Toeplitz matrices: spectral properties and preconditioning in the CG method

Stefano Serra-Capizzano *

November 20, 2007

Abstract

We consider multilevel Toeplitz matrices $T_n(f)$ generated by Lebesgue integrable functions f defined over I^d, $I = [-\pi, \pi)$, $d \geq 1$. We are interested in the solution of linear systems with coefficient matrix $T_n(f)$ when the size of $T_n(f)$ is large. Therefore the use of iterative methods is recommended for computational and numerical stability reasons. In this note we focus our attention on the (preconditioned) conjugate gradient (P)CG method and on the case where the symbol f is known and univariate ($d = 1$): the second section treat spectral properties of Toeplitz matrices $T_n(f)$; the third deals with the spectral behavior of $T_n^{-1}(g)T_n(f)$ and the fourth with the band Toeplitz preconditioning; in the fifth section we consider the matrix algebra preconditioning through the Korovkin theory. Then in the sixth section we study the multilevel case $d > 1$ by emphasizing the results that have a plain generalization (those in the Sections 2, 3, and 4) and the results which strongly depend on the number d of levels (those in Section 5): in particular the quality of the matrix algebra preconditioners (circulants, trigonometric algebras, Hartley etc.) deteriorates sensibly as d increases.

A section of conclusive remarks and two appendices treating the theory of the (P)CG method and spectral distributional results of structured matrix sequences.

key words Linear system, conjugate gradient method, Toeplitz matrix, structured matrix, preconditioner.

Contents

*Dipartimento di Chimica, Fisica e Matematica, Università dell'Insubria, Via Valleggio 11, 22100 Como (ITALY). Email: stefano.serrac@uninsubria.it; serra@mail.dm.unipi.it

1 Introduction

Toeplitz matrices and operators arise in a wide variety of fields of pure and applied mathematics such as probability theory, harmonic analysis, statistics, Markov chains, signal theory, image processing etc. A matrix is called Toeplitz matrix (of finite, infinite or bi-infinite order) if its (i, j) entry depends only on the difference $i - j$ of the subscripts. In this note we are interested in finite dimensional Toeplitz problems in which the $n \times n$ Toeplitz matrix $T_n(f)$ is an n–section of an infinite Toeplitz operator whose entries on the k–th diagonal are the Fourier coefficients a_k of an assigned Lebesgue integrable function f defined on the fundamental interval $I = [-\pi, \pi)$ and periodically extended to the whole real axis:

$$a_k = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x) e^{-\mathbf{i}kx} dx, \quad \mathbf{i}^2 = -1.$$

Indeed, if f is a real valued function we have $a_k = \bar{a}_{-k}$ and, consequently, $T_n(f)$ is Hermitian; moreover, if $f(x) = f(-x)$, then the coefficients a_k are real and $T_n(f)$ is symmetric. We emphasize (see [23]) that the generating function f is given in many applications (such as finite difference discretization of partial differential equations (PDEs), some linear estimation problems etc.) but it is unknown in many others (such as some inverse problems arising in signal/image restoration applications etc.). In the following we suppose to know some information about the analytic properties of f: the case where f is unknown in briefly sketched at the end of Section 4 where some relevant references are also provided. Of particular interest in the applications are the solution of Toeplitz linear systems and the analysis of the extremal behavior of the spectra of such matrices. By using a combination of elementary techniques of linear algebra and calculus, we discuss linear algebra tools to obtain estimates regarding the smallest and the greatest eigenvalue of a Toeplitz matrix. Moreover we stress that the knowledge of the asymptotic behavior of the spectra of the family of matrices $\{T_n(f)\}$ is crucial in order to understand how ill–conditioned are these matrices (see Section 2 and [14, 66]): this information is, obviously, useful to design a suitable solver for the related Toeplitz linear system. Actually, since we deal with very large problems, it is imprudent to use direct methods [17]; consequently, for this reason and for computational convenience both in sequential and in parallel model of computation, a lot of attention has been paid to the application of iterative methods, such as preconditioned conjugate gradient (PCG) and, more recently, multigrid methods [33, 35]. In the following, we analyze in detail the application of a class of PCG methods whose preconditioners $T_n(g)$ are positive definite Toeplitz matrices generated by essentially nonnegative functions g. In view of this, we deeply analyze the properties of the spectra of the preconditioned matrices $T_n^{-1}(g)T_n(f)$ in terms of the dimension n and of the analytic behavior of the functions f and g (see Section 4 and [18, 30, 61, 64, 62]). In this way we arrive at the design of efficient algorithms for the solution of linear systems of the form $T_n(f)\mathbf{x} = \mathbf{b}$. In the second part we consider preconditioners of $O(n\log(n))$ arithmetic cost coming from matrix algebras (circulants, trigonometric matrix algebras, Hartley matrices etc.) for which we introduce the Frobenius optimal approach introduced by Tony Chan [24]: their analysis is carried out in a unified way by means of the Korovkin theory of which we present the classical version [49] in Approximation Theory and our matrix version [71]. Finally we consider the more challenging multilevel case which often appear in multivariate problems (imaging, PDEs etc.): we focus our attention on the band approach and on the matrix algebra approach by emphasizing both the generalizable parts of the theory and the intrinsic difficulties of the multilevel setting.

The organization of these notes follows the scheme given in the index. In particular, in Section 2 we analyze the localization, the distribution and the extremal properties of the spectra of Hermitian Toeplitz matrix sequences. In Section 3 we make the same analysis about the spectra of the preconditioned matrices. These results are crucial to design efficient preconditioned iterative methods and, in fact, in Section 4 we perform a detailed analysis of the various cases ($f \geq 0$, $f \leq 0$ in Subsections 4.2 and 4.3, f with nondefinite sign in Subsection 4.4, f with zeros of odd orders in Subsection 4.5, f with zeros of generic orders in Subsection 4.6, and more involved cases in Subsection 4.7), obtaining suitable band Toeplitz preconditioners and good algorithms to solve a given linear system $T_n(f)\mathbf{x} = \mathbf{b}$. Section 5 is devoted to the matrix algebra preconditioning which is introduced through the Korovkin theory. Section 6 deals with the multilevel case by emphasizing the results that have a plain generalization (those in the Sections 2, 3, and 4) and the results which strongly depend on the number d of levels (those in Section 5). A section of conclusions 7 and two appendices on convergence theory of (P)CG methods and advanced spectral results conclude these notes.

2 Spectral features of $T_n(f)$

The main purpose of this section is the study of the eigenvalues and especially of the extreme eigenvalues of $T_n(f)$, i.e., $\lambda_1^{(n)}$ and $\lambda_n^{(n)}$: we write $\lambda_j^{(n)}(T_n(f))$ in place of $\lambda_j^{(n)}$ when the matrix is not clear from the context and in any case we always consider a nondecreasing order i.e. $\lambda_1^{(n)} \leq \cdots \leq \lambda_n^{(n)}$. Starting from the pioneering work by Szegö (see e.g. [37]), we know that many spectral properties of $T_n(f)$ are well understood by considering its generating function f. In the following we denote by essinf f and esssup f the essential infimum and the essential supremum of f, i.e., $\inf f$ and $\sup f$ up to a zero–measure set; we denote by C_0 the set of the continuous functions with bounded support defined on the real line, by $m\{\cdot\}$ the usual Lebesgue measure, and by $M_N(\mathbf{C})$ the set of all N by N matrices with complex entries. The following result due to Szegö in $L^\infty(I)$ and generalized by Tyrtyshnikov and Zamarashkin in $L^1(I)$ holds true.

Theorem 2.1 *Let $\lambda_j^{(n)}$, $j = 1, \ldots, n$, be the eigenvalues of $T_n(f)$ with f real valued and belonging to $L^1(I)$. Then for every $F \in C_0$ the following limit relation holds:*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{j=1}^{n} F\left(\lambda_j^{(n)}\right) = \frac{1}{2\pi} \int_I F\left(f(x)\right) dx. \tag{1}$$

The most suggestive interpretation of relation (1) is as follows: a suitable ordering of the eigenvalues $\lambda_j^{(n)}$, $j = 1, \ldots, n$, can be seen as an approximation of the function $f(\cdot)$ sampled on an equispaced grid on the domain $I = [-\pi, \pi)$. Therefore it is quite evident that the symbol f decides the asymptotic inertia (asymptotic definiteness), the asymptotic size of the ill-conditioned spaces (and its nature) of the associated matrices $T_n(f)$. These features are discussed in more detail in Subsection 2.1

2.0.1 Extensions and generalizations of Theorem 2.1

Some possible extensions of the class of the test functions have been considered. For instance in [67], the test functions have not necessarily bounded support since they are obtained as intersection of the continuous functions with the L^∞ closure of $\mathcal{J}$ where $\mathcal{J}$ is the set of the linear combinations of characteristic functions of unbounded intervals: the resulting space is given by the continuous functions having finite limits at ∞ and $-\infty$ (see item **d** of Theorem 3.1). Furthermore, in [91] relation (1) is proved for any $F \in UCB$ (i.e. uniformly continuous and bounded on $\mathbf{R}$). We observe that this class strictly contains the class of test functions $C_{\text{blimits}} = \overline{\mathcal{J}}^{L^\infty}$ given in [67].
A bigger class (in some sense the biggest class) is considered in [76] for which the considered limit relation holds for any function $f \in L^p(I)$: the considered class is $C(p)$ which defined as the set of all continuous functions F on the real line such that $\frac{F(z)}{1+|z|^p}$ belongs to $L^\infty(\mathbf{R})$. The meaning is that the test functions should satisfy a growth condition at infinity which is vacuous when the symbol f is in $L^\infty(I)$ and is stronger when $f \in L^1(I)$ (in this case F can have at most a linear growth). The crucial facts used for proving such results are contained in the following inequalities obtained in [83] and valid also for complex valued (and multivariate) symbols: for any $f \in L^p(I)$ with $p \in [1, \infty)$, we have

$$\|T_n(f)\|_p^p \leq (2\pi)^{-1}\, n \|f\|_{L^p}^p \;\text{ and }\; \|T_n(f)\| \leq \|f\|_\infty \;\; (\text{if } f \in L^\infty(I)), \tag{2}$$

where $\|X\|_p = \left[\sum_{j=1}^{N} \sigma_j^p\right]^{1/p}$ denotes the Schatten p norm of the matrix $X \in M_N(\mathbf{C})$ with singular values $\{\sigma_j\}$ (see e.g. [7]) and $\|X\| = \max_j \sigma_j$. We use inequalities (2) for giving a simple matrix

theoretic proof of Theorem 2.1 in Appendix **B.1**.

In a completely different direction we can consider the case of complex valued symbols: in this case the finite sections $T_n(f)$ are not necessarily Hermitian and the natural extension of relation (1) holds with the eigenvalues replaced by the singular values and with the symbol f replaced by $|f|$ in the righthand side (see Theorem 9.1 and Definition 9.1 in Appendix **B**).

Very deep results are derived by Tilli [89] for the eigenvalues in the non Hermitian case: in this case the (1) is satisfied for every f if we restrict the test functions F to the set of the analytic functions on a certain annulus of the complex plane. This set of test functions is very poor (the analyticity is a restrictive condition) and gives very weak information, consequently, if we want as test functions the class C_0 on $\mathbf{C}$, then we have to focus our attention on symbols f such that the essential range does not disconnect the complex plane (see the beautiful paper [92]): an interesting and somehow strange consequence of this statement is that we cannot have good distribution results for most of the regular symbols such as the polynomials.

Finally mention has to be made to notion of approximating class of sequences [74]: it represents a basic tool for developing an approximation theory for matrix sequences and it has been successfully used for obtaining global distribution results for more involved matrix sequences including multilevel Toeplitz structures, discretization of partial differential equations with variable coefficients and over general domains, the algebra generated by (multilevel) Toeplitz matrices with $L^\infty(I^d)$ symbols etc. A sketch of the use of this tool and of the related results is given in Appendix **B**.

2.1 Localization and extreme spectral results

Here is a sample of some spectral results that are described through the symbol f: many of these can be proven by using Theorem 2.1.

Theorem 2.2 *Let $\lambda_j^{(n)}$ be the eigenvalues of $T_n(f)$ sorted in nondecreasing order, and m_f =essinf f, M_f =esssup f.*

a. *If $m_f < M_f$ then $\lambda_j^{(n)} \in (m_f, M_f)$ for every j and n; if $m_f = M_f$ then f is constant and trivially $T_n(f) = m_f I_n$ with I_n identity of size n;*

b. *the following asymptotic relationships hold:* $\lim_{n\to\infty} \lambda_1^{(n)} = m_f$, $\lim_{n\to\infty} \lambda_n^{(n)} = M_f$.

