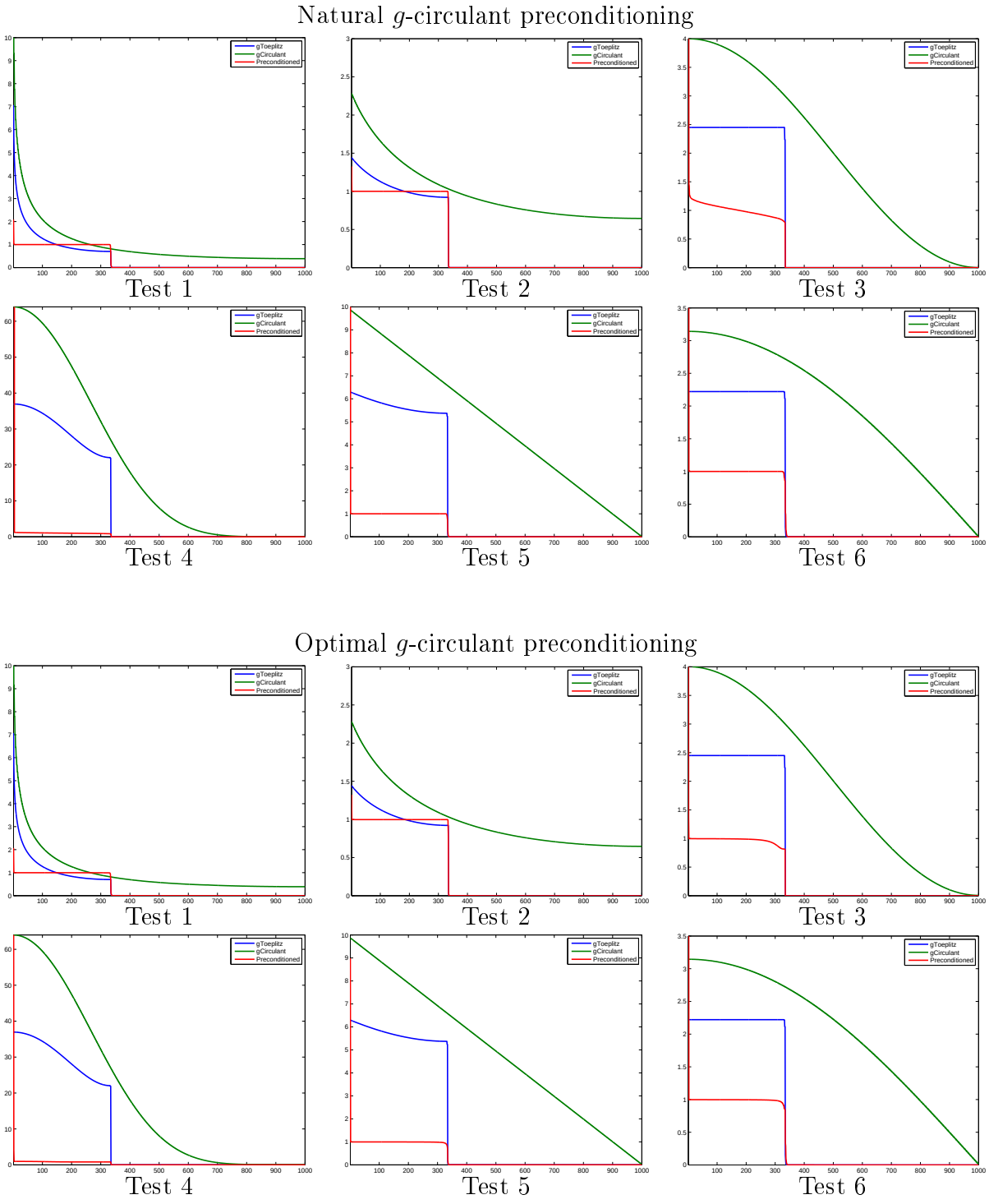

Figure 9.5: $g = 3$ (coprime case) - Singular values of g-Toeplitz matrices A, Natural (top) and Optimal (bottom) g-circulant preconditioners P and corresponding preconditioned matrices $P^{-1}A$.

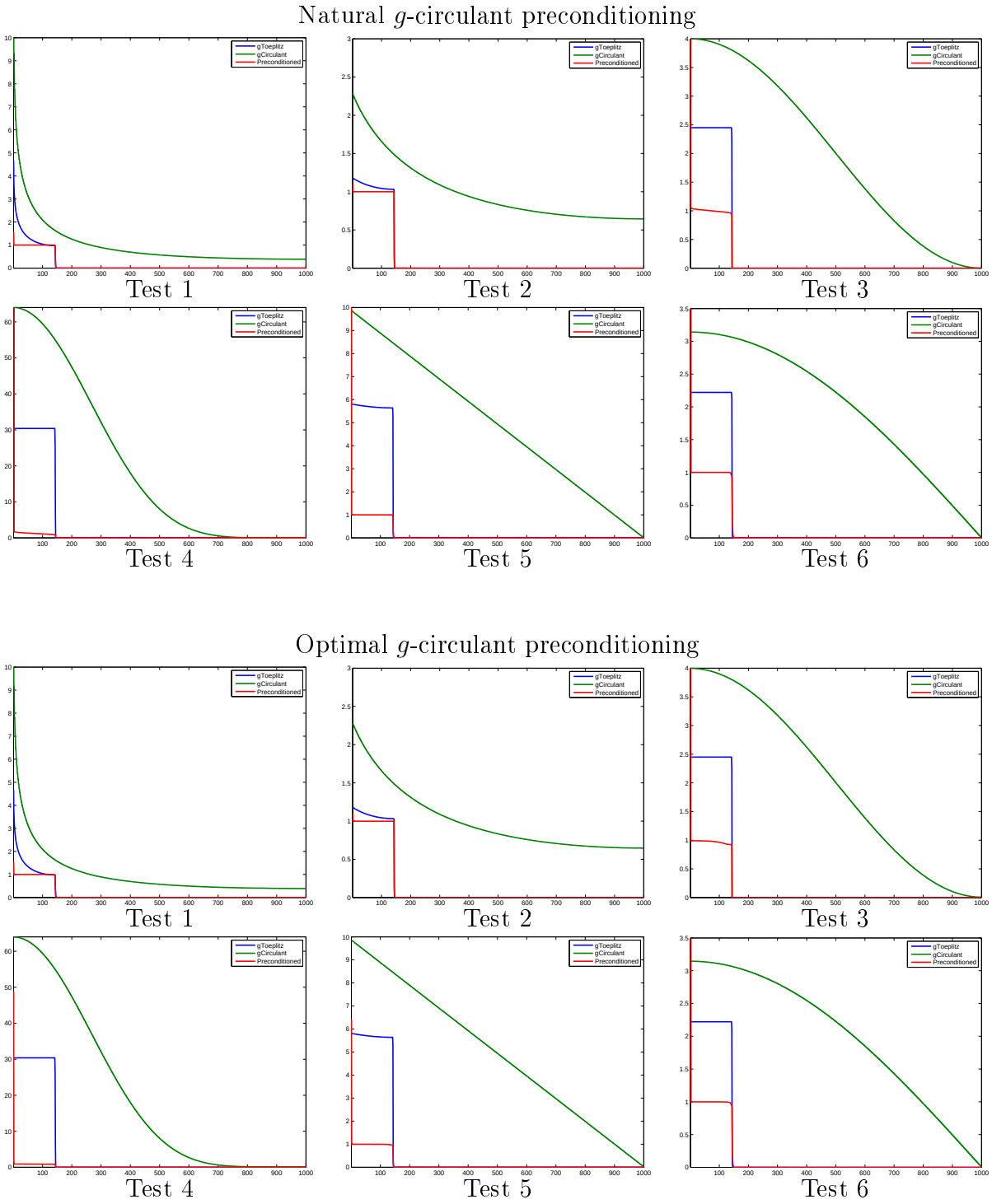

Figure 9.6: $g = 7$ (coprime case) - Singular values of g-Toeplitz matrices A, Natural (top) and Optimal (bottom) g-circulant preconditioners P and corresponding preconditioned matrices $P^{-1}A$.

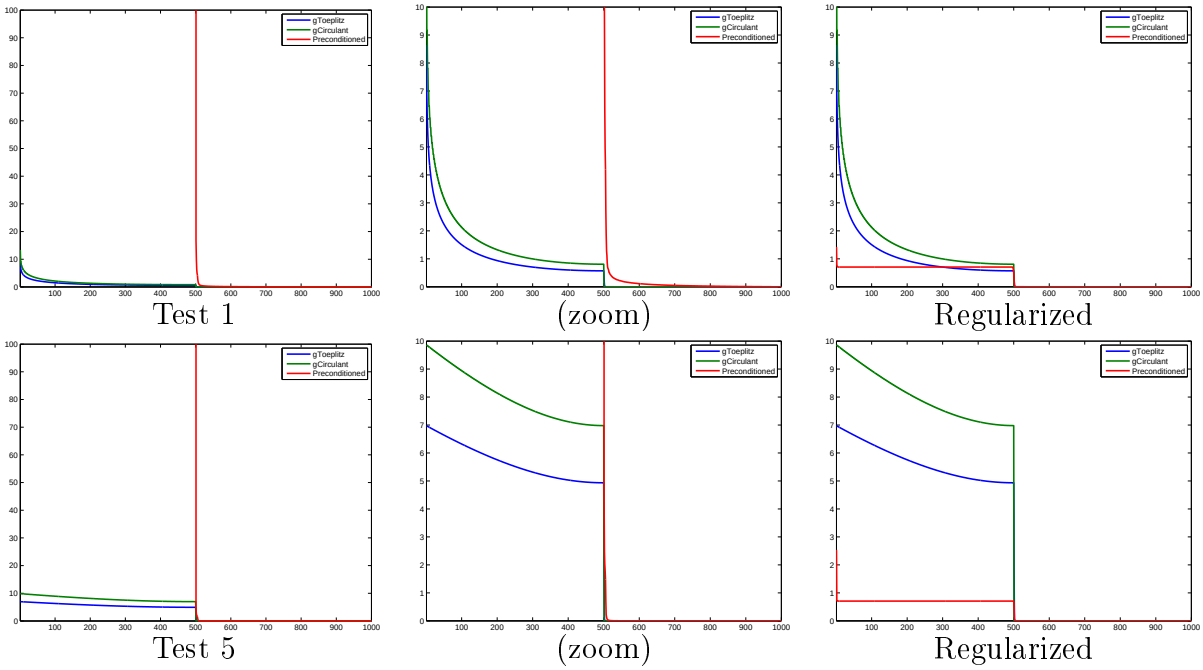

Figure 9.7: $g = 2$ (non-coprime case) - Singular values of g-Toeplitz matrices A, optimal g-circulant preconditioners P and corresponding preconditioned matrices $P^{\dagger}A$ (left), zoom on the small values (center), and analogous spectral distributions related to the regularized preconditioners (right).

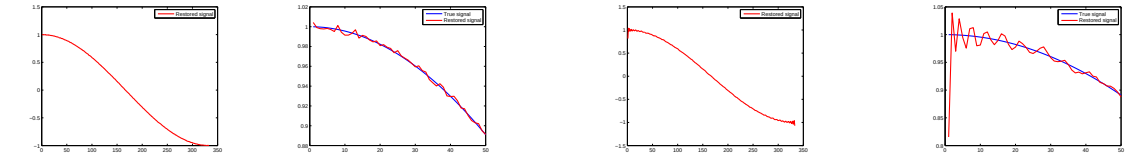

Figure 9.8: Restored signal with (P)CG on the normal equations (1% of data noise, $g = 3$). Left: without preconditioning. Right: with Optimal g-circulant preconditioning.

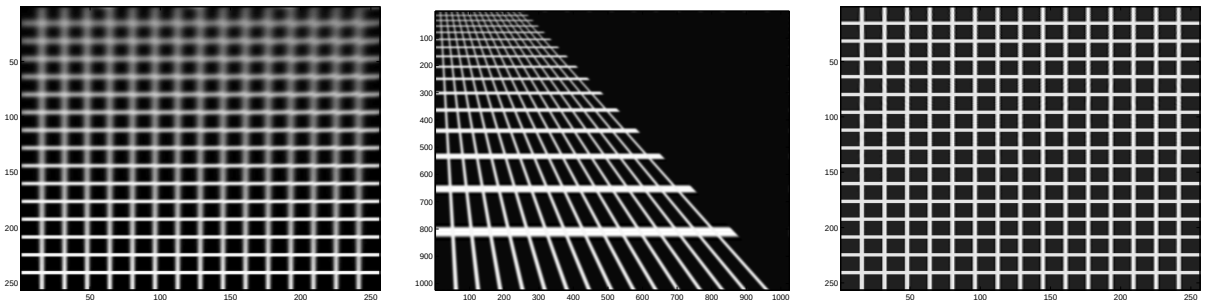

Figure 9.9: Shift-variant blurred data, projected data (shift-invariant blur), deblurred data.