Proof. Concerning the first item, we observe that $m_f = M_f$ implies that f is constant almost everywhere (a.e.) and hence $a_0 = m_f$ and $a_k = 0$ for all $k \neq 0$. Therefore $T_n(f) = m_f I_n$ and any eigenvalue of $T_n(f)$ coincides with m_f.
When $m_f < M_f$ the proof is given as follows. Any eigenvalue $\lambda_j^{(n)}$ can be viewed as a special Rayleigh quotient $\mathbf{u}^H T_n(f)\mathbf{u}$ with unitary vector $\mathbf{u}$ (more precisely $\mathbf{u}$ is a unitary eigenvector related to $\lambda_j^{(n)}$). Therefore the thesis claimed in the first item is proven if we prove that $\mathbf{u}^H T_n(f)\mathbf{u} \in (m_f, M_f)$ for every unitary vector $\mathbf{u}$ i.e., by linearity of $T_n(\cdot)$, if we prove that $\mathbf{u}^H T_n(f - m_f)\mathbf{u} > 0$ and $\mathbf{u}^H T_n(M_f - f)\mathbf{u} > 0$ for every unitary vector $\mathbf{u}$. We consider the first inequality the second one being completely similar:

$$\mathbf{u}^H T_n(f - m_f)\mathbf{u} \;=\; \sum_{j,k=1}^{n} \bar{u}_j \left(T_n(f - m_f)\right)_{j,k} u_k$$

$$
\begin{aligned}
&= \sum_{j,k=1}^{n} \bar{u}_j u_k \frac{1}{2\pi} \int_I (f(x) - m_f) e^{-\mathbf{i}(j-k)x} dx \\
&= \frac{1}{2\pi} \int_I (f(x) - m_f) \sum_{j,k=1}^{n} \bar{u}_j e^{-\mathbf{i}jx} u_k e^{\mathbf{i}kx} dx \\
&= \frac{1}{2\pi} \int_I (f(x) - m_f) \sum_{j,k=1}^{n} \bar{u}_j e^{-\mathbf{i}(j-1)x} u_k e^{\mathbf{i}(k-1)x} dx \\
&= \frac{1}{2\pi} \int_I (f(x) - m_f) \left| \sum_{k=1}^{n} u_k e^{\mathbf{i}(k-1)x} \right|^2 .
\end{aligned}
$$

Therefore $\mathbf{u}^H T_n(f - m_f)\mathbf{u}$ is nonnegative since $f(x) - m_f$ is nonnegative a.e. and $\left|\sum_{k=1}^{n} u_k e^{\mathbf{i}(k-1)x}\right|^2$ is nonnegative being the square of a polynomial. To prove the strict inequality we observe that the set A^+ where $f(x) - m_f$ is strictly positive has positive measure since $M_f > m_f$ and therefore the integral over A^+ of $(f(x) - m_f)\left|\sum_{k=1}^{n} u_k e^{\mathbf{i}(k-1)x}\right|^2$ must be positive by the fundamental theorem of algebra since the complex polynomial $\sum_{k=1}^{n} u_k z^{k-1}$ (which is not identically zero due to the normalization condition $\mathbf{u}^H\mathbf{u} = \sum_{k=1}^{n} |u_k|^2 = 1$) can have at most $n-1$ zeros in A^+.
For item **b**, we observe that for every positive n the matrix $T_n(f)$ is a principal submatrix of $T_{n+1}(f)$. Therefore, by the Cauchy interlace Theorem [7, 55], we obtain that $\lambda_1^{(n)} > m_f$ is a non increasing sequence and $\lambda_n^{(n)} < M_f$ is a nondecreasing sequence. As a consequence both the sequences have limits and, more precisely, we deduce

$$
\lim_{n\to\infty} \lambda_1^{(n)} = m \geq m_f, \quad \lim_{n\to\infty} \lambda_1^{(n)} = M \leq M_f.
$$

By contradiction we assume $m > m_f$ (or $M < M_f$). Then we can construct a continuous nonnegative function $F \in C_0$ such that $F(z) = 0$ for $z \geq m$ or $z \leq m_f - 1$, with $F(m_f) = 1$ and being linear in $[m_f - 1, m_f]$ and $[m_f, m]$ (or $F(z) = 0$ for $z \leq M$ or $z \leq M_f + 1$, with $F(M_f) = 1$ and being linear in $[M_f, M_f + 1]$ and $[M, M_f]$): in such a way we have

$$
\begin{aligned}
\frac{1}{2\pi} \int_I F\left(f(x)\right) dx &\geq \frac{1}{2\pi} \int_{\{x\in I:\ f(x)\in[m_f,(m+m_f)/2]\}} F\left(f(x)\right) dx \\
&\geq \frac{1}{2\pi} \int_{\{x\in I:\ f(x)\in[m_f,(m+m_f)/2]\}} \frac{1}{2} dx \\
&= \frac{1}{4\pi} m\{x \in I :\ f(x) \in [m_f, (m + m_f)/2]\} > 0
\end{aligned}
$$

while

$$
\frac{1}{n} \sum_{j=1}^{n} F\left(\lambda_j^{(n)}\right) = 0, \ \ \forall n \geq 1
$$

since no eigenvalue of $T_n(f)$ lies below m. Obviously the latter two relations cannot be simultaneously true due to (1). •

It is worthwhile observing that a simple variation on the theme of the proof of the second item allows to obtain the following quite strong results:

$$
\lim_{n\to\infty} \lambda_{i(n)}^{(n)} = m_f, \ \ \lim_{n\to\infty} \lambda_{n+1-i(n)}^{(n)} = M_f, \quad \forall\, i(n) \geq 1 \text{ with } \ i(n) = o(n). \tag{3}
$$

Further interesting results can be obtained by manipulating relation (1).

Theorem 2.3 *If* $m\{x \in I :\ f(x) = a\} = m\{x \in I :\ f(x) = b\} = 0$ *then*

$$\lim_{n\to\infty} \frac{\#\{j : \lambda_j^{(n)} \in [a,b]\}}{n} = \frac{m\{x \in I :\ f(x) \in [a,b]\}}{2\pi}.$$

Proof. We observe that the wanted relation is the same as (1) where the test function F is the characteristic function of the set $[a,b]$ i.e. $F(z) = 1$ if $z \in [a,b]$ and 0 elsewhere: the problem is that this function is not continuous as required by the hypotheses of Theorem 2.1. However, by using equalities $m\{x \in I :\ f(x) = a\} = m\{x \in I :\ f(x) = b\} = 0$, it is possible to prove that for every $\epsilon > 0$ the number of the eigenvalues belonging to a ϵ neighborhood of a and b are bounded by $n\theta(\epsilon)$ with $\theta(\epsilon)$ infinitesimal as ϵ becomes infinitesimal. The latter is used with relation (1) with the globally continuous test functions F_+^ϵ and F_-^ϵ. Here $F_+^\epsilon(z) = F(z)$ for $z \in [a,b] \cup (\mathbf{R}\backslash[a-\epsilon, b+\epsilon])$, is linear on $[a-\epsilon, a]$ and on $[b, b+\epsilon]$; analogously $F_+^\epsilon(z) = F(z)$ for $z \in [a+\epsilon, b-\epsilon] \cup (\mathbf{R}\backslash[a,b])$, is linear on $[a, a+\epsilon]$ and on $[b-\epsilon, b]$. It is clear that

$$F_-^\epsilon(z) \le F(z) \le F_+^\epsilon(z) \tag{4}$$

and, by relation (1), we have

$$\lim_{n\to\infty} \frac{1}{n} \sum_{j=1}^{n} F_\pm^\epsilon\left(\lambda_j^{(n)}\right) = \frac{1}{2\pi} \int_I F_\pm^\epsilon\left(f(x)\right) dx.$$

Therefore, by (4), for every $\epsilon > 0$ we deduce

$$\begin{aligned}
\frac{1}{2\pi} \int_I F_+^\epsilon\left(f(x)\right) dx &\ge \limsup_{n\to\infty} \frac{1}{n} \sum_{j=1}^{n} F\left(\lambda_j^{(n)}\right) \\
&\ge \liminf_{n\to\infty} \frac{1}{n} \sum_{j=1}^{n} F\left(\lambda_j^{(n)}\right) \\
&\ge \frac{1}{2\pi} \int_I F_-^\epsilon\left(f(x)\right) dx.
\end{aligned}$$

Because F is the L^1 limit of both F_-^ϵ and F_+^ϵ we have

$$\lim_{\epsilon\to 0} \frac{1}{2\pi} \int_I F_-^\epsilon\left(f(x)\right) dx = \lim_{\epsilon\to 0} \frac{1}{2\pi} \int_I F_+^\epsilon\left(f(x)\right) dx = \frac{1}{2\pi} \int_I F\left(f(x)\right) dx,$$

and therefore the desired result follows since both the $\liminf_{n\to\infty}$ and the $\limsup_{n\to\infty}$ of $\frac{1}{n} \sum_{j=1}^{n} F\left(\lambda_j^{(n)}\right)$ coincide with $\frac{1}{2\pi} \int_I F\left(f(x)\right) dx$. •

Let $\mathcal{ER}(f)$ be the essential range of f defined by $y \in \mathcal{ER}(f)$ if and only if $\forall \epsilon > 0$ we have $m\{x :\ f(x) \in (y-\epsilon, y+\epsilon)\} > 0$). Let $B(A,\delta)$ be the closed δ-fattening of A with positive δ and $A \subset \mathbf{R}$, i.e., $B(A,\delta) = \bigcup_{x\in A} \{y : |y-x| \le \delta\}$. From Theorem 2.1 it follows

Proposition 2.1 *Let $\delta >$ and $X = (m_f, M_f)/B(\mathcal{ER}(f), \delta)$, then $\bigcup_{n=1}^{\infty} \bigcup_{j \le n} \lambda_j^{(n)}$ in dense in $\mathcal{ER}(f)$ while only "few" eigenvalues of $T_n(f)$ belong to X. Actually, for every positive δ independent of n, $\#\{j : \lambda_j^{(n)} \in X\} = o(n)$ and $n - \#\{j : \lambda_j^{(n)} \in B(\mathcal{ER}(f), \delta)\} = o(n)$.*

Proof. The main idea is to use the function G defined as

$$G(z) = \mathrm{dist}(z, \mathcal{ER}(f))$$

where $\mathrm{dist}(K_1, K_2)$ is the distance in Euclidean norm between two sets K_1 and K_2. It is easy to prove that G is continuous (indeed it is Lipschitz continuous). For $\epsilon > 0$, we consider the functions

$$F^\epsilon = \exp(-G(z)/\epsilon)$$

for $z \in [m_f - 1, M_f + 1]$: outside the interval $[m_f - 1, M_f + 1]$ we complete the function F^ϵ is such a way that it belongs to C_0. Since G is zero if and only if $x \in \mathcal{ER}(f)$ and positive elsewhere, it follows that $F^\epsilon \ge F$ with F being the characteristic function of the set $\mathcal{ER}(f))$ and F is the L^1 limit of F^ϵ. Therefore the use of Theorem 1 with test function F^ϵ and an asymptotic argument with ϵ arbitrarily small complete the proof. •

Therefore almost all the eigenvalues of $T_n(f)$ are in every fixed fattening of the essential range of f; anyway in [101] the following somehow surprising result is proved.

Theorem 2.4 *With the previous notations, the set $\bigcup_{n=1}^{\infty} \bigcup_{j \le n} \lambda_j^{(n)}$ is dense in $[m_f, M_f]$.*

2.2 More on the extreme eigenvalues

We present simple linear algebra techniques (see [66, 63]) in order to evaluate the rate of convergence of $\lambda_1^{(n)}$ to m_f (and $\lambda_n^{(n)}$ to M_f). By using these tools we obtain asymptotic results which extend previous estimates provided by Kac, Murdoch, Szegö [46], Parter [56], Widom [99] under certain regularity assumptions on the generating functions. In fact they obtained that $\lambda_1^{(n)} - m_f \sim n^{-2k}$ in the case where f is globally continuous on I and of class C^{2k} in a suitable neighborhood of the unique zero x_0 of $f - m_f$ with $f(x) - m_f \sim |x - x_0|^{2k}$ (in other words the key assumption is that $f - m_f$ has at x_0 a zero of order $2k$). The smoothness hypotheses impose severe restrictions which might be hard to verify or they may be not satisfied in some areas of application such as prediction theory of stationary processes and signal processing, where f is viewed as a spectral density of a stationary stochastic process [57]. However it is simple to prove that the smoothness features do not have any influence on the convergence speed of $\lambda_1^{(n)}$ to m_f and indeed also the uniqueness of the point where the minimum is attained is a removable assumption. We remark that the statements concern the smallest eigenvalue of $T_n(f)$ but the same holds for the biggest since it is enough to consider the smallest eigenvalue of $T_n(-f)$.

The following theorem is useful both for devising good preconditioners and for obtaining extremal spectral results.

Theorem 2.5 *Let f and g be two real valued integrable functions where g is nonnegative and with positive essential supremum; then for any n the matrix $T_n^{-1}(g)T_n(f)$ has eigenvalues in the open interval (r, R) where $r =$essinf $\frac{f}{g}$ and $R =$esssup $\frac{f}{g}$ if $r < R$. In the case where $r = R$ we have $T_n^{-1}(g)T_n(f) = rI_n$ with I_n being the identity matrix and trivially every eigenvalue of $T_n^{-1}(g)T_n(f)$ coincides with r.*

Proof. From the first item of Theorem 2.2, the matrix $T_n(g)$ is Hermitian and positive definite and therefore by the Schur normal form Theorem [7] there exists its square root and is a Hermitian and positive definite matrix. Thus $G_n = T_n^{-1}(g)T_n(f)$ is similar to $T_n^{-1/2}(g)T_n(f)T_n^{-1/2}(g)$ which is Hermitian thanks to the Hermitianity of $T_n^{1/2}(g)$ and of $T_n(f)$: consequently any eigenvalue of $T_n^{-1}(g)T_n(f)$ is a real number. Let λ be an eigenvalue of $T_n^{-1}(g)T_n(f)$. Then the matrix $C_n(\lambda) = T_n(f) - \lambda T_n(g)$ is singular and $C_n(\lambda)$ is a Toeplitz matrix generated by $c_\lambda(x) = f(x) - \lambda g(x)$ where we assume that $c_\lambda(x)$ is not zero a.e. (i.e. $r < R$). In view of the first item of Theorem 2.2, $c_\lambda(x)$ cannot have essentially constant sign. Therefore $f - \lambda g$ cannot be nonnegative a.e. and $f - \lambda g$ cannot be nonpositive a.e., i.e., $f - \lambda g$ is essentially nondefinite ($m\{x \in I : c_\lambda(x) < 0\},\ m\{x \in I : c_\lambda(x) > 0\} > 0$). Since $g \geq 0$ a.e. and $c_\lambda = g c_\lambda^*$ we find that $c_\lambda^* = \frac{f}{g} - \lambda$ is essentially nondefinite. This means that essinf $c_\lambda^* < 0$ and esssup $c_\lambda^* > 0$, that is, essinf $\frac{f}{g} < \lambda <$ esssup $\frac{f}{g}$. The second part of the theorem is trivial and its proof is left to the reader (it corresponds to the case $c_\lambda(x)$ zero a.e. i.e. $r = R$). •

Now we can introduce the following definitions:

Definition 2.1 *Let f, g be two nonnegative integrable functions on I (not essentially zero), then $f \sim g$ if there exists a constant $r > 0$ such that $\frac{f}{g}, \frac{g}{f} \geq r$, a.e. Moreover $f \preceq g$ if esssup $f/g = \infty$, essinf $f/g = r > 0$.*

With these definitions the following simple result is true.