Chapter Ten

General Conclusions

This dissertation has described a detailed study of preconditioning, approximation and spectral analysis of special classes of large structured linear systems. The structured linear systems considered in this Thesis arise in many applications. For examples, in solutions to differential and integral equations, spline functions, problems and method in physics, mathematics, digital signal processing, such as linear prediction and estimation [86, 97, 98], image restoration [66], the approximation by radial basis functions (RBFs) of constant coefficients elliptic boundary value problems, in wavelet analysis [50] and subdivision algorithm or, equivalently, in the associated refinement equations, see [58] and references therein (cases of Toeplitz, g-Toeplitz and g-circulant systems); in many applications involving the discrete Fourier transform (DFT) and the study of cyclic codes for error correction (case of circulant systems). The spectral analysis considered here can be viewed in the sense of singular values, eigenvalues and eigenvectors of such related matrices. The problem of preconditioning and approximation approached in this Thesis concerns: the optimal approximation via the Korovkin-type theory, the distribution results in the sense of the singular values and the approximation in the sense of the Weyl-Tyrtyshnikov equal distribution.

Concerning the spectral analysis, we have determined the eigenvalues, singular values and eigenvectors of circulant and g-circulant matrices, have done a detailed study of the optimal approximation by the Korovkin-type theory for finite Toeplitz operators via matrix algebra in the case of Toeplitz sequences and we have provided the distribution result in the sense of singular values for the g-Toeplitz sequences.

Our second solved problem is specifically based on the preconditioning g-Toeplitz sequences via g-circulants and that of collocation matrices approximating elliptic boundary value problems. In the case of the approximation of g-Toeplitz sequences, we have studied the singular values of matrix sequences obtaining by preconditioning. The main point was that the standard preconditioning works only in the classical setting but, the surprise was that when, the stepsize g is positive, a regularizing preconditioning can be obtained by a clever choice of the g-circulant sequences. Furthermore, the results on the Weyl-Tyrtyshnikov equal distribution and the Perron-Frobenius theory were fundamental for the determination of preconditioners and a detailed study of the spectral radii of collocation matrices approximating elliptic boundary value problems. From this study, it followed that one of the advantages of meshless methods based on the radial basis functions with respect to another, is high decrease of computational volume that arises when changing multi-dimensions to one dimension, further, the use of the globally supported radial basis functions, reaches to the large linear systems, poorly condition number and full matrices.

An application of the preconditioned conjugate gradient method to symmetric block Toeplitz matrices with symmetric Toeplitz blocks generated by unbounded functions, which are equally distributed and equally localized as the collocation matrices, associated with some numeric results are presented in chapter 9 of this dissertation.

Finally, the difficulties met during the study of the g-circulant matrices and collocation matrices have obliged us to restrict our researches on the singular values and eigenvalues in the case of g-circulant and to impose some requirements on the shape parameter "c" figuring in the radial basis functions for the collocation matrices. In the future works, we will delete this constraint and then will try to look for another theory in order to study the spectral

radii of such collocation matrices. The study of eigenvectors of the g-circulant matrices and the distribution result in the sense of eigenvalues of the related g-Toeplitz sequences will also be the subject of our researches.

References

[1] A. Aricò, Marco Donatelli, S. Serra-Capizzano, "V-cycle optimal convergence for certain (multilevel) structured linear systems", SIAM J. Matrix Anal. Appl., 26(2004), pp. 186 – 214.

[2] F. Avram, "On bilinear forms on Gaussian random variables and Toeplitz matrices, Probab. Th. Related Fields" 79(1988), 37 – 45.

[3] W.E., Arnoldi, "The principe of minimized iteration in the solution of the matrix eigenvalue problem. Quart. Appl. Math. " 9(1951), pp. 17 – 29.

[4] O. Axelsson, "Solution of linear systems of equations: Iterative methods". In Barker (1977).

[5] O. Axelsson and V. Barker, "Finite Element Solution of Boundary Value Problems, Theory and Computation", Academic, New York, 1984.

[6] O. Axelsson and G. Lindskog, "On the rate of convergence of the preconditioned conjugate gradient method", Numer. Math., 52(1986), PP. 499 – 523.

[7] O. Axelsson and M. Neytcheva, "The algebraic multilevel iteration methods-theory and applications ", in proceedings of the 2nd International Colloquiun on Numerical Analysis (D. Bainov, Ed.), Plovdiv, Bulgaria, Aug. 1993, PP. 13 – 23.

[8] E.H. Bareiss, "Numerical solution of linear equations with Toeplitz system and vector Toeplitz matrices", Numer. Math., 13(1969), pp. 404 – 424.

[9] A. Berman and R.J. Plemons, "Nonnegative Matrices in the Mathematical Sciences". Classics in Applied Mathematics. SIAM, Philadelphia, PA, 1994.

[9$^{(1)}$] M. Bertero and P. Boccacci, "Introduction to Inverse Problems in Imaging", Institute of Physics Publ., Bristol, 1998.

[10] D. Bini, M. Capovani, "Spectral and computational properties of band symmetric Toeplitz matrices." Linear Algebra Appl., 52/53 1983, 99 – 126.

[11] D. Bini, F. Di Benedetto, "A new preconditioner for the parallel solution of positive definite Toeplitz linear systems." Proc. 2nd SPAA conf., Crete (Greece), 1990, pp. 220 – 223.

[12] D. Bini, P. Favati, "On a matrix algebra related to the discrete Hartley transforms". SIAM J. Math. Anal. Appl. 14(1993), 500 – 507.

[13] *Å* Börck, "Least squares methods". In: Ciarlet, Lions, Eds. (1990), 465 – 647. Linear Algebra Appl., 52/53 1983, 99 – 126.

[14] A. Bottcher, S. Grudsky, E. Maksimenko, "The Szögo and Avram-Parter theorems for general test functions", Comptes Rendus Acad. Sci. Paris-Ser. I, in print.

[15] A. Bottcher, J. Gutiérrez-Gutiérrez, P. Crespo, "Mass concentration in quasicommutators of Toeplitz matrices", J. Comp. Appl. Math., 205(2007), pp. 129 – 148.

[16] A. Bottcher, B. Silbermann, "Introduction to Large Truncated Toeplitz Matrices", Springer-Verlag, New York 1999.

[17] E. Bozzo, "Matrix Algebras and Discrete Transforms", PhD Thesis in Comput. Sci., Dip. Inf., Pisa 1994.

[18] D. Braess, "Finite element", Berlin, Heidelberg, New York, Springer-Verlag. 1997.

[19] J.H. Bramble, "Multigrid Methods". Harlow: Longman, 1993.

[20] A. Brandt, "Multi-level adaptative solutions to boundary value problems". Math. of Comput. 31, 333 – 390, 1997.

[21] R.P. Brent, "Algorithms for minimizations without Derivatives". Englewood Cliffs, N. J.: Prentice-Hall, 1973.

[22] C. Brezinski, "Algorithmes d'Accélérations de la convergence. Etude Numrique". Paris: Edition Technip. 1978.

[23] C. Brezinski, "Quasi-Newton-methods and their applications to function minimization". Math. Comput. 21, 368 – 381, 1967.

[24] C. Brezinski, "The convergence of a class of double rank minimization algorithms: 1. General considerations, 2. The new algorithm"". J. Inst. Math. Appl. 6, 76 – 90, 222 – 231, (1970).

[24$^{(1)}$] P. Brianzi, F. Di Benedetto and C. Estatico, "Improvement of space-invariant image deblurring by preconditioned Landweber iterations", SIAM J. Sci. Comp., 2008, 30 : 1430 – 1458.

[25] W. L. Briggs, "A Multigrid Tutorial". Philadelphia: SIAM (1987).

[26] J.R. Bunch and B.N. Parlett, "Direct methods for solving symmetric indefinite systems of linear equations", SIAM J. Numer. Anal. 8, pp. 639 – 655, 1971.

[27] P. Businger, G.H. Golub "Linear least squares solutions by Householder transformations". Numer. Math. 7, 269 – 276, 1965.

[28] B.L. Buzbee, G.H. Golub, C.W Nielson, "On direct methods for solving poisson's equations". SIAM J. Numer. Anal. 7, 627 – 656, 1970.

[29] S. Carnachan Naqvi and J.J. Mc Donald, "The Combinatorial structure of eventually nonnegative matrices". The Electronical Journal of Linear Algebra 9, 255 – 269, 2002.

[30] R.H. Chan, "Circulant preconditioners for Hermitian Toeplitz system", SIAM J. Matrix Anal. Appl., 10(1989), pp. 542 – 550.

[31] R.H. Chan, "Toeplitz preconditioners for Toeplitz systems with nonnegative generating functions". IMA J. Numer. Anal. 11(1991), pp. 333 – 345.

[33] R.H. Chan and T.F. Chan, "Circulant preconditioners for elliptic problems", Tech. Report, Department of Mathematics, University of California, Los Angeles, CA, Dec. 1990.

[34] R.H. Chan, T.F. Chan and C. Wong, "Cosine transform based preconditioners for total variation minimization problems in image processing". Proc. 2nd IMACS conf. on Iterative Methods in Linear Algebra, Vassilevsky, P. (ed), Blagoevvgrad (Bulgaria), pp. 311 – 329, 1995.

[35] R.H. Chan and W. Ching, "Toeplitz-circulant preconditioners for Toeplitz systems and their applications to queuing network with batch arrivals", SIAM J. Sci. Comput., 17 : 762 – 772, (1996).

[36] R.H. Chan and M.K. Ng, "Conjugate gradient methods for Toeplitz systems", SIAM Rev., 38(1996), pp. 427 – 482.

[37] R.H. Chan and G. Strang, "Toeplitz equations by conjugate gradients with circulant preconditioner", SIAM J. Sci. Statist. Comput., 10(1989), pp. 104 – 119.

[38] T.F. Chan, "An optimal circulant preconditioner for Toeplitz systems", SIAM J. Sci. Statist. Comput., 9(1988), pp. 766 – 771.

[39] J.J. Clement and C. Perea, "Some Comparison theorems for weak nonnegative Splitting of bounded operators", Linear Algebra Appl., 275 – 276, (1998).

[40] R.H. Chan and X. Jin, "A family of block preconditioners for block systems". SIAM J. Sci. Statist. Comput. 13(1992) 1218 – 1235.

[41] R.H. Chan, X. Jin and M.C. Yeung, "The circulant operator in the Banach algebra of matrices". Linear Algebra Appl. 149(1991) 41 – 53.

[42] R.H. Chan and P. Tang, "Fast band-Toeplitz preconditioners for Hermitian Toeplitz systems". SIAM J. Sci. Comput. 15(1994) 164 – 171.

[43] R.H. Chan and M.C. Yeung, "Circulant preconditioners for Toeplitz matrices with positive continuous generating functions". Math. Comput. 58(1992) 233 – 240.

[44] R.H. Chan and M.C. Yeung, "Jackson's theorem and circulant preconditioned Toeplitz systems". J. Approx. Theory 70(1992) 191 – 205.

[45] R.H. Chan and M.C. Yeung, "Circulant preconditioners constructed from kernels". SIAM J. Numer. Anal. 29(1992) 1093 – 1103.

[46] L.F. Chaparro and E.I. Jury, "Rational approximation of 2-D linear discrete systems", IEEE Trans. Acoust., Speech Signal Process., 30(1982), pp. 766 – 787.

[47] J. Cullum and R.A. Willoughby, "Lanczos Algorithms for Large Symmetric Eigenvalue Computations, Vol.I: Theory, Vol.II: Programs. Progress in Scientific Computing, Vol. 3, 4". Bassel: Birkhäusser, 1985.

[48] E. Cheney, "Introduction to Approximation Theory", Mc Graw-Hill, New York, 1966.

[49] J.W. Daniel, W.B. Gragg, L. Kaufmann and G.W. Stewart, "Reorthogonalization and stable algo. for updating the Gram-Schmidt QR factorization". Math. Comput. 30(1976), pp. 772 – 795.

[50] I. Daubechies, "Ten Lectures on Wavelets". CBMS-NSF Regional Conference Series in Applied Mathematics, 61, SIAM, Philadelphia, 1992.

[51] P. Davis, "Circulant Matrices", J. Wiley and Sons, New York, 1979.

[52] Fabio Di Benedetto, "Analysis of preconditioning techniques for ill-conditioned Toeplitz matrices", SIAM J. Sci. Comput. 16(682 – 697), (1995).