Theorem 2.6 *Let f, g be two integrable functions on I satisfying the condition $f - m_f \sim g - m_g$, where m_f (and analogously m_g) is defined as in Theorem 2.2, then $\lambda_j^{(n)}(T_n(f)) - m_f \sim \lambda_j^{(n)}(T_n(g)) - m_g$, $j = 1, \ldots, n$. In particular for $j = 1$ we deduce that the minimal eigenvalue of $T_n(f)$ tends to m_f with the same asymptotic speed of the convergence of $\lambda_1^{(n)}(T_n(g))$ to m_g.*

Proof. By the minmax characterization it is well known that the j-th eigenvalue ($\lambda_1(A) \leq \cdots \leq \lambda_n(A)$) of a Hermitian matrix $A \in M_n(\mathbf{C})$ is described as

$$\lambda_j^{(n)}(A) = \min_{\dim(\mathcal{V})=j} \max_{\mathbf{x}\in\mathcal{V},\ \mathbf{x}\neq 0} \frac{\mathbf{x}^H A \mathbf{x}}{\mathbf{x}^H \mathbf{x}}.$$

By the assumption, we know that there exists positive r and R such that $r(g - m_g) \leq f - m_f \leq R(g - m_g)$ and therefore by the first item of Theorem 2.2 and by the linearity of the operator $T_n(\cdot)$, we have

$$rT_n(g - m_g) \leq T_n(f - m_f) \leq RT_n(g - m_g).$$

We recall that for Hermitian matrices A and B, $A \leq B$ is equivalent to say that $B - A$ is positive semidefinite. Consequently, for every $j = 1, \ldots, n$

$$\begin{aligned}
\lambda_j^{(n)}(T_n(f)) - m_f &= \lambda_j^{(n)}(T_n(f) - m_f I_n) \\
&= \min_{\dim(\mathcal{V})=j} \max_{\mathbf{x}\in\mathcal{V},\ \mathbf{x}\neq 0} \frac{\mathbf{x}^H (T_n(f - m_f))\mathbf{x}}{\mathbf{x}^H \mathbf{x}} \\
&\leq \min_{\dim(\mathcal{V})=j} \max_{\mathbf{x}\in\mathcal{V},\ \mathbf{x}\neq 0} \frac{\mathbf{x}^H R(T_n(g - m_g))\mathbf{x}}{\mathbf{x}^H \mathbf{x}} \\
&= \min_{\dim(\mathcal{V})=j} \max_{\mathbf{x}\in\mathcal{V},\ \mathbf{x}\neq 0} R\frac{\mathbf{x}^H (T_n(g - m_g))\mathbf{x}}{\mathbf{x}^H \mathbf{x}} \\
&= R\lambda_j^{(n)}(T_n(g) - m_g I_n) = R(\lambda_j^{(n)}(T_n(g)) - m_g).
\end{aligned}$$

In a completely analogous way we deduce that

$$\lambda_j^{(n)}(T_n(f)) - m_f \geq r(\lambda_j^{(n)}(T_n(g)) - m_g)$$

i.e. $\lambda_j^{(n)}(T_n(f)) - m_f \sim \lambda_j^{(n)}(T_n(g)) - m_g, \forall j = 1, \ldots, n.$ •
If f and g are such that $f - m_f \preceq g - m_g$, by means of the same arguments of the previous theorem, it is easy to show that

$$\lambda_1^{(n)}(T_n(f)) - m_f = o(\lambda_1^{(n)}(T_n(g)) - m_g).$$

Consequently, the set of all the functions $f \in L^\infty(I)$ such that $f - m_f$ has a unique zero of order ρ, is a class of equivalence $\mathcal{Z}_\rho$ with respect to the equivalence relation $f\mathcal{R}g$ if and only if $f - m_f \sim g - m_g$, that is, $f - m_f$ and $g - m_g$ have a zero of the same order ρ. In this way all the Toeplitz matrices T_n possessing the generating function f in $\mathcal{Z}_\rho$ have minimal eigenvalue which tends to m_f with the same asymptotical rate of convergence depending only on ρ. Now we are ready to generalize a result of Kac, Szegö and Murdoch [46] where it is proven that if $f - m_f \sim (\sin^{2k}((x - x_0)/2)$ for some $x_0 \in I$, $I = [-\pi, \pi)$ and $f \in C(\bar{I}) \cap C^{2k}(J)$, for some neighborhood J of x_0, then $\lambda_1^{(n)}(T_n(f)) - m_f \sim n^{-2k}$. The following more general result can be easily proven.

Corollary 2.1 *$f - m_f \sim \sin^{2k}((x - x_0)/2)$, then $\lambda_1^{(n)}(T_n(f)) - m_f \sim n^{-2k}$.*

Proof. The thesis holds for $g(x) = \sin^{2k}((x - x_0)/2)$ [46] which is infinitely differentiable. Now, since $f - m_f \sim g$, by using Theorem 2.6 we have $\lambda_1^{(n)}(T_n(f)) - m_f \sim \lambda_1^{(n)}(T_n(g)) \sim n^{-2k}$. •
In addition by following the same reasoning we obtain

Corollary 2.2 *If $\sin^{2k}((x - x_0)/2) \preceq f - m_f$, for every positive integer k (i.e., the order of the zero of $f - m_f$ is ∞), then $\lambda_1^{(n)}(T_n(f)) - m_f = o\left(n^{-2k}\right)$ for every positive k. Therefore we have super polynomial convergence of $\lambda_1^{(n)}(T_n(f))$ to m_f.*

Finally we end this subsection mentioning the case of several minima. It is interesting to point out that in some areas of application such as prediction theory of stationary processes and signal processing, where f is viewed as the spectral density of a stationary process, the density f may have many zeros and essinf $f = 0$ is attained at several points, so that the assumption of a unique minimum is not fulfilled [57]. In this case as well the same type of result holds. More specifically, if $f - m_f$ has a finite number of (essential) zeros of finite order with maximal order α, then $\lambda_1^{(n)}(T_n(f)) - m_f \sim n^{-\alpha}$ (for such a type of results see [14]).

As an example, let us consider the symbol $f(x) = x^2$ which has a unique zero of order two at $x_0 = 0$. The matrix $T_n(f)$ is a dense one (since the expansion is not finite $x^2 = \frac{\pi^2}{3} + 2\sum_{k=1}^{\infty} \frac{(-1)^k}{k^2}(e^{\mathbf{i}kx} + e^{-\mathbf{i}kx})$) and its eigenvalues are not explicitly known. Conversely, the simple band Toeplitz matrix (occurring in the Finite Differences discretization of the one dimensional Laplace operator with Dirichlet boundary conditions)

$$T_n(g) = \begin{bmatrix} 2 & -1 & & & \\ -1 & \ddots & \ddots & & \\ & \ddots & & \ddots & \\ & & \ddots & \ddots & -1 \\ & & & -1 & 2 \end{bmatrix}, \tag{5}$$

with $g(x) = 2 - 2\cos(x) = 4\sin^2(x/2)$, belongs to the matrix algebra [10, 33] diagonalized by the sine transform DST I: its eigenvalues are explicitly known and are given by a sample of g over the grid $\frac{j\pi}{n+1}$. Therefore the minimal eigenvalue is given by

$$4\sin^2\left(\frac{\pi}{2(n+1)}\right).$$

In conclusion by Theorem 2.6 we deduce that $\lambda_1^{(n)}(T_n(f))$ is in the open interval $\left(4\sin^2\left(\frac{\pi}{2(n+1)}\right), \pi^2\sin^2\left(\frac{\pi}{2(n+1)}\right)\right)$ since $1 = \min f/g$ and $\pi^2/4 = \max f/g$. Moreover the maximal eigenvalue $\lambda_1^{(n)}(T_n(f))$ converges to $\max f = \pi^2$ by the second item of Theorem 2.2 and therefore the spectral condition number of $T_n(f)$ grows asymptotically as n^2: we mention that the matrix $T_n(f)$ appears in the discretization of the one dimensional Laplace operator with Dirichlet boundary conditions when using the super polynomially convergent Sinc Galerkin method [52].

3 Spectral properties of $T_n^{-1}(g)T_n(f)$

In this section we take a nonnegative not identically zero function g and we consider the positive definite matrix $T_n(g)$ (the positive definiteness follows from Theorem 2.2). The motivation is computational since for a given Toeplitz matrix $T_n(f)$ we choose a function g such that $T_n(g)$ is a good preconditioner for $T_n(f)$. In particular for analyzing the performances in terms of convergence speed of such a preconditioner we need to study the spectral behavior of the sequence $\{T_n^{-1}(g)T_n(f)\}$. The present section deals with the eigenvalue analysis of $G_n = T_n^{-1}(g)T_n(f)$ where we show that almost all the results of Section 2 can be generalized to G_n where the crucial role of the generating function is played by f/g.

Theorem 3.1 *Let f and g be two integrable functions over I and let us suppose that g is nonnegative and not identically zero. Let us order the eigenvalues $\lambda_j^{(n)}$ of $G_n = T_n^{-1}(g)T_n(f)$ in nondecreasing way and let r and R be the essential infimum and the essential supremum of f/g ; the following relations hold.*

a. *G_n has eigenvalues in the open set (r, R) if $r < R$ and it coincides with rI_n if $r = R$.*

b. *If $m\{x \in I : g(x) = 0\} = 0$ then $\bigcup_{n=1}^{\infty}\bigcup_{j\le n}\lambda_j^{(n)}$ is dense in $\mathcal{ER}(f/g)$ where $\mathcal{ER}(f/g)$ is the essential range of f/g (we recall that $y \in \mathcal{ER}(h)$ if and only if for any $\epsilon > 0$ the Lebesgue measure of the set $\{x \in I : h(x) \in (y-\epsilon, y+\epsilon)\}$ is positive).*

c. *The extreme eigenvalues of G_n are such that $\lim_{n\to\infty}\lambda_1^{(n)} = r$, $\lim_{n\to\infty}\lambda_n^{(n)} = R$.*

d. *Let $C_{\text{blimits}} = \{F : \mathbf{R} \to \mathbf{R},\ F$ continuos and with finite limits at $\pm\infty\}$; then for every $F \in C_{\text{blimits}}$ we have*

$$\lim_{n\to\infty}\frac{1}{n}\sum_{j=1}^{n}F\left(\lambda_j^{(n)}\right) = \frac{1}{2\pi}\int_I F\left(f(x)/g(x)\right)dx. \tag{6}$$

Proof. Part **a** is contained in Theorem 2.5.
For part **b**, we demonstrate the general result under the only hypothesis that g is essentially nonnegative and $m\{x \in I :\ g(x) = 0\} = 0$. Actually, the thesis is equivalent to the following statement:

$$\forall\alpha \in \mathcal{ER}(f/g),\ \forall\epsilon > 0,\ \ \exists n \in \mathbf{N} \text{ and } \lambda \in \Sigma_n \text{ such that } |\lambda - \alpha| < \epsilon. \tag{7}$$

Here, we indicate by Σ_n the set of all the eigenvalues of G_n. Let $H_{n,\alpha} = T_n(f) - \alpha T_n(g)$: if $H_{n,\alpha}$ is singular for some value n then there exists $\lambda \in \Sigma_n$ such that $\lambda = \alpha$ and (7) is fulfilled. Otherwise $H_{n,\alpha}$ is nonsingular for any positive integer n. Moreover $H_{n,\alpha} = T_n(c_\alpha(x))$ where the function $c_z(x)$ is defined as $f(x) - zg(x)$, with z real parameter. Now we consider $m_\epsilon^\alpha = m\{x \in I : f - (\alpha+\epsilon)g < 0\}$ and $m_{-\epsilon}^\alpha = m\{x \in I : f - (\alpha-\epsilon)g < 0\}$. Since $g > 0$ a.e. we have $f - (\alpha+\epsilon)g < f - (\alpha-\epsilon)g$ a.e., that is, $f/g - (\alpha+\epsilon) < f/g - (\alpha-\epsilon)$ a.e.. But $\alpha \in \mathcal{ER}(f/g)$ and therefore $m_\epsilon^\alpha > m_{-\epsilon}^\alpha$. By Theorem 2.3, for every a, b, $a < b$, and for every $f \in L^1(I)$ such that $m\{x \in I : f(x) = a \text{ or } f(x) = b\} = 0$, we have

$$\lim_{n\to\infty} \frac{\#\{j : \lambda_j^{(n)}(T_n(f)) \in (a,b)\}}{n} = \frac{m\{x \in I : f(x) \in (a,b)\}}{2\pi}.$$

Consequently

$$\lim_{n\to\infty} \#\{j : \lambda_j^{(n)}(T_n(c_{\alpha+\epsilon})) < 0\} = \frac{m_\epsilon^\alpha n}{2\pi}, \tag{8}$$

$$\lim_{n\to\infty} \#\{j : \lambda_j^{(n)}(T_n(c_{\alpha-\epsilon})) < 0\} = \frac{m_{-\epsilon}^\alpha n}{2\pi}. \tag{9}$$

By using the relation $m_\epsilon^\alpha > m_{-\epsilon}^\alpha$ it follows that, for n large enough, "many" eigenvalues of $T_n(c_z)$ move from positive values to negative values when the parameter z moves from $\alpha - \epsilon$ to $\alpha + \epsilon$. As a consequence, for a large value of n, by using a continuity argument, we have to find $\lambda(n) \in (\alpha - \epsilon, \alpha + \epsilon)$ such that the matrix $T_n(c_{\lambda(n)}(x))$ is singular, i.e., $\lambda(n) \in \Sigma_n$. Therefore part **b** is proved. In equations (8), (9) we have supposed that $m\{x \in I : f - (\alpha+\epsilon)g = 0\} + m\{x \in I : f - (\alpha-\epsilon)g = 0\} = 0$. In the case where this assumption is not verified we can obviously choose ϵ^*, $0 < \epsilon^* < \epsilon$ such that

$$m\{x \in I : f - (\alpha+\epsilon^*)g = 0\} = m\{x \in I : f - (\alpha-\epsilon^*)g = 0\} = 0.$$