[53] Fabio Di Benedetto, "Preconditioning of block Toeplitz matrices by sine transforms", SIAM J. Sci. Comput. 18(495 – 515), (1997).

[54] Fabio Di Benedetto and S. Serra Capizzano, "A unifying approach to abstract matrix algebra preconditioning", Numer. Math. in press.

[55] Fabio Di Benedetto and S. Serra Capizzano, "Optimal and superoptimal matrix algebra operators". TR nr. 360, Dept. of Mathematics-Univ. of Genova, (1997).

[56] Fabio Di Benedetto, G. Fiorentino and S. Serra Capizzano, "C.G. preconditioning for Toeplitz matrices", Comput. Math. Appl. 25 : 35 – 45 (1993).

[57] M. Donatelli, S. Serra Capizzano and D. Sesana, "Multigrid methods for Toeplitz linear systems with different size reduction", preprint available from http://*arxiv.org/abs*/1010.5730*v*1, (2010).

[57$^{(1)}$] M. Donatelli, C. Estatico, A. Martinelli and S. Serra Capizzano, "Improved image deblurring with anti-reflective boundary conditions and re-blurring", Inverse Problems, 2006, 22 : 2035 – 2053.

[58] N. Dyn and D. Levin, "Subdivision schemes in geometric modelling", Acta Numerica, 11(2002), pp. 73 – 144.

[58$^{(1)}$] H.W. Engl, M. Hanke and A. Neubauer, "Regularization of Inverse Problems", Kluwer Academic Press, Dordrecht, 1996.

[58$^{(2)}$] C. Estatico, "Shift-invariant approximations of structured shift-variant blurring matrices", Dipartimento di Matematica e Informatica, Universit*à* degli Studi di Cagliari, Via Ospedale 72, 09124 Cagliari, Italy, estatico@unica.it, in progress.

[58$^{(3)}$] C. Estatico, "A classification scheme for regularizing preconditioners, with application to Toeplitz systems", Linear Algebra Appl., 397 (2005), pp. 107–131.

[58$^{(4)}$] C. Estatico, "Preconditioners for ill-conditioned Toeplitz matrices with differentiable generating functions", Numer. Linear Algebra Appl., 16 (2009), pp. 237–257.

[58$^{(5)}$] C. Estatico, E. Ngondiep, S. Serra-Capizzano and D. Sesana, "A note on the (regularizing) preconditioning of g-Toeplitz sequences via g-circulants", to appear.

[59] D. Fasino, P. Tilli "Spectral clustering properties of block multilevel Hankel matrices", Linear Agebra Appl., 306(2000), pp. 155 – 163.

[60] G. Fiorentino and S. Serra Capizzano, "Multigrid methods for Toeplitz matrices", calcolo, 28 – 34, (1991), 283 – 305.

[61] G. Fiorentino and S. Serra Capizzano, "Multigrid methods for symmetric positive definite block Toeplitz matrices with nonnegative generating functions", SIAM J. sci, Comput. 17(1996), 1068 – 1081.

[62] G. Fiorentino and S. Serra Capizzano, "Tau Preconditioners for (high order) elliptic problems, in: P. Vassilevski (Ed.), Preceedings of the 2*nd* IMACS Conference on Iterative Methods in Linear Algebra", Blagoevgrad, Bulgaria, June 1995, PP. 342 – 353.

[63] R. Fletcher, "Conjugate gradient methods for indefinite systems". In: G.A. Watson (ed.), preceedings of the Dundee Biennial Conference on Numerical Analysis 1974, pp. 73 – 89. New York: Springer-Verlag 1975.

[64] R. Fletcher and M.J.D. Powell, "A rapidly convergence descent method for minimization". Comput. J. 6, 163 – 168.

[65] R.W. Freund and N.M. Natchigal, "QMR: a quasi-minimal residual method for non-Hermitian linear systems". Numerische Mathematiko 60, 315 – 339 (1991).

[66] R.W. Freund and T. Szeto, "A transpose-free quasi-minimal residual squared algorithm for non-Hermitian linear systems", Advances in Computer Methods for Partial Differential Equations-VII (R. Vichnevetsky, D. Knight, and G. Richter, eds), pp. 258 – 264 (1992).

[67] W. Gautschi, "The condition of Vandermonde-like matrices involving orthogonal polynomials". Linear Algebra Appl. 52/53 (1983).

[68] A. George, "Nested dissection of a regular finite element mesh". SIAM J. Numer. Anal. 10, 345 – 363 (1973).

[70] P.E. Gill, G.H. Golub, W. Murray and M.A. Saunders, "Methods for modifying matrix factorizations". Math. Comput. 28, 505 – 535 (1974).

[70] I. Gohberg and I. Feldman, "Convolution Equations and Projection Methods for their Solutions", Transl. Math. Monogr. 41, Amer. Math. Soc., Providence (1974).

[71] L. Golinskii and S. Serra Capizzano, "The asymptotic properties of the spectrum of non symmetrically perturbed Jacobi matrix sequences". J. Approx. Theory, 144 – 1 (2007), pp. 84 – 102.

[72] G. Golub and C. Van Loan, "Matrix Computations". The Johns Hopkins University Press, Baltimore (1983).

[73] R.M. Gray, "On the asymptotic eigenvalue distribution of Toeplitz matrices", IEEE Transactions on Information Theory, Vol. 18, November 1972, pp. 725 – 730.

[74] R.M. Gray, "On unbounded Toeplitz matrices and Nonstationary Time series with an Application to Information Theory", Information and Control, 24, pp. 181 – 196, 1974.

[75] R.M. Gray, "Toeplitz and circulant matrices", A review, Department of Electronical Engineering Stanford 94305, USA, rmgray@stanford.edu.

[76] U. Grenander and M. Rosemblatt, "Statistical Analysis of Stationary Time series". Second edition, Chelsea, New York 1984.

[77] U. Grenander and G. Szegö, "Toeplitz forms and their Applications". Second edition, Chelsea, New York (1984).

[78] W. Hackbusch, "Multigrid Methods and Applications". Berlin, Heidelberg, New York, Springer-Verlag, (1985).

[79] W. Hackbusch and U. Trottenberg (Eds), "Multigrid methods". Lectures Notes in Mathematics. Vol. 960. Berlin, Heidelberg, New York: Springer-Verlag (1982).