Moreover if the thesis of part **b** is proved for ϵ^* such that $0 < \epsilon^* < \epsilon$ then the thesis holds for ϵ.
Concerning part **c** we first prove that the extreme eigenvalues are monotone sequences. Indeed the matrix $T_n(g)$ is Hermitian and positive definite and therefore by the Schur normal form Theorem its square root is Hermitian positive definite. Therefore

$$\begin{aligned}
\lambda_1^{(n)} &= \lambda_1^{(n)}(T_n^{-1/2}(g)T_n(f)T_n^{-1/2}(g)) \\
&= \min_{\mathbf{x}\neq 0} \frac{\mathbf{x}^H T_n^{-1/2}(g)T_n(f)T_n^{-1/2}(g)\mathbf{x}}{\mathbf{x}^H\mathbf{x}} \\
&= \min_{\mathbf{z}=T_n^{-1/2}(g)\mathbf{x},\ \mathbf{x}\neq 0} \frac{\mathbf{z}^H T_n(f)\mathbf{z}}{\mathbf{z}^H T_n(g)\mathbf{z}} \\
&\leq \min_{\mathbf{z}=(\mathbf{w},0),\ \mathbf{w}\neq 0,\ \text{size}(\mathbf{w})=n-1} \frac{\mathbf{z}^H T_n(f)\mathbf{z}}{\mathbf{z}^H T_n(g)\mathbf{z}} \\
&= \min_{\mathbf{w}\neq 0} \frac{\mathbf{w}^H T_{n-1}(f)\mathbf{w}}{\mathbf{w}^H T_{n-1}(g)\mathbf{w}} \\
&= \lambda_1^{(n-1)}(T_{n-1}^{-1/2}(g)T_{n-1}(f)T_{n-1}^{-1/2}(g)) = \lambda_1^{(n-1)}.
\end{aligned}$$

Analogously $\lambda_n^{(n)} \geq \lambda_{n-1}^{(n-1)}$. Part **c** is a consequence of the monotonicity of $\lambda_1^{(n)}$ and $\lambda_n^{(n)}$ and of part **a** and **b**.

Finally we prove the last item. By Theorem 2.3, $\forall s \in \mathbf{R}$ such that $m\{x \in I : f(x) - sg(x) = 0\} = 0$ we have

$$\lim_{n\to\infty} \frac{\#\{j : \lambda_j^{(n)}(T_n(f) - sT_n(g)) > 0\}}{n} = \frac{m\{x \in I : f(x) - sg(x) > 0\}}{2\pi}. \tag{10}$$

Since $m\{x \in I : g(x) = 0\} = 0$ and since g is nonnegative, the set $\{x \in I : f(x) - sg(x) > 0\}$ coincides with $\{x \in I : f(x)/g(x) > s\}$ up to zero measure sets. Moreover the matrices

$$T_n(f) - sT_n(g) \text{ and } T_n^{-1}(g)T_n(f) - sI_n$$

have the same inertia since $T_n^{-1}(g)T_n(f) - sI_n$ is similar to

$$T_n^{-1/2}(g)T_n(f)T_n^{-1/2}(g) - sI_n$$

and the latter coincides with

$$T_n^{-1/2}(g)[T_n(f) - sT_n(g)]T_n^{-1/2}(g).$$

Therefore equation (10) is equivalent to

$$\lim_{n\to\infty} \frac{\#\{j : \lambda_j^{(n)}(G_n) > s\}}{n} = \frac{m\{x \in I : f(x)/g(x) > s\}}{2\pi} \tag{11}$$

$\forall s$ such that $m\{x \in I : f(x)/g(x) > s\} = 0$, i.e., setting F_s the characteristic function of the set $(s, +\infty)$, we find

$$\lim_{n\to\infty} \frac{1}{n} \sum_{j=1}^{n} F_s\left(\lambda_j^{(n)}(G_n)\right) = \frac{1}{2\pi} \int_I F_s\left(f(x)/g(x)\right) dx. \tag{12}$$

$\forall s$ such that $m\{x \in I : f(x)/g(x) > s\} = 0$. Now we observe that the set of the real number s such that $m\{x \in I : f(x)/g(x) = s\} \neq 0$ is at most countable since

$$\sum_s m\{x \in I : f(x)/g(x) = s\} \leq m\{I\} = 2\pi.$$

Therefore the set of the real number s such that $m\{x \in I : f(x)/g(x) = s\} = 0$ must be dense in $\mathbf{R}$. The latter remark is sufficient to conclude the proof because the infinity norm closure of the functional space $\mathcal{I}$ spanned by F_s coincides with C_{blimits} under the assumption that the numbers s can be chosen densely in $\mathbf{R}$. •

Even for the preconditioned matrices, it is interesting to analyze the extremal properties of the spectrum of $T_n^{-1}(g)T_n(f)$. In the former theorem two properties concerning the extremal eigenvalues are proven: $\lambda_1^{(n)} > r$, $\lambda_n^{(n)} < R$ and $\lambda_1^{(n)} \to r$, $\lambda_n^{(n)} \to R$. Now the question is: which are the rates of convergence of $\lambda_1^{(n)}$ to r and $\lambda_n^{(n)}$ to R? A partial answer in contained in the following result.

Proposition 3.1 [62] *Let* $f = |x - x_0|^\alpha$, $g = |x - x_0|^\beta$ *with* $0 < \beta < \alpha$ (r =essinf $f/g = 0$ *and* $0 < R$ =esssup $f/g < \infty$) *and* $G_n = T_n^{-1}(g)T_n(f)$. *Then* $\lambda_1^{(n)}(G_n) \sim n^{-(\alpha-\beta)}$.

We observe that the latter property is very useful to design good preconditioners for the problem $T_n(f)\mathbf{x} = \mathbf{b}$ in the case where f has zeros of noninteger orders too (see Subsection 4.6).

3.1 The spectral behavior of $T_n(f)$ and $T_n^{-1}(g)T_n(f)$

What Theorem 3.1 and Proposition 3.1 emphasize is that, except for minor differences in the assumptions, the spectral behavior of the preconditioned sequences is formally similar to the one of Toeplitz sequences: part **a** of Theorem 3.1 corresponds to part **a** of Theorem 2.2; part **b** of Theorem 3.1 corresponds to Theorem 2.3; part **c** of Theorem 3.1 corresponds to part **b** of Theorem 2.2; part **d** of Theorem 3.1 corresponds to the Szegö Theorem 2.1: more precisely part **d** of Theorem 3.1 contains Theorem 2.1 since $T_n^{-1}(g)T_n(f) = T_n(f)$ for $g \equiv 1$ a.e. and generalizes Theorem 2.1 since C_0 is a proper subset of C_{blimits}. Finally the analogous of Proposition 3.1 is expressed by the results in Subsection 2.2. In particular we stress that the role played by the generating function f is played by f/g. We notice that while f has to be Lebesgue integrable, the function f/g is only measurable and in fact for every measurable function h defined over I we can find f and g in $L^1(I)$ with nonnegative g such that $h = f/g$.

The only result which is not mentioned is Theorem 2.4 for which a counterpart in the preconditioned case cannot be found. Indeed, while the closure of $\bigcup_{n=1}^{\infty} \bigcup_{j \le n} \lambda_j^{(n)}$ with $\lambda_j^{(n)}$ being the eigenvalues of $T_n(f)$ coincides with $[\text{essinf } f, \text{esssup } f]$ (even when the set $\mathcal{ER}(f/g)$ is not connected), the closure of $\bigcup_{n=1}^{\infty} \bigcup_{j \le n} \lambda_j^{(n)}$ with $\lambda_j^{(n)}$ being the eigenvalues of $T_n^{-1}(g)T_n(f)$ is a closed subset of $[\text{essinf } f/g, \text{esssup } f/g]$ (containing $\mathcal{ER}(f/g)$) which can be disconnected if $\mathcal{ER}(f/g)$ is disconnected (see [42]). This property has a computational impact in the preconditioning algorithms for nondefinite systems as discussed in Subsection 4.4.

4 The band Toeplitz preconditioning

The main idea of this section is to use $T_n(g)$ as preconditioner for $T_n(f)$. Of course this proposal makes sense only if the cost of solving a generic system with coefficient matrix $T_n(g)$ is sensibly lower than the cost of solving a generic system with coefficient matrix $T_n(f)$: for instance if $T_n(f)$ is full and $T_n(g)$ is banded then by using a band solver for the last one (by exploiting the band structure and by ignoring the Toeplitz structure), we have a sensible gain if the spectrum of the preconditioned matrix $T_n^{-1}(g)T_n(f)$ is far away from zero and from infinity uniformly with respect to n.

Consider the function $f(x) = x^2$ which has a unique zero of order two at $x_0 = 0$. The matrix $T_n(f)$ is full and ill-conditioned (as shown a the end of Section 2 its spectral condition number is asymptotic n^2), but can be optimally preconditioned by the simple band Toeplitz matrix

$$T_n(g) = \begin{bmatrix} 2 & -1 & & & \\ -1 & \ddots & \ddots & & \\ & \ddots & & \ddots & \\ & & \ddots & \ddots & -1 \\ & & & -1 & 2 \end{bmatrix}$$

with $g(x) = 2 - 2\cos(x) = 4\sin^2(x/2)$. By the parts **a** and **d** of Theorem 3.1, we deduce that the eigenvalues of $T_n^{-1}(g)T_n(f)$

1. are contained in the interval $(1, \pi^2/4)$, $1 = \min f/g$, $\pi^2/4 = \max f/g$ and
2. are globally distributed as the function f/g.

Therefore, from item 1. and by invoking classical convergence results on the preconditioned conjugate gradient (PCG) method [3], we know that a constant number of iterations is sufficient for reaching the solution within a preassigned accuracy. Moreover the distribution result in item 2. tells us that the estimate that we obtain from the Axelsson and Linskog bounds are tight (see also [5]) since the eigenvalues are more or less equally distributed in $(1, \pi^2/4)$.

In the next subsections first we introduce the notion of asymptotically difficult problems and of optimal methods and then we discuss how to obtain optimal PCG methods for (difficult) Toeplitz problems

4.1 Iterative methods and Toeplitz matrices

Often in applications a (large) linear system $A_n\mathbf{x}_n = \mathbf{b}_n$ is obtained as an approximation (discretization) of a problem in a infinite dimensional space: this situation typically occurs in the case of PDEs, integral equations etc. Then usually the larger is the dimension the more accurate is the solution but, in most of the cases, the condition number of A_n diverges to infinity as n tends to infinity: in the positive definite case, without loss of generality, we can assume that the maximal eigenvalue tends to a positive constant and therefore we say that the sequence of problems with coefficient matrices in $\{A_n\}$ is difficult if the minimal eigenvalue $\lambda_1^{(n)}$ tends to zero as n tends to infinity. In the Toeplitz setting, by Theorem 2.4, we know that this phenomenon arises if and only if zero in contained in the convex hull of $\mathcal{ER}(f)$. If the symbol f is nonnegative the latter is equivalent to say that f has essential zeros. In the case of difficult problems the direct methods can be unstable and often are too costly since they do not exploit the structure while the iterative ones are more accurate and have moderate cost per iteration since in general it is quite easy to exploit the structure of the problem: however they can be very slow due to the vanishing eigenvalues.

Therefore for these types of difficult problems we are interested in optimal methods that is in methods such that the cost of the (inverse) problem of solving $A_n\mathbf{x}_n = \mathbf{b}_n$ with generic $\mathbf{b}_n$ is at most proportional to the cost of multiplying the coefficient matrix A_n by a generic vector $\mathbf{c}_n$. In the context of iterative solvers, the latter can be translated in the following definition (see also notions **opt1** and **opt2** in Appendix **A** and [4]).

Definition 4.1 *An iterative method is optimal for a class of problems*

$$A_n\mathbf{x}_n = \mathbf{b}_n,$$

if, uniformly with respect to the dimension n of the problem, we have:

1. *the cost per iteration is proportional to the matrix vector product with a generic vector;*
2. *for any fixed accuracy ϵ, the number of iterations for reaching the solution within the given accuracy is bounded by a constant independent of n and possibly depending on ϵ.*

Focusing our attention on the preconditioned conjugated gradient (PCG) method [41] as an iterative solver, the above definition implies that for every n we should be able to find a preconditioner P_n such that **a)** the solution of a generic system $P_n\mathbf{y}_n = \mathbf{c}_n$ has computational cost bounded by the matrix vector product with matrix A_n and **b)** the spectrum of $P_n^{-1}A_n$ is bounded away from zero and from infinity uniformly with respect to n: for more details on the (P)CG methods and on their convergence results see Appendix **A** and especially Theorem 8.3. Clearly, the two issues **a)** and **b)** are often conflicting since when a matrix P_n is too close to A_n (requirement **b)**) it also requires the same computational effort to invert (so contradicting requirement **a)**). However in the case of Toeplitz

matrices associated to a symbol ($A_n = T_n(f)$) a satisfactory solution can be found: tools for dealing with requirement **b)** have been reported in the last section while for dealing with requirement **a)** we have study the computational cost of a matrix vector product when a Toeplitz matrix is involved. This computational cost is proportional to $n\log(n)$ and can be achieved by using the Fast Fourier Transform (FFT) [97] as reported in the next subsection.

4.1.1 Toeplitz, circulants and the Fourier matrix: the matrix vector product

A generic circulant matrix A_n of size n is defined by n parameters $a_0, \ldots, a_{n-1}$ and is characterized by a circular structure:

$$A_n = \begin{pmatrix} a_0 & a_1 & a_2 & \ldots & a_{n-3} & a_{n-2} & a_{n-1} \\ a_{n-1} & a_0 & a_1 & a_2 & \ldots & a_{n-3} & a_{n-2} \\ & \ddots & \ddots & \ddots & & & \\ \vdots & & \ddots & \ddots & \ddots & & \vdots \\ & & & \ddots & \ddots & \ddots & \\ a_2 & \ldots & a_{n-3} & a_{n-2} & a_{n-1} & a_0 & a_1 \\ a_1 & a_2 & \ldots & a_{n-3} & a_{n-2} & a_{n-1} & a_0 \end{pmatrix}. \tag{13}$$

The j-th of A_n is the periodic forward shift of the preceding $(j-1)$-th row (periodic since the last element of the row $j-1$ becomes the first of the subsequent row j): we observe that also the concepts of "preceding" and "subsequent" have to be intended periodically because for $j = 1$ the preceding row is the n-th and analogously for $j = n$ the subsequent row is the first row.

We denote by $\mathcal{C}_n$ the class of circulant matrices

$$\mathcal{C}_n = \{A_n \in M_n(\mathbf{C}) \text{ such that } A_n \text{ is of the form (13) with } a_j \in \mathbf{C}\}. \tag{14}$$

It is is easy to verify that $\mathcal{C}_n$ is a vector space since it is closed under complex linear combinations. By denoting by $\mathcal{T}_n$ the space of complex n by n Toeplitz matrices, it is evident that $\mathcal{C}_n$ is a proper set of $\mathcal{T}_n$.

By a direct check in the structure displayed in (13), every $A_n \in \mathcal{C}_n$ can be written as

$$A = a_0 Z_0 + a_1 Z_1 + a_2 Z_2 + \ldots + a_{n-1} Z_{n-1}, \tag{15}$$

where

$$Z_0 = \begin{pmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{pmatrix}, \qquad Z_1 = \begin{pmatrix} 0 & 1 & 0 & \ldots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & 0 \\ 0 & & & \ddots & 1 \\ 1 & 0 & \ldots & \ldots & 0 \end{pmatrix},$$

$$Z_2 = \begin{pmatrix} 0 & 0 & 1 & 0 & 0 \\ \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & \ddots & \ddots & \ddots & 1 \\ 1 & \ddots & \ddots & \ddots & 0 \\ 0 & 1 & 0 & \ldots & 0 \end{pmatrix}, \ldots, Z_{n-1} = \begin{pmatrix} 0 & \ldots & \ldots & 0 & 1 \\ 1 & \ddots & & & 0 \\ 0 & \ddots & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \ldots & 0 & 1 & 0 \end{pmatrix}.$$

The matrices $\{Z_j : j = 0, \ldots, n-1\}$ are special circulant matrices which are linear independent since every Z_j has ones in positions where all the Z_k, $k \neq j$, have null values. Therefore $\{Z_j : j = 0, \ldots, n-1\}$ is a canonical basis for $\mathcal{C}_n$ and its dimension is n (the number of the parameters). Again by direct computation it can be proven that the matrices Z_j, $j = 0, \ldots, n-1$, are related each other; more precisely we have

$$\begin{aligned} Z_j &= Z_1^j \qquad \forall j = 0, \ldots, n-1, \\ Z_1^j &= Z_1^{j \bmod n} \qquad \forall j \in \mathbf{Z} \end{aligned} \tag{16}$$

and hence the generic circulant matrix A_n is represented as

$$A = \sum_{j=0}^{n-1} a_j Z_1^j \equiv p(Z_1) \qquad p(z) = \sum_{j=0}^{n-1} a_j z^j, \ z \in \mathbf{C}. \tag{17}$$

A consequence of (17) and of (16) is that $\mathcal{C}_n$ is closed under the matrix product and therefore, by the Cayley-Hamilton, it is closed under inversion: we give an explicit proof of these two statements. Let $A_n, B_n \in \mathcal{C}_n$, then

$$\begin{aligned} A_n &= \sum_{j=0}^{n-1} \alpha_j Z_1^j, \\ B_n &= \sum_{j=0}^{n-1} \beta_j Z_1^j \end{aligned}$$

and therefore

$$A_n B_n = \sum_{j=0}^{2n-2} \gamma_j Z_1^j$$

with $\gamma_j = \sum_{k+t=j} \alpha_k \beta_t$. Now we recall (16) i.e. $Z_1^n = I$, $Z_1^{jn+k} = Z_1^k$ and we conclude

$$A_n B_n = \sum_{j=0}^{n-1} (\gamma_j + \gamma_{j+n}) Z_1^j \in \mathcal{C}_n \qquad \gamma_{2n+1} = 0$$

and in addition $A_n B_n = B_n A_n$. For the closure under inversion it is enough to use the closure under matrix product and the Cayley-Hamilton theorem (see e.g. [86]).

Theorem 4.1 *Let $A \in M_n(\mathbf{C})$ and let p_A its characteristic polynomial defined as $p_A(t) = \det(A - tI)$, then*

a. $p_A(A) = 0$;

b. *moreover $p_A(t) = \det(A) - t q_A(t)$ with q_A suitable polynomial of degree at most $n-1$:*

Therefore by combining items **a** and **b** of the Cayley-Hamilton Theorem 4.1, the inverse of an invertible matrix A is a polynomial of the matrix and more precisely we have $A^{-1} = \det^{-1}(A) q_A(A)$. Finally, in the circulant case, by the closure under matrix products and linear combinations, the inverse of an invertible circulant matrix is still circulant.

Now we prove a spectral decomposition of a generic circulant matrix A_n. The key tool is the Fourier matrix

$$F_n = \frac{1}{\sqrt{n}} \left(e^{\frac{\mathrm{i}2\pi kj}{n}} \right)_{j,k=0}^{n-1}.$$

A simple computation shows the following features:

p1. F_n is unitary i.e. $F_n^H F_n = I$;

p2. $F_n^H = PF_n$ with P permutation matrix with $P_{i,j} = 1$ if and only if $(i+j) \mathrm{mod}\, n = 2$;

p3. F_n is complex symmetric i.e. $F_n = F_n^T$ and therefore $F_n^H = F_n P$ by **p2**.

Now, if we define $\mathbf{f}_j$ the j-th column of F_n

$$\mathbf{f}_j = \frac{1}{\sqrt{n}} \left(e^{\frac{i2\pi kj}{n}} \right)_{k=0}^{n-1},$$

we have

$$Z_1 \mathbf{f}_j = \frac{1}{\sqrt{n}} \begin{pmatrix} e^{\frac{i2\pi j}{n}} \\ e^{\frac{i2\pi j2}{n}} \\ \vdots \\ e^{\frac{i2\pi j(n-1)}{n}} \\ e^{\frac{i2\pi j0}{n}} \end{pmatrix} = \frac{1}{\sqrt{n}} e^{\frac{i2\pi j}{n}} \begin{pmatrix} e^{\frac{i2\pi j0}{n}} \\ e^{\frac{i2\pi j1}{n}} \\ \vdots \\ e^{\frac{i2\pi j(n-2)}{n}} \\ e^{\frac{i2\pi j(n-1)}{n}} \end{pmatrix} = e^{\frac{i2\pi j}{n}} \mathbf{f}_j,$$

and therefore $e^{\frac{i2\pi j}{n}}$ is an eigenvalue of Z_1 with unitary eigenvector $\mathbf{f}_j$, i.e., setting $\omega_1 = e^{\frac{i2\pi}{n}}$, we deduce

$$Z_1 \mathbf{f}_j = \omega_1^j \mathbf{f}_j, \qquad \forall j = 0, \ldots, n-1.$$

By putting together all the previous relations, we find $Z_1 = F_n \Lambda F_n^H$ where Λ is diagonal with j-th diagonal entry given by $\omega_1^j, \quad j = 0, \ldots, n-1$: in other words the eigenvalues of Z_1 are all the n-th roots of unity. Now we observe that $Z_1^j = F_n \Lambda^j F_n^H$, $j = 0, \ldots, n-1$, and we conclude that

$$\begin{aligned} A_n &= \sum_{j=0}^{n-1} a_j Z_1^j \\ &= \sum_{j=0}^{n-1} a_j F_n \Lambda_n^j F_n^H \\ &= F_n \left(\sum_{j=0}^{n-1} a_j \Lambda_n^j \right) F_n^H \\ &= F_n \Lambda_n(a) F_n^H \end{aligned}$$

where

$$(\Lambda_n(a))_{kk} = \sum_{j=0}^{n-1} a_j (\omega_1^k)^j, \tag{18}$$

i.e., calling $\underline{\lambda}$ the vector whose k-th entry is $(\Lambda_n(a))_{kk}$ and $\mathbf{a}$ the vector whose k-th entry is a_k we have

$$\underline{\lambda} = \sqrt{n} F_n \mathbf{a}. \tag{19}$$

In conclusion the circulants coincide with the set of all the matrices which are simultaneously diagonalized by the discrete Fourier transform F_n. Therefore if we have to compute a matrix vector product or we have solve a linear system with circulant coefficient matrix A_n, then, thanks to the Schur decomposition $A_n = F_n \Lambda_n(a) F_n^H$, thanks to item **p2** ($F_n^H = PF_n$) and thanks to relation (19), three discrete Fourier transforms are sufficient for performing the above mentioned calculations. In view of the complexity of a Fast Fourier Transform (see Subsection 4.1.2), the circulant matrix vector product can be performed in $9/2n\log(n) + O(n)$ arithmetic operations.

Now we prove that the Toeplitz matrix vector product can be achieved in $O(n\log(n))$ operations by using a suitable circulant embedding.

Theorem 4.2 *Let $T_n \in \mathcal{T}_n$ be a generic Toeplitz matrix and let $\mathbf{v} \in \mathbf{C}^n$. The matrix vector product $T_n\mathbf{v}$ can be performed in $O(f(n)log(f(n)))$ floating point operations where $f(n)$ is a function bounded from below by $2n$ and bounded by above by $4n-4$. More precisely the computation of $T_n\mathbf{v}$ is carried out by performing a circulant matrix vector product $A_{f(n)}\mathbf{v}^*$ with $A_{f(n)} \in \mathcal{C}_{f(n)}$ and $\mathbf{v}^* \in \mathbf{C}^{f(n)}$ ($A_{f(n)}$ depending on T_n and $\mathbf{v}^*$ depending on $\mathbf{v}$).*

proof Let $T_n \in \mathcal{T}_n$; then

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

The idea. We "embed" the matrix T_n into a circulant matrix $A_{\hat{f}(n)} \in \mathcal{C}_{\hat{f}(n)}$ with $\hat{f}(n)$ minimal dimension:

$$A_{\hat{f}(n)} = \left(\begin{array}{ccccc|ccc} t_0 & t_{-1} & \dots & t_{-(n-2)} & t_{-(n-1)} & t_{n-1} & \dots & t_1 \\ t_1 & \ddots & \ddots & \ddots & t_{-(n-2)} & t_{-(n-1)} & \ddots & t_2 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ t_{n-2} & \ddots & \ddots & \ddots & t_1 & t_2 & \ddots & t_{n-1} \\ t_{n-1} & t_{n-2} & \dots & t_1 & t_0 & t_{-1} & \ddots & t_{-(n-1)} \\ \hline t_{-(n-1)} & \ddots & \ddots & \ddots & t_1 & t_0 & \ddots & t_{-(n-2)} \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ t_{-1} & t_{-2} & \dots & t_{-(n-1)} & t_{n-1} & t_{n-2} & \dots & t_0 \end{array}\right).$$

The minimal dimension is exactly $\hat{f}(n) = 2n-1$. The problem is that $2n-1$ is an odd number and the classical FFT implementation of $O(n\log(n))$ cost is obtained on dimensions which are powers of 2 (this assumption can be relaxed in many ways [97]). Therefore we add in the first row of $A_{\hat{f}(n)}$ as

many zeros between $t_{-(n-1)}$ and t_{n-1} so that we reach a suitable power of 2. More precisely we are interested in $f(n) = 2^k$, such that does not exist h for which $2n-1 < 2^h < 2^k$ (we take as $f(n)$ the closest power of 2 which approximates from above $2n-1$): the optimal configuration is when $n = 2^q$ for a given positive integer q. In such a case, $2n-1 = 2^{q+1}-1$ and it is enough to choose $f(n) = 2^{q+1} = 2n$; the worst case is observed for $n = 2^q + 1$ for a given positive integer q. In such a case $2n-1 = 2^{q+1}+1$ and therefore we have $f(n) = 2^{q+2} = 4n-4$.