[79$^{(1)}$] M. Hanke, A. Neubauer and O. Scherzer, "A convergence analysis of Landweber iteration for nonlinear ill-posed problems", Numer. Math., 1995, 72 : 21 – 37.

[79$^{(2)}$] R.J. Hanisch and R.L. White, eds., "The Restoration of HST Images and Spectra - II", Space Telescope Science Institute, Baltimore, 1994.

[79$^{(3)}$] P.C. Hansen, J.G. Nagy, D.P.O. Leary, "Deblurring images: matrices, spectra, and filtering", SIAM, Philadelphia, 2006.

[80] M.R. Hestenes and E. Stiefel, "Methods of conjugate gradients for solving linear systems". Nat. Bur. Standards, J. of Res. 49, 409 – 436, (1952).

[81] A.J. Hoffman and H.W. Wielandt, "The variation of the spectrum of a normal matrix". Duke Math. J., Vol. 20. pp. 37 – 39, (1953).

[82] A.S. Householder, "The Numerical Treatment of a single Non-linear Equation". New York, Mc Gram-Hill, (1970).

[83] T. Huckle, "Circulant and skew-circulant matrices for solving Toeplitz matrix problems", in Cooper Mountain Conference on Iterative Methods, Cooper Mountain, Colorado, (1990).

[84] D. Jackson, "The Theory of Approximation". Americam Mathematical Society, New York (1930).

[85] X.A. Jin, "Hartley preconditioners for Toeplitz systems generated by positive continuous functions". BIT 34, 367 – 371, (1994).

[86] J.H. Justice, "A Levinson-type algorithm for two-dimensional Wiener filtering using bivariate Szegö polynomials", Proc. IEEE, 65(1977), pp. 882 – 886.

[87] S. Kaniel, "Estimates for some computational techniques in linear algebra". Math. Comput. 20, 369 – 378, (1966).

[88] E.J. Kansa, "Multiquadric. A scattered data approximation scheme with applications to computational fluide dynamics: II solution to parabolic, hyperbolic, and elliptic problems", Comput. Math. Appl. 61 – 68, 32(1996).

[89] H.B. Keller, "Numerical Methods for two-point Boundary value problems". London: Blaisdell, (1968).

[90] P.P. Korovkin, "Linear Operators and Application Theory (English translation)", Hindustan Publishing Co., Delhi, (1960).

[91] M.G. Krein, "On some new Banach algebras and Wiener-Levy theorems for Fourier series and integrals". Amer. Math. Soc. Transl., 93, 177 – 199 (1970).

[92] T.K. Ku and C.J. Kuo, "Design and analysis of preconditioners", Tech. Report 155, Signal and Image Processing Institute, University of Southern California, May 1990; IEEE Trans. Signal Processing, 40(1992), pp. 129 – 141.

[93] A.B.J. Kuijlaars and S. Serra Capizzano, "Asymptotic zero distribution of orthogonal polynomials with discontinuously varying recurrence coefficients", J. Approx. Theory, 113(2001), pp. 142 – 155.

[94] C. Lanczos, "Solution of systems of linear equations by minimized iterations". J. Res. Nat. Bur. Standards 49, pp. 33 – 53 (1975).

[95] C. Lanczos, "An iterative method for the solution of the eigenvalue problem of linear differential and linear integrable equations". J. Res. Nat. Bur. Standards 45, pp. 255 – 282 (1950).

[96] C.L. Lawson and H.J. Hanson, "Solving least squares problems". Englewood Cliffs, N.J.: Prentice-Hall, (1974).

[97] B.C. Levy, M.B. Adams and A.S. Willsky, "Solution and linear estimation of 2-D nearest-neighbor models". Proc. IEEE, 78(1990), pp. 627 – 641.

[98] T.L. Marzetta, "Two-dimensional linear prediction: Autocorrelation arrays, minimum-phase prediction error filters, and reflection coefficients arrays", IEEE Trans. Acoust., Speech, Signal Process., 28(1980), pp. 725 – 733.

[98$^{(1)}$] S.R. McNown and B.R. Hunt, "Approximate shift-invariance by warping shift-variant systems The Restoration of HST Images and Spectra II", Space Telescope Science Institute, Baltimore, 1994, 181 – 187.

[99] J.A. Meijerink and H.A. Van der Vorst, "An iterative solution method for linear systems of which the coefficient matrix is a symmetric M-matrix". Math. Comput. 31, pp. 148 – 162, (1977).

[100] E.H. Moore, "General Analysis". Part I, Amer. Phill. Society, Philadelphia, (1935).

[101] V.A. Muller and M. Meymann, "A note on comparison theorems for nonnegative matrices", Numer. Math. 47(1985), 427 – 434.

[101^{1}] J.G. Nagy and D. P.O. Leary, "Restoring Images Degraded by Spatially-Variant Blur", SIAM J. Sci. Comp., 1997, 19 : 1063 – 1082.

[102] I.P. Natanson, "Constructive Function Theory", I. Frederick Ungar Publishing Co., New York, (1964).

[103] E. Ngondiep, S. Serra Capizzano and D. Sesana, "Spectral features and asymptotic properties of α-circulant and α-Toeplitz sequences: Theoretical results and examples", preprint available from http://arxiv.org/abs/0906.2104, (2009).

[104] E. Ngondiep, S. Serra Capizzano and D. Sesana, "Spectral features and asymptotic properties of g-circulant and g-Toeplitz sequences", SIAM J. Matrix Anal. Appl., 31–4 (2010), pp. 1663 – 1687.

[105] D. Noutsos, "On the Perron-Frobenius property of matrices having some negative entries", Lin. Algebra Appl., 412(2006), pp. 132 – 153.

[106] C.C. Paige, "The Computation of eigenvalues and eigenvectors of very large sparse matrices". PhD Thesis, London University, (1971).

[107] C.C. Paige and M.A. Saunders, "Solution of sparse indefinite systems of linear equations". SIAM J. Numer. Analysis 12, pp. 617 – 624, (1975).

[108] B.N. Parlett and D.S. Scott, "The Lanczos algorithm with selective orthogonalization". Math. Comput. 33, 217 – 238, (1979).

[109] S.V. Parter, "On the distribution of singular values of Toeplitz matrices", Lin. Algebra Appl., 80 : 115 – 130, (1986).

[110] S.V. Parter, "Extreme eigenvalues of Toeplitz forms and applications to elliptic difference equations", Trans. Amer. Math. Soc. 99 : 153 – 162, (1996).