Therefore the matrix $A_{f(n)}$ is defined and it can be partitioned as

$$A_{f(n)} = \left(\begin{array}{c|c} T_n & X_1 \\ \hline X_2 & X_3 \end{array}\right);$$

as vector $\mathbf{v}^*$ we consider

$$\begin{pmatrix} \mathbf{v} \\ \mathbf{w} \end{pmatrix}$$

with $\mathbf{w} = 0$. Therefore

$$A_{f(n)}\mathbf{v}^* = \begin{pmatrix} T_n\mathbf{v} + X_1\mathbf{w} \\ X_2\mathbf{v} + X_3\mathbf{w} \end{pmatrix}$$

and therefore by recalling that $\mathbf{w} = 0$ we conclude that the first n entries of $A_{f(n)}\mathbf{v}^*$ form exactly the vector $T_n\mathbf{v}$. In that way, thanks to the cost analysis of the FFT reported in (20), the total arithmetic cost is given by $\frac{9}{2}f(n)(\log(f(n))) - \frac{f(n)}{2}$ which is bounded by $18n\log(n) + o(n)$. •

4.1.2 The product $F_n\mathbf{v}$: the basics of the FFT algorithm

Let n be a power of two, $n = 2^k$, k positive integer, and let $N = \frac{n}{2}$. We consider the product

$$\mathbf{y} = V_n\mathbf{x}$$

with $V_n = \sqrt{n}F_n$ scaled Fourier matrix and $\mathbf{x}$ generic vector of $\mathbf{C}^n$. The k-th entry y_k of the resulting vector $\underline{y}$ is given by

$$y_k = \sum_{j=0}^{n-1} e^{\frac{i2\pi kj}{n}} x_j.$$

Let us decompose the whole into two partial sums the first containing all the even indices j and the second containing all the odd indices j:

$$\begin{aligned} y_k &= \sum_{j=0}^{N-1} e^{\frac{i2\pi k2j}{n}} x_{2j} + \sum_{j=0}^{N-1} e^{\frac{i2\pi k(2j+1)}{n}} x_{2j+1} \\ &= \sum_{j=0}^{N-1} e^{\frac{i2\pi kj}{N}} x_{2j} + \left(\sum_{j=0}^{N-1} e^{\frac{i2\pi kj}{N}} x_{2j+1}\right) e^{\frac{i2\pi k}{n}}. \end{aligned}$$

We observe that the first summation for $k = 0, \ldots, N-1$ represents the k-th entry of product between V_N and the vector $\mathbf{x}_{\text{even}}$ whose k-th entry is given by the $(2k)$-th entry x_{2k} of $\mathbf{x}$. Analogously, setting $D(n) = \text{diag}_{0\le k\le N-1}\left(e^{\frac{i2\pi k}{n}}\right)$, the second summation for $k = 0, \ldots, N-1$ represents the

product between $D(n)V_N$ and $\mathbf{x}_{\rm odd}$ whose k-th entry is given by the $(2k+1)$-th entry x_{2k+1} of $\mathbf{x}$. Therefore in matrix vector notation we have

$$(\mathbf{y})_{k=0}^{N-1} = V_N \mathbf{x}_{\rm even} + D(n) V_N \mathbf{x}_{\rm odd},$$

For the remaining part of the vector $\mathbf{y}$ that is for the indices $k = N, \ldots, n-1$, we need to use a slightly modified argument. We set $k = k' + N$ con $k' = 0, \ldots, N-1$, and we get

$$e^{\frac{i2\pi k}{n}} = e^{\frac{i2\pi(k'+N)}{n}} = e^{\frac{i2\pi k'}{n}} e^{\frac{i2\pi N}{n}} = e^{\frac{i2\pi k'}{n}} e^{\frac{i2\pi n}{2n}} = e^{\frac{i2\pi k'}{n}} e^{i\pi} = -e^{\frac{i2\pi k'}{n}},$$

so that

$$(\mathbf{y})_{k=N}^{n-1} = V_N \mathbf{x}_{\rm even} - D(n) V_N \mathbf{x}_{\rm odd}.$$

From the preceding expansions we clearly see the recursive structure of the matrix V_n to which it is naturally associated a recursive (divide and conquer) algorithm:

$$V_n = (V_2 \otimes I_N) \begin{pmatrix} I_N & \\ & D(n) \end{pmatrix} (I_2 \otimes V_N) P_{\rm even-odd}$$

with $P_{\rm even-odd}$ permutation matrix that puts in the first N positions the ordered vector of the even numbers and in the last N positions the ordered vector of the odd numbers.

Therefore to compute a discrete Fourier transform of size n we have compute two discrete Fourier transforms of order $N = \frac{n}{2}$ plus a linear amount of multiplicative and additive operations. From this we can derive the computational cost of this algorithm which represents the essence of the FFT: we denote by $C_{\rm m}(n)$ and $C_{\rm a}(n)$ the multiplicative and additive cost of a FFT of length n. A FFT of size n requires:

- two FFTs of size $N = \frac{n}{2}$,
- one product $D(n)[V_N \mathbf{x}_{\rm odd}]$,
- two sums of vectors of length $N = \frac{n}{2}$.

Therefore

$$\begin{aligned} C_{\rm m}(n) &= 2C_{\rm m}\left(\frac{n}{2}\right) + \frac{n}{2} && \text{(first recursive call)} \\ &= 2\left(2C_m\left(\frac{n}{4}\right) + \frac{n}{4}\right) + \frac{n}{2} \\ &= 4C_{\rm m}\left(\frac{n}{4}\right) + 2\frac{n}{2} && \text{(second recursive call)} \\ &= 2^j\left(C_{\rm m}\left(\frac{n}{2^j}\right)\right) + j\frac{n}{2} && (j\text{-th recursive call}). \end{aligned}$$

The maximal number of recursive calls is $\log_2(n) - 1$, since the multiplicative cost of a FFT of length 2 is zero due to the special structure of

$$V_2 = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}.$$

In summary we find

$$C_{\mathrm{m}}(n) = (\log_2(n) - 1)\frac{n}{2}.$$

For the evaluation of the additive cost we proceed analogously:

$$\begin{aligned} C_{\mathrm{a}}(n) &= 2C_{\mathrm{a}}\left(\frac{n}{2}\right) + n && \text{(first recursive call),} \\ &= 2\left(2C_{\mathrm{a}}\left(\frac{n}{4}\right) + \frac{n}{2}\right) + n \\ &= 4C_{\mathrm{a}}\left(\frac{n}{4}\right) + 2n && \text{(second recursive call),} \\ &= 2^j\left(C_{\mathrm{a}}\left(\frac{n}{2^j}\right)\right) + j \cdot n && (j\text{-th recursive call).} \end{aligned}$$

In this case the reasonable maximal number of recursive calls is $\log_2(n)$ and consequently

$$C_{\mathrm{a}}(n) = n\log_2(n).$$

In conclusion the total cost is exactly given by

$$\begin{aligned} C_{\mathrm{algorithm}} &= C_{\mathrm{m}}(n) + C_{\mathrm{a}}(n) \\ &= (\log_2(n) - 1)\frac{n}{2} + n\log_2(n) = \frac{3}{2}n\log_2(n) - \frac{n}{2}, \end{aligned} \tag{20}$$

if we assume that one multiplication/division takes approximately the same time as an algebraic sum (see the subsequent remark).

Remark The use of the primary school algorithms for making a division and a multiplication between numbers with t digits emphasizes that their cost in terms of digit operations is proportional to t^2 while the digit cost of an algebraic sum is proportional to t. Therefore it seems unreasonable to assume that the multiplicative unit cost and the additive unit cost are roughly equal. However thanks to the use of FFT based algorithms it has been shown that the product and the division between numbers of t digits can be performed in $tp(\log(t))$ digit operations where p is a suitable low degree polynomial (see e.g. [13]). In the light of these results, it makes sense to assume that the 4 basic operations have approximately the same time cost.

4.2 The case of f (essentially) nonnegative

In the case of $f \geq 0$ (and similarly $f \leq 0$), since we are looking for band Toeplitz preconditioners, we consider nonnegative trigonometric polynomials g having the same zeros of f and the same orders. But g is a polynomial and therefore infinitely differentiable: it follows that it can have only a finite number of zeros of even orders. As a consequence we have to restrict our attention to the case where f is nonnegative and has only a finite number of zeros of even orders. Under to above mentioned assumptions, in the light of Corollary 2.1 (and its generalizations to many zeros [14]) and of the item **a** of Theorem 3.1, the condition number of the system moves from $O(n^{2k})$ -condition number of the original coefficient matrix $T_n(f)$- to $O(1)$ -the (spectral) condition number of $T_n^{-1}(g)T_n(f)$- where $2k$ is the maximum of the orders of the zeros of f. The PCG method proposed in the case of $f \geq 0$ a.e. with zeros of even orders is the following.

Algorithm 4.1($f \geq 0$, zeros of even order)

Input. The zeros $x_1, \ldots, x_j$ of f with orders $2k_1, \ldots, 2k_j$, $T_n(f)$, $\mathbf{b}$.

Step 1. Consider the polynomial $g = g_{\min}$ of smallest degree with the same zeros of f and with the same orders: $g_{\min} = \prod_{i=1}^{j}(2 - 2\cos(x - x_i))^{l_i}$, $\quad b = \text{degree}(g_{\min}) = \sum_{i=1}^{j} 2k_i$.

Step 2. Apply the PCG method to the system $T_n(f)\mathbf{x} = \mathbf{b}$ using $T_n(g)$ as preconditioner.

The calculation of the coefficients of $T_n(g)$, $g = g_{\min}$ can be done in a time independent of n as its bandwidth is $2b+1$ which is constant with respect to n. Since $0 < r \leq \frac{f}{g} \leq R < \infty$, it is obvious that the number of iterations $N(\epsilon)$ performed in order to reach the solution within a preassigned accuracy ϵ, is bounded by a constant and, in particular, the following relation holds (see Theorem 8.3):

$$N(\epsilon) = k^*(r, R, \epsilon) = \left\lceil \log\left[2\epsilon^{-1}\right] / \log\left[\frac{\sqrt{R} + \sqrt{r}}{\sqrt{R} - \sqrt{r}}\right] \right\rceil. \tag{21}$$

Finally by using a standard band solver for the systems whose coefficients matrix is $T_n(g)$ we have a sequential cost of $O(n\log(n))$ arithmetic ops, due to the applications of few FFTs in the product of $T_n(f)$ by a vector. In a parallel PRAM model of computation Algorithm 4.1 takes $O(\log(n))$ steps if we use one of the algebraic methods proposed in [8] in the solution of the preconditioning systems.

As an example, consider the function $f(x) = x^2$. The matrices $T_n(f)$ show a condition number growing as n^2 as proven at the end of Section 2. From the discussion above, the preconditioner of minimal bandwidth is given by

$$g(x) = 2 - 2\cos x.$$

In the following, for $n = 200$ we report the error reduction in ∞-norm when we choose I_n and $T_n(g)$ as preconditioners; all the calculations have been made in MATLAB.

Table 1: PCG, original coefficient matrix $T_n(f)$, $f(x) = x^2$, convergence history with no preconditioning and preconditioner $T_n(g)$, $g(x) = 2 - 2\cos(x)$

step	I_n	$T_n(g)$
1	$9.49644E-01$	$3.92508E-01$
2	$9.44697E-01$	$1.27115E-01$
3	$9.03174E-01$	$3.56149E-02$
4	$8.28545E-01$	$7.74028E-03$
5	$7.96437E-01$	$1.77635E-03$
6	$7.63530E-01$	$4.34434E-04$
7	$7.37671E-01$	$1.02930E-04$
8	$7.13113E-01$	$2.35099E-05$
9	$6.99331E-01$	$5.97104E-06$
10	$6.81484E-01$	$1.36921E-06$

The mean reduction rates are 0.9264 and 0.2592 respectively: the difference is not trivial since after 10 steps we have $(0.9264)^{10} \approx 0.6815$ when no preconditioning is used and $(0.2592)^{10} \approx 1.369 * 10^{-6}$ when an appropriate band Toeplitz preconditioner is applied.

4.3 Fast band Toeplitz preconditioning

When f has zeros of even orders ($f \geq 0$), as discussed in Subsection 4.1, the main idea is to find a nonnegative trigonometric polynomial g for which $0 < r \leq \frac{f}{g} \leq R < \infty$ a.e. The associated band Toeplitz preconditioner $T_n(g)$ is the desired preconditioner in the sense that the spectrum of $T_n^{-1}(g)T_n(f)$ lies (see Theorem 3.1) in the open interval (r, R) for any dimension n.

A further possibility (see [23]) is to to increase the bandwidth of $T_n(g)$ to get extra degrees of freedom. The generating function $g = g_{\rm opt}$ is computed by using an adapted version of Remez algorithm with the aim of minimizing the relative error $h = \|(f-g)/f\|_\infty$ over all the polynomials g of fixed degree l. In [65] it is proven that this minimization property enables one not only to match the zeros of f but also to minimize R/r and consequently $N(\epsilon)$: indeed in the light of (21) it is evident that the quantity $k^*(r, R, \epsilon)$ is an increasing function of R/r. Therefore, from items **a** and **c** of Theorem 3.1, we obtain that $T_n(g)$, $g = g_{\rm opt}$, is the best band Toeplitz preconditioner in the class of all the band Toeplitz matrices of fixed bandwidth $2l + 1$.

However the Remez algorithm can be heavy from a computational point of view, it is not easy to implement and can suffer from instabilities since its basis in our specific context is made by $\{1, \sin(qx)/f(x), \cos(qx)/f(x) : q = 1, \ldots, l\}$ where $1/f$ is unbounded. There exist quasi optimal alternatives that we can consider [65] where the idea is again to minimize "in a certain sense" $(f-g)/f$. If $g_{\rm min}$ is the polynomial of minimum degree k containing all the zeros of f with their orders, then the generating function g of our preconditioners it is chosen in the following way:

$$g = g_{\rm min} g_{\,l-b}, \quad \text{degree}(g) = l \geq b, \;\; b = \text{degree}(g_{\rm min}).$$

$g_{\,l-b}$ is a trigonometric polynomial of degree $l - b$ and is defined, for example, in the light of these two strategies.

A $g_{\,l-b}$ is the best Chebyshev approximation of $\hat{f} = f/g_{\rm min}$, i.e.,

$$\|\hat{f} - g_{\,l-b}\|_\infty = \min_{g \in \mathbf{P}_{l-b}} \|\hat{f} - g\|_\infty.$$

In this case we set $g = g_A$.

B $g_{\,l-b}$ is the trigonometric polynomial of degree at most $l - b$ interpolating $\hat{f}$ at the $l - b + 1$ zeros of the $(l - b + 1)$–th Chebyshev polynomial of the first kind. In this case we set $g = g_B$.

Observe that we cannot choose g directly like the best Chebyshev approximation of f for two reasons: we are not guaranteed that g is nonnegative since f has zeros (as a consequence, by the second item of Theorem 2.2, $T_n(g)$ is not positive definite for n large enough and cannot be used as preconditioner) and we are not sure that f/g and g/f are bounded because, in general, g has different zeros with respect to f. From a computational point of view we remark that $g_{\,l-b}$ in **A** can be calculated by using the standard Remez algorithm [58] with respect to the classical trigonometric basis $\{1, \sin(qx), \cos(qx) : q = 1, \ldots, l-b\}$, while the calculation of $g_{\rm opt}$ in [23] is performed by using a modified version of the Remez algorithm [88] with the basis $\{1, \sin(qx)/f(x), \cos(qx)/f(x) : q = 1, \ldots, l\}$, in this case it is possible to observe instability problems due to the fact that f has zeros. On the other hand, for the calculation of $g_{\,l-b}$ in **B** we do not observe computational problems: this polynomial can be calculated, very easily, with few FFTs of order $(l - b)$ by means of a classical trigonometric representation of the interpolating polynomial at Chebyshev zeros. On the other hand, by defining

$$\begin{array}{ll} r^* = \inf_{x\in I} f(x)/g_{\rm opt}(x), & R^* = \sup_{x\in I} f(x)/g_{\rm opt}(x), \\ r^A = \inf_{x\in I} f(x)/g_A(x), & R^A = \sup_{x\in I} f(x)/g_A(x), \\ r^B = \inf_{x\in I} f(x)/g_B(x), & R^B = \sup_{x\in I} f(x)/g_B(x), \end{array}$$

and

$$\mu^* = \frac{R^*}{r^*},\ \ \mu^A = \frac{R^A}{r^A},\ \ \mu^B = \frac{R^B}{r^B},$$

it is easy to prove that

$$\mu^* \leq \mu^A \leq \mu^B \tag{22}$$

and therefore by relation (21) we have that the performances (in terms of convergence speed) of the strategy **A** are better than those of strategy **B** and of course worse than those of the optimal strategy: however, as shown in some subsequent numerical tests, the convergence speed in the three cases is practically equal and therefore the third strategy results to be the best one since it is the cheapest and the simplest.

In conclusion in both the strategies **A** and **B** the polynomial g is easier to calculate and the preconditioned systems have an $O(1)$ condition number for which upper bounds, depending on l, n and on the regularity features of f, can be derived by using standard approximation theory tools (see [65]). Therefore we can estimate the number of iterations to reach the solution within a preassigned accuracy ϵ; on the other hand, the solution of a system $T_n(g)\mathbf{y} = \mathbf{c}$ can be obtained in $O(l^2 n)$ arithmetic operations (ops) by using a classic band solver [36]. Hence, balancing the cost of a single iteration of the PCG and the number of required iterations, it is possible to estimate the optimal bandwidth l, which allows to minimize the *total amount* of calculations to reach the solution of $T_n(f)\mathbf{x} = \mathbf{b}$ within a preassigned tolerance ϵ. Moreover by choosing $l = l(n)$ diverging function of n and $g \in \{g_{\rm opt}, g_A, g_B\}$, if f is smooth enough then

$$\lim_{n\to\infty} \mu^* = \lim_{n\to\infty} \mu^A = \lim_{n\to\infty} \mu^B = 1$$

and therefore for n large enough by equation (21) only one iteration is needed (the method becomes a direct one). We can say that such a kind on method is a *superlinearly* convergent PCG method becoming faster as n becomes larger. For the solution of the systems related to the preconditioner we can use a Golub band solver of cost $O(l^2 n)$ and therefore we have to choose $l(n) = O(\log^{1/2}(n))$; if we use a multigrid strategy (see [33]) of cost $O(ln)$ then we have to set $l(n) = O(\log(n))$: with these positions the total cost of $O(n\log(n))$ arithmetic ops.

Finally, in all the cases $g = g_{\rm min}, g = g_{\rm opt}, g = g_A, g = g_B$ we stress that we use the basic Algorithm 4.1 defined in the former Subsection 4.1.

Now we present some numerics in order to substantiate our claims. More specifically we compare the convergence rate of the band Toeplitz preconditioner (strategy **B**), with the optimal band Toeplitz preconditioner [23] and with the optimal circulant preconditioner [24] on three different generating functions having zeros. They are $(x-1)^2(x+1)^2$, $1-e^{-x^2}$ and x^4 and are associated to ill-conditioned matrices $T_n(f)$ having Euclidean condition numbers equal to $O(n^2)$, $O(n^2)$ and $O(n^4)$ respectively (see Corollary 2.1). The matrices $T_n(f)$ are formed by evaluating the Fourier coefficients of the generating functions by using FFTs (see [23]). In the considered tests, the vector of all ones is the right-hand side vector, the zero vector is the initial guess and the stopping criterion is $\|\mathbf{r}_q\|_2/\|\mathbf{r}_0\|_2 \leq 10^{-7}$, where $\mathbf{r}_q$ is the residual vector after q iterations. All computations are done by using MATLAB.

In the subsequent Tables 2, 3, and 4, I_n denotes that no preconditioning is used, $\Phi_{n,\mathcal{C}}$ is the circulant *Frobenius optimal preconditioner* (see [24] and Subsection 5.2), $B^*_{n,l}$ is the optimal band Toeplitz preconditioner [23] and $B^B_{n,l}$ is the band Toeplitz preconditioner defined according to the strategy **B**; here l denotes the half-bandwidth of the band preconditioners.

We do not make explicit comparison with the preconditioner related to the strategy **A** because, by virtue of equation (22), the associated PCG method has a convergence speed between the R. Chan, P. Tang one and the "**B**" one.

We observe that the "optimal" and the "**B**" band Toeplitz PCG methods perform, substantially, in the same way, but the second one is much more economical with respect to the computation of the related generating function. This fact is not so considerable when the bandwidth is fixed, but it becomes crucial in order to increase l, say, as $\log(n)$. Actually, in this case, for any dimension n, it is not expensive to calculate a different preconditioner $T_n(g^B(l))$, since the related cost $O(\log(n)\log(\log(n)))$ is strongly dominated by the cost $O(n\log(n))$ of each PCG iteration.
Finally, the reduction of the number of required iterations, as the dimension increases, shown in Table 5 gives a numerical evidence of the superlinear convergence when $l = l(n)$ is a (mildly) diverging function. We stress that the exceptional convergence behavior of the PCG algorithm related to $B^B_{n,6}$ is explained by the good approximation properties of the first-kind Chebyshev interpolation: to have a practical measure of this, it is sufficient to notice that the reduction of the condition number from $T_n(f)$ to $(B^B_{n,6})^{-1}T_n(f)$, for $n = 256$ and $f(x) = 1 - e^{-x^2}$, is from $2.7 * 10^4$ to $1 + 5 * 10^{-4}$.

Table 2: PCG, comparison with generating function $f(x) = (x^2 - 1)^2$

n	I	$\Phi_{n,\mathcal{C}}$	$B^*_{n,3} = B^B_{n,3}$	$B^*_{n,4}$ $B^B_{n,4}$	$B^*_{n,5}$ $B^B_{n,5}$	$B^*_{n,6}$ $B^B_{n,6}$
16	11	9	9	9 7	8 6	7 6
32	27	14	13	11 9	9 7	7 6
64	74	17	16	11 10	8 8	7 7
128	193	22	18	11 11	8 8	7 7
256	465	28	19	11 11	8 9	7 7
512	> 1000	34	19	11 11	8 8	7 7

Table 3: PCG, comparison with generating function $f(x) = 1 - e^{-x^2}$

n	I	$\Phi_{n,\mathcal{C}}$	$B^*_{n,2} = B^B_{n,2}$	$B^*_{n,3}$ $B^B_{n,3}$	$B^*_{n,4}$ $B^B_{n,4}$	$B^*_{n,5}$ $B^B_{n,5}$
16	9	6	9	7 8	4 4	3 3
32	14	7	15	7 8	5 5	3 3
64	24	8	17	8 9	5 5	3 3
128	42	10	17	8 9	5 5	3 3
256	77	13	17	8 9	5 5	3 3
512	143	17	17	8 9	5 5	3 3

Table 4: PCG, comparison with generating function $f(x) = x^4$

n	I	$\Phi_{n,\mathcal{C}}$	$B^*_{n,3} = B^B_{n,3}$	$B^*_{n,4}$ $B^B_{n,4}$	$B^*_{n,5}$ $B^B_{n,5}$	$B^*_{n,6}$ $B^B_{n,6}$
16	12	10	9	9 8	9 7	7 6
32	34	16	15	10 10	11 8	9 7
64	119	26	21	13 12	11 10	9 8
128	587	77	24	15 15	12 11	10 10
256	> 1000	179	27	16 16	12 13	10 10
512	> 1000	406	29	16 16	13 13	10 11

Table 5: Superlinear PCG with generating function $f(x) = 1 - e^{-x^2}$, Preconditioner= $B^B_{n,l(n)}$, $l(n) = \log_2(n) - 2$

n	16	32	64	128	256	512
$l(n)$	2	3	4	5	6	7
Iter	9	7	5	3	2	2

4.4 The case of f with (essentially) nondefinite sign

In some applications of Toeplitz matrices, we have observed that $T_n(f)$ is guaranteed to be positive definite, but in other fields such as eigenfilter problems, linear prediction theory, eigenvalue computation etc. the matrices may also be indefinite [25] and hence, in this subsection, we assume that the generating function f has (essentially) nondefinite sign. Following [64], the original nondefinite sysyem (potentially singular) is transformed into an equivalent positive definite (at least nonnegative definite) system whose coefficients matrix is somehow related to the Toeplitz structure: the convergence analysis is strongly related to the tools of Section 3. The method is outlined by the following steps:

Algorithm 4.4(f with nondefinite sign)

Input. The zeros of f, $T_n(f)$, $\mathbf{b}$.

Step 1. Find g such that $\mathcal{ER}\left(\frac{f}{g}\right)$ is contained in $[\alpha^-, \beta^-] \bigcup [\alpha^+, \beta^+]$ where $\alpha^- \leq \beta^- < 0 < \alpha^+ \leq \beta^+$; for instance, if we set $g = |f|$ then we have $\alpha^- = \beta^- = -1,\ \alpha^+ = \beta^+ = 1$.

Step 2. Compute the Toeplitz matrix $T_n(g)$, which is Hermitian and positive definite, and consider the equivalent system $G_n\mathbf{x} = \hat{\mathbf{b}}$ where $\hat{\mathbf{b}} = T_n^{-1}(g)\mathbf{b}$ and $G_n = T_n^{-1}(g)T_n(f)$. The vector $\hat{\mathbf{b}}$ can be calculated in $O(n\log(n))$ arithmetic ops and $O(\log n)$ parallel steps if we use the PCG procedures of the former two subsections and if f has only zeros of even order.

Step 3. Consider the new equivalent (spectrally) positive definite system (at least (spectrally) nonnegative definite) $G_n^2\mathbf{x} = \tilde{\mathbf{b}}$ where $\tilde{\mathbf{b}} = G_n\hat{\mathbf{b}}$ and solve it by the PCG method : here $T_n(g)$ is the preconditioner and $T_n(f)T_n^{-1}(g)T_n(f)$ is the new coefficient matrix.

Observe that G_n^2 is similar to $T_n^{-1/2}(g)T_n(f)T_n^{-1}(g)T_n(f)T_n^{-1/2}$ which is at least semidefinite positive (G_n is singular if and only if $T_n(f)$ is singular). By using the fact that the product between a Toeplitz matrix and a vector costs $O(n\log(n))$ ops ($O(\log(n))$ parallel steps with n processors in the PRAM model) it is easy to prove that **Step 2** costs $O(n\log(n))$ ops, provided that the entries of $T_n(g)$ can be computed within this time. Concerning **Step 3** we have that its cost is $O(N(\epsilon)n\log(n))$ ops where $N(\epsilon)$ is the number of iterations required by the PCG method to reach the solution within a preassigned accuracy. Thus the main goal is to evaluate $N(\epsilon)$. Now, by the second item of Theorem 3.1 we know that the union of the spectra of the matrices G_n is dense in $\mathcal{ER}\left(\frac{f}{g}\right)$ and, by the first item of the same theorem, it is contained in (α^-, β^+). Moreover in [64] it is shown that, in the case where f and g are rational symmetric functions, setting $c^- = \min\{\alpha^+, |\beta^-|\}$, $c^+ = \max\{|\alpha^-|, \beta^+\}$, there exists a constant q independent of n such that $\Sigma(G_n^2) \subset \{\lambda_1^{(n)}, \ldots, \lambda_q^{(n)}\} \bigcup [c^-, c^+]$, where $\lambda_1^{(n)}, \ldots, \lambda_q^{(n)} \in [0, c^-)$ and $\Sigma(X)$ denotes the set of the eigenvalues of X. In the general case we have $q = o(n)$, but for sufficiently regular nonrational functions f and g it can be proved that

the eigenvalues stay uniformly away from zero: this phenomenon known as gap phenomenon has been observed and proved in some special cases in [41]. Therefore by applying the result of [3] (see Theorem 8.3), the conjugate gradient method applied to the system $G_n^2\mathbf{x} = \hat{\mathbf{b}}$, converges to the solution with a preassigned accuracy ϵ in $N(\epsilon) = k^*(c, [c^+]^2, \epsilon) + q$ iterations where the function k^* is the one defined in Theorem 8.3 and c is any bound from below for the minimal eigenvalue of of G_n^2. Since ϵ is fixed if c is an absolute constant independent of n, then the desired precision is obtained through a constant number of iterations, an arithmetic cost of $O(n\log(n))$ and $O(\log(n))$ parallel steps. The evaluation of c is tricky but, as shown in the next examples, it seems that the minimal eigenvalues of G_n^2 stay away from zero (or approaches zero very slowly). Here we discuss some examples in order to substantiate the previous claims. Let

$$f(x) \equiv x = \sum_{k=1}^{\infty} \frac{\mathbf{i}(-1)^k}{k}(e^{\mathbf{i}kx} - e^{-\mathbf{i}kx}), \quad x \in I$$

and choose $T_n(g)$ generated by

$$g(x) \equiv |x| = \frac{\pi}{2} - \sum_{k=1}^{\infty} \frac{1-(-1)^k}{\pi k^2}(e^{\mathbf{i}kx} + e^{-\mathbf{i}kx}), \quad x \in I.$$

According to the Theorem 3.1, we expect that the eigenvalues of $G_n = T_n^{-1}(g)T_n(f)$ form two clusters around -1 and 1 since $f/g =$sign(x): for $n = 16$ we have

$$\Sigma(G_n) = \{\pm 1.000 \text{ (4 times)}, \pm 0.9997, \pm 0.9946, \pm 0.9287, \pm 0.4773\}. \tag{23}$$