[111] J.K. Reid (Ed.), "On the method of conjugate gradients for the solution of large sparse systems of linear equations", 231 – 252, (1971).

[111$^{(1)}$] A. Rieder, "On the regularization of nonlinear ill-posed problems via inexact Newton iterations", Inverse Problems, 1999, 15 : 309 – 327.

[111$^{(2)}$] G.M. Robbins, "Image restoration for a class of linear spatially-variant degradations", Pattern Recognition, 1970, 2 : 91 – 103.

[112] R. Penrose, "A generalized inverse for matrices", Proc. Cambridge Phil. Soc., 51(1955), pp. 406 – 413.

[113] J. Rissanen, "Algorithm for triangular decomposition of block Hankel and Toeplitz matrices with application to factoring positive matrix polynomials", Math. Comp., 27(1973), pp. 147 – 154.

[114] Y. Saad, "Iterative methods for sparse linear systems". Boston: PWS Pushing Company (1996).

[115] Y. Saad, "On the rate of convergence of the Lanczos and block Lanczos methods". SIAM J. Num. Anal. 17, 687 – 706, (1980).

[115$^{(1)}$**]** Saad, Y.: SPARSKIT, "A basic tool kit for sparse matrix computations". Technical Report 90.20, RIACS, NASA Ames Research Center, May 1990.

[116] Y. Saad and M.H. Schultz, "GMRES: a generalized minimal residual algorithm for solving nonsymmetric linear systems". SIAM J. Scientific and Statistical Computing, 7, 856 – 869, (1986).

[116$^{(1)}$] A.A. Sawchuk, "Space variant system analysis of image motion", J. Opt. Soc. Am., 1973, 63 : 1052 – 1063.

[117] S. Serra Capizzano, "Preconditioning strategies for asymptotically ill-conditioned block Toeplitz systems", BIT 34(1994), 579 – 594.

[118] S. Serra Capizzano, "Preconditioning strategies for Hermitian Toeplitz systems with nondefinite generating functions". SIAM J. Matrix Anal. Appl. 17 – 4, 1007 – 1019, 1996.

[119] S. Serra Capizzano, "Optimal, quasi-optimal and superlinear band Toeplitz preconditioners for asymptotically ill-conditioned positive definite Toeplitz systems", Math. Comput. 66(1997), 651 – 665.

[120] S. Serra Capizzano, "Sulle proprietà spettrali di matrici preconditionate di Toeplitz". Boll. Un. Mat. Ital. 11 – 7, 651 – 665, (1997).

[121] S. Serra Capizzano, "The extension of concept of generating function to a class of preconditioned Toeplitz matrices", Linear. Algebra Appl., 267, 139 – 161, (1997).

[122] S. Serra Capizzano, "Superlinear PCG methods for symmetric Toeplitz systems". Math. Comput., in press

[123] S. Serra Capizzano, "A Korovkin-type Theory for finite Toeplitz operators via matrix algebra", Numer. Math. 82 – 1, 117 – 142, (1999).

[124] 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, pp. 1831 – 1857, (1999).

[125] S. Serra Capizzano, "Distribution results on the algebra generated by Toeplitz sequences: finite dimensional approach", Linear Algebra Appl., 328(2001), pp. 121 – 130.

[126] S. Serra Capizzano, "Convergence analysis of two-grid methods for elliptic Toeplitz and PDEs matrix sequences", Numer. Math. 92(2002), pp. 433 – 465.

[127] S. Serra Capizzano, "Test functions, growth conditions and Toeplitz matrices", Rend. Circolo Mat. Palermo, II-68 (2002), pp. 791 – 795.

[128] S. Serra Capizzano, "Generalized Locally Toeplitz sequences: spectral analysis and applications to discretized Partial Differential Equations", Linear Algebra Appl., 366(2003), pp. 371 – 402.

[129] S. Serra Capizzano, "A note on antireflective boundary conditions and fast deblurring models", SIAM J. Sci. Comput., 25 – 4, (2003), pp. 1307 – 1325.

[130] S. Serra Capizzano, "The GLT class as a Generalized Fourier Analysis and applications", Linear Algebra Appl., 419(2006), pp. 180 – 233.

[131] S. Serra Capizzano, "The spectral approximation of multiplication operators via asymptotic (structured) linear algebra", Linear Algebra Appl., 424(2007), pp. 154 – 176.

[132] S. Serra Capizzano, "The rate of convergence of Toeplitz based on PCG methods for second order nonlinear boundary value problems", Numer. Math. 18(3), (1999) 461 – 495.

[133] S. Serra Capizzano, "How to choose the best iterative strategy for symmetric Toeplitz systems?" SIAM J. Numer. Anal., in press.

[134] S. Serra Capizzano, "Korovkin Theorems and linear positive Gram matrix algebras approximation of Toeplitz matrices". Linear Algebra Appl., 284, (1998), pp. 307 – 334.

[135] S. Serra Capizzano, "Some theorems on linear positive operators and functionals and their applications". TR nr.26, Dept. of Mathematics (LAN)-Univ. of Calabria. (1997).

[136] S. Serra Capizzano, "Spectral behavior of matrix sequences and discretized boundary value problems", Linear Algebra Appl. Vol. 337, PP. 37 – 78, 2001.

[137] S. Serra Capizzano, "Conditioning and solution of Hermitian (block) Toeplitz systems by means of preconditioned conjugate gradient methods", in Proceedings in Advanced Signal Processing Algorithms, Architectures, and Implementations-SPIE conference (F. Luk, Ed.), San Diego, July 1995, pp. 326 – 337.

[138] S. Serra Capizzano, "On the extreme eigenvalues of Hermitian (block) Toeplitz matrices, Linear Algebra Appl.", 270(1998), 109 – 129.

[139] S. Serra Capizzano, "On the extreme spectral properties of Toeplitz matries generated by L^1 functions with several minima (maxima)", BIT, 36 : 135 – 142, (1996).

[140] S. Serra Capizzano, "New PCG based methods for Hermitian Toeplitz systems", Calcolo, in press.

[141] S. Serra Capizzano and C. Tablino Possio, "Multigrid methods for multilevel circulant matrices", SIAM J. Sci. Comput., 26 – 1 (2004), pp. 55 – 85.

[142] S. Serra Capizzano and C. Tablino Possio, "High precision finite difference schemes and Toeplitz based preconditioners for elliptic problems", Electr. Trans. Numer. Anal. 11(2000), 55 – 84.

[143] S. Serra Capizzano and E. Tyrtyshnikov, "Any circulant-like preconditioner for multilevel matrices is not optimal". TR nr.28, Dept. of Mathematics (LAN)-Univ. of Calabria. 1997.

[144] S. Serra Capizzano and E. Tyrtyshnikov, "Any preconditioner belonging to partially equimodular spaces for multilevel matrices is not optimal". TR nr.30, Dept. of Mathematics (LAN)-Univ. of Calabria. 1998.

[145] S. Serra Capizzano and E. Tyrtyshnikov, "Multilevel Toeplitz matrices and approximation by matrix algebras in: F. Luk (Ed.), preceedings in Advanced Signal Processing Algorithms, Architectures, and Implementations", $VIII-$ SPIE Conference, San Diego, CA, July 1998, PP.393 – 404.

[146] B. Silbermann and O. Zabroda, "Asymptotic behavior of generalized convolutions: an algebraic approach", J. Integral Equ. Appl., 18 – 2 (2006), PP. 169 – 196.

[147] P. Sonneveld: CGS, "A fast Lanczos-type solver for nonsymmetric linear systems". SIAM J. Scientific and Statistical Computing, 10 PP. 36 – 52, (1989).

[148] J. Stoer, "On the convergence rate of imperfect minimization algorithms in Broyden's β-class". Math. Programming 9, PP. 313 – 335, (1975).

[149] J. Stoer and R. Burlish, "Introduction to Numerical Analysis", Vol. I, Springer-Verlag, Berlin, (1970).

[150] G. Strang, "Wavelets and dilation equations: a brief introduction". SIAM Rev., 31(1989), PP. 614 – 627.

[150$^{(1)}$] G. Strang, "A proposal for Toeplitz matrix calculations", Stud. Appl. Math., 74 (1986), pp. 171–176.

[151] V. Strela and E. Tyrtyshnikov, "Which circulant preconditioner is better?" Math. Comput., 65, PP. 137 – 150, (1996).

[152] P. Tarazaga, M Raydan and A. Hurman, "Perron-Frobenius theorem for matrices with some negative entries", Linear Algebra Appl. 328(2001), 57 – 68.

[153] P. Tilli, "Singular values and eigenvalues of non Hermitian block Toeplitz matrices", Linear Algebra Appl., 272, PP. 59 – 89, (1998).

[154] P. Tilli, "On the asymptotic distribution of the eigenvalues of Hermitian Toeplitz matrices with Toeplitz blocks", Math. Comput., in press.

[155] P. Tilli, "Locally Toeplitz matrices: spectral theory and applications", Linear Algebra Appl., 278(1998), PP. 91 – 120.

[156] P. Tilli, "A note on the spectral distribution of Toeplitz matrices", Linear Multilin. Algebra, 45(1998) PP. 147 – 159.

[157] P. Tilli, "Some results on complex Toeplitz eigenvalues", J. Math. Anal. Appl., 239 – 2 (1999) PP. 390 – 401.

[158] L.N. Trefethen, "Approximation theory and numerical linear algebra, in algorithms for Approximation II", M. Cox and J.C. Mason, eds, Chapman and Hall, London, (1988).

[159] W.F. Trench, "Asymptotic distribution of the even and odd spectra of real symmetric Toeplitz matrices", Linear Algebra, Appl., Vol., 302 – 303, PP. 155 – 162, 1999.

[160] W.F. Trench, "Simplification and strenghening of Weyl′s definition of equal distribution of two families of finite sets", cubo, A Mathematical Journal, Vol., 06, N3 (2004), PP. 47 – 54.

[161] W.F. Trench, "Properties of multilevel block α-circulants", Linear Algebra Appl., 431 (2009), PP. 1833 – 1847. 1999.

[$161^{(1)}$] W. F. Trench, "Properties of unilevel block circulants", Linear Algebra Appl., 430 (2009), pp. 2012 – 2015.

[162] U. Trottenberg, C.W. Oosterlee and A. Schüller, "Multigrid", Academic Press, London, 2001.

[163] E. Tyrtyshnikov, "Optimal and superoptimal circulant preconditioners", SIAM J. Matrix Anal. Appl., 13 (1992), pp. 459 – 473.

[164] E. Tyrtyshnikov, "A unifying approach to some old and new theorems on distribution and clustering", Linear Algebra Appl. 232(1996), 1 – 43.

[165] E. Tyrtyshnikov, "Circulant preconditioners with unbounded inverses". Linear Algebra Appl. 216 (1995), 1 – 23.

[166] E. Tyrtyshnikov and N. Zamarashkin, "Thin structure of eigenvalue clusters for non-Hermitian matrices", Linear Algebra Appl. 292(1999), 297 – 310.

[167] E. Tyrtyshnikov and N. Zamarashkin, "Spectra of multilevel Toeplitz matrices", Advanced theory via simple matrix relationships, private communication, (1996).

[168] Van der Vorst, "Bi-CGSTAB: A fast and smoothly converging variant of Bi-CG for the solution of non-symmetric linear systems". SIAM J. Scientific and Statistical Computing 12, 631 – 644, (1992)

[169] G.A. Watson, "An algorithm for the inverse of block matrices of Toeplitz forms", J. Comput. Mach., 20(1973), pp. 409 – 415.

[170] H. Widom, "Toeplitz matrices", in Studies in Real and Complex Analysis (I. Hirshman, Jr., Ed.), Math. Assoc. Amer., 1995.

[171] H. Widom, "Szegö limit theorem: The higher dimensional matrix case", J. Funct. Anal. 39 : 182 – 198, (1980).

[172] H. Widom, "On the singular values of the Toeplitz matrices", Z. Anal. Anwendungen 8 : 221 – 229 (1989).

[173] J.H. Wilkinson, "The algebraic Eigenvalue Problem". Claredon Press, Oxford (1965).

[174] J.H. Wilkinson and C. Reinsch, "Linear algebra. Handbook for automatic Computation, Vol. II Grundlehren der mathematischen Wissenschaften in Einzeldarstellungen", Bd. 186. Berlin, Heidelberg, New York: Springer-Verlag (1971).

[175] Z. Woźnicki, "Nonnegative Splittings Theory". Japan Journal of Industrial and Applied Mathematics 11(1998), 289 – 342.

[176] A. Zygmund, "Trigonometric Series", Cambridge U.P., Cambridge, 1959. 11(1998), 289 – 342.