For $n = 64$ we find

$$\begin{aligned} \Sigma(G_n) \quad = \quad & \{\pm 1.000 \text{ (27 times)}, \pm 0.9995, \pm 0.9963, \pm 0.9737, \\ & \pm 0.8470, \pm 0.3830\}. \end{aligned} \tag{24}$$

As a second example, let us consider

$$\begin{aligned} f(x) \quad \equiv \quad & \text{sign}(x)x^2 = \textstyle\sum_{k=1}^{\infty} \frac{\mathbf{i}}{\pi k^2}\left((-1)^k\pi^2 + \frac{2}{k^2}(1+(-1)^{(k+1)}\right) \\ & (e^{\mathbf{i}kx} - e^{-\mathbf{i}kx}), \quad x \in I; \end{aligned}$$

we propose two different functions g_1 and g_2:

$$g_1(x) \equiv x^2 = \frac{\pi^2}{3} + 2\sum_{k=1}^{\infty} \frac{(-1)^k}{k^2}(e^{\mathbf{i}kx} + e^{-\mathbf{i}kx}), \quad x \in I,$$

$$g_2(x) = 2 - 2\cos(x), \quad x \in I.$$

According to the results of the previous section we expect that $\Sigma(G_n) = \Sigma(T_n^{-1}(g_1)T_n(f))$ forms two clusters around -1 and 1 since $f/g_1 =$sign(x).
For $n = 16$ we have

$$\Sigma(G_n) = \{\pm 1.000 \text{ (4 times)}, \pm 0.9998, \pm 0.9961, \pm 0.9412, \pm 0.500\}. \tag{25}$$

For $n = 64$ we have

$$\begin{aligned} \Sigma(G_n) \quad = \quad & \{\pm 1.000 \text{ (27 times)}, \pm 0.9997, \pm 0.9972, \pm 0.9787, \\ & \pm 0.8640, \pm 0.4002\}. \end{aligned} \tag{26}$$

In the case of $G_n = T_n^{-1}(g_2)T_n(f)$, from Theorem 3.1, we expect that most of the eigenvalues belong to $\mathcal{ER}(f/g_2) = [-\pi^2/4, -1] \bigcup [1, \pi^2/4]$. For $n = 16$ we obtain

$$\begin{array}{ll} 14 & \text{eigenvalues in } [-\pi^2/4, -1] \bigcup [1, \pi^2/4], \\ 2 & \text{eigenvalues} = \pm 0.7078 \text{ in } (-1, 1). \end{array}$$

For $n = 64$ we have

$$\begin{array}{ll} 60 & \text{eigenvalues in } [-\pi^2/4, -1] \bigcup [1, \pi^2/4], \\ 2 & \text{eigenvalues} = \pm 0.9938 \text{ in a small neighbourhood} \\ & \text{of } -1 \text{ and } 1 \text{ respectively, and} \\ 2 & \text{eigenvalues} = \pm 0.5698 \text{ in } (-1, 1). \end{array}$$

Furthermore let

$$\begin{array}{rcl} f(x) \equiv e^x - 1 & = & \sum_{k=-\infty}^{\infty} \frac{(-1)^k (e^{\pi} - e^{-\pi})}{2\pi(1+k^2)} \\ & & (1 + \mathbf{i}k) e^{\mathbf{i}kx} - 1, \quad x \in I; \end{array}$$

and

$$\begin{array}{rcl} g(x) \equiv |e^x - 1| & = & \sum_{k=-\infty}^{\infty} t_k + \frac{1+\mathbf{i}k}{2\pi(1+k^2)} \\ & & ((e^{\pi} - e^{-\pi})(-1)^k - 2) e^{\mathbf{i}kx}, \quad x \in I, \end{array}$$

where t_k is $\dfrac{2\mathbf{i}}{\pi k}$ if k is odd and 0 elsewhere.
According to the Theorem 3.1 we expect two clusters around -1 and 1 for the spectrum of $G_n = T_n^{-1}(g)T_n(f)$: for $n = 16$ we obtain

$$\begin{array}{rcl} \Sigma(G_n) & = & \{\pm 1.000 \text{ (4 times)}, 0.9950, -0.9987, 0.9741, \\ & & -0.9740, 0.6755, -0.6842, 0.3395, -0.3384\}. \end{array} \tag{27}$$

For $n = 64$ we find

$$\begin{array}{rcl} \Sigma(G_n) & = & \{1.000 \text{ (27 times)}, -1 \text{ (26 times)}, 0.9999, \\ & & -0.9998, 0.9993, -0.9978, 0.9950, -0.9824, \\ & & 0.9710, -0.8764, 0.4212, -0.4076\}. \end{array} \tag{28}$$

Now we consider a function with a higher order zero: let

$$f(x) \equiv x^3 \quad = \quad \sum_{k=1}^{\infty} \frac{\mathbf{i}(-1)^k}{k} \left(\pi^2 - \frac{6}{k^2}\right) (e^{\mathbf{i}kx} - e^{-\mathbf{i}kx})$$

and

$$g(x) \equiv |x|(2 - 2\cos(x)) \quad = \quad \sum_{k=-\infty}^{\infty} c_k e^{\mathbf{i}kx},$$

where $c_0 = \pi - 2a_1(|x|)$, $c_j = 2a_j(|x|) - a_{j-1}(|x|) - a_{j+1}(|x|) = c_{-j}$ and $a_j(|x|)$ are the Fourier coefficients of the function $|x|$ shown in the first example.
According to the theoretical results, we expect that most of the eigenvalues $G_n = T_n^{-1}(g)T_n(f)$ belong to $\mathcal{ER}(f/g) = [-\pi^2/4, -1] \bigcup [1, \pi^2/4]$ which is the closure of the image of f/g. For $n = 16$ we obtain

$$\begin{array}{ll} 14 & \text{eigenvalues in } [-\pi^2/4, -1] \bigcup [1, \pi^2/4], \\ 2 & \text{eigenvalues} = \pm 0.7409 \text{ in } (-1, 1). \end{array}$$

For $n = 64$ we have

$$\begin{array}{ll} 60 & \text{eigenvalues in } [-\pi^2/4, -1] \bigcup [1, \pi^2/4], \\ 2 & \text{eigenvalues} = \pm 0.9980 \text{ in a small neighbourhood} \\ & \text{of } -1 \text{ and } 1 \text{ respectively, and} \\ 2 & \text{eigenvalues} = \pm 0.5937 \text{ in } (-1, 1). \end{array}$$

It is very interesting to remark that f/g_2 in the second example coincides with f/g in this example and, as a consequence, we have that the behavior of the spectra of the related matrices G_n, $n = 16, 64$, are practically the same.

Let

$$\begin{array}{rcl} f(x) & \equiv & x\left(x - \frac{\pi}{2}\right) = \sum_{k=-\infty}^{\infty} a_k(f) e^{\mathbf{i}kx}, \\ a_k(f) & = & a_k(x^2) - \frac{\pi}{2} a_k(x), \end{array}$$

be a function having two zeros in the fundamental interval I. Let

$$g(x) \equiv \left|x\left(x - \tfrac{\pi}{2}\right)\right| \quad = \quad \sum_{k=-\infty}^{\infty} c_k e^{\mathbf{i}kx},$$

where the values c_k are obtained by the Fourier coefficients of f in the following way:

$$\begin{array}{rcl} c_0 & = & \frac{17\pi^2}{48}, \\ r_k & = & \frac{\pi}{2k}\mathbf{i}e^{-\mathbf{i}k\pi/2} + \frac{1}{k^2}\left(e^{-\mathbf{i}k\pi/2} - 1\right), \ k \neq 0, \\ s_k & = & \frac{\pi^2}{4k}\mathbf{i}e^{-\mathbf{i}k\pi/2} - \frac{2}{k}\mathbf{i}r_k, \ k \neq 0, \\ c_k & = & a_k(f) - \frac{1}{\pi}s_k + \frac{1}{2}r_k, \ k \neq 0. \end{array}$$

According to the Theorem 3.1, we expect two clusters around -1 and 1 for the spectrum of $G_n = T_n^{-1}(g)T_n(f)$: for $n = 16$ we have

$$\begin{array}{rcl} \Sigma(G_n) & = & \{1.000 \text{ (3 times)}, 0.9999 \text{ (6 times)}, 0.9993, 0.9863, \\ & & 0.7947, -0.9999, -0.9967, -0.9244, -0.2314\}. \end{array} \tag{29}$$

For $n = 64$ we find

$$\begin{array}{rcl} \Sigma(G_n) & = & \{1.000 \text{ (35 times)}, -1 \text{ (4 times)}, \\ & & 0.9999 \text{ (9 times)}, -0.9999 \text{ (7 times)}, 0.9992, \\ & & -0.9997, 0.9933, -0.9972, 0.9467, -0.9750, \\ & & 0.6624, -0.8199, -0.1723\}. \end{array} \tag{30}$$

We remark that the interval where $f/g = -1$ is small with respect to I and, in fact, the number of the negative eigenvalues of G_n close to -1 is less than the number of those close to 1: this is a numerical evidence of the result displayed in item **d** of Theorem 3.1. Moreover, it is worth pointing out that the presence of two zeros causes a partial deterioration of the clustering property of the spectrum of G_n with respect to the case where the generating function f has a unique zero.

In the cases (23)–(24), (25)–(26),(27)–(28), f/g is the function sign(x) (because $g(x) = |f(x)|$ and $f(0) = 0$) and it is interesting to compare these spectra with the spectrum of $T_n(\text{sign}(x))$; the similarities are very deep: for $n = 16$ we have

$$\Sigma(T_n(\text{sign}(x))) = \{\pm 1.000 \text{ (4 times)}, \pm 0.9995, \pm 0.9913, \pm 0.9013, \pm 0.4294\}.$$

For $n = 64$ we have

$$\begin{array}{rcl} \Sigma(T_n(\text{sign}(x))) & = & \{\pm 1.000 \text{ (26 times)}, \pm 0.9999, \pm 0.9993, \pm 0.9945, \\ & & \pm 0.9636, \pm 0.8122, \pm 0.3487\}. \end{array}$$

4.5 The case of f with zeros of odd order

In this subsection we deal with the case of zeros of odd order (they occur frequently when f is nondefinite), in the special case where $T_n(f)$ is a symmetric real matrix, i.e., $f(x) = f(-x)$. We use a linear algebra trick to transform the system $T_n(f)\mathbf{x} = \mathbf{b}$ into a new system in which we have to solve the linear system $T_n(fg)\mathbf{x} = \mathbf{c}$ where fg has only zeros of even orders. Moreover, if $f \geq 0$ then fg has nondefinite sign, otherwise if f is nondefinite then fg is nonnegative. Therefore, in the first case, we use Algorithm 4.4, otherwise Algorithm 4.1. Finally, since the costs of this transformation have been proved to be bounded by $O(n\log(n))$ arithmetic ops and $O(\log(n))$ parallel steps (see [62]), then the total cost is the one between Algorithm 4.1 and Algorithm 4.4. The proposed algorithm is outlined in the following steps:

Algorithm 4.5(f symmetric having a unique zero of odd order)

Input. $f,\ T_n(f),\ \mathbf{b}$ where f is a function having, for the sake of simplicity, a unique zero of odd order $2k+1$.

Step 1. Find a polynomial g of degree 1 with the same zero (of first order) of f; $T_n(g)$ is a symmetric tridiagonal matrix whose eigenvalues and eigenvectors are known ($T_n(g)$ belongs to the algebra [10] of all the symmetric matrices simultaneously diagonalized by the discrete sine transform DST I).

Step 2. If $T_n(g)$ is singular we take the matrix $E = \mathbf{v}_i\mathbf{v}_i^T,\ \epsilon > 0$ and we define the nonsingular matrix $\tilde{A}_n(g) = T_n(g) + E$, $\mathbf{v}_i$ being the unitary eigenvector of $T_n(g)$ associated with the null eigenvalue; we set $\tilde{A}_n(g) = T_n(g)$ in the case where $T_n(g)$ is nonsingular.

Step 3. Consider the equivalent systems $\tilde{A}_n(g)T_n(f)\mathbf{x} = \tilde{\mathbf{b}};\ T = \tilde{A}_n(g)T_n(f)$ where $\tilde{\mathbf{b}} = \tilde{A}_n(g)\mathbf{b}$. Observe that $\tilde{A}_n(g)T_n(f) = T_n(fg) + L$ where L is a low rank correction matrix (in particular $\mathrm{rank}(L) = 2$ if $T_n(g)$ is nonsingular and $\mathrm{rank}(L) = 3$ if $T_n(g)$ is singular).

Step 4. Solve the former system by the Sherman–Morrison–Woodbury (see e.g. [36]) formula which involves the solution of some systems where the coefficients matrix is $T_n(fg)$, $|fg| \geq 0$ and fg has a zero of even order $2(k+1)$. Therefore, if f is nonnegative then fg is nondefinite and we can apply the PCG Algorithm 4.4 proposed in [64], otherwise if f is nondefinite then fg is nonnegative and we apply Algorithm 4.1. In both the cases we observe a total arithmetic cost of $O(n\log(n))$ and $O(\log(n))$ parallel steps.

4.6 The case of f with zeros of any order

If we have to solve the linear system $T_n(f)\mathbf{x} = \mathbf{b}$ where f is nonnegative and has, for the sake of simplicity, a unique zero x_0 of order $\rho > 0$, then, in Algorithm 4.1, we choose $T_n((x-x_0)^{2k})$ as preconditioner, $2k$ being the even number which minimizes the distance from ρ. In the light of Proposition 3.1, the condition number of $T_n^{-1/2}((x-x_0)^{2k})T_n(f)T_n^{-1/2}((x-x_0)^{2k})$ grows as $n^{|\rho-2k|}$ and therefore, by (21), the related PCG method requires a number of steps $N(\epsilon)$ proportional to $\log(1/\epsilon)\cdot n^{\frac{|\rho-2k|}{2}}$), i.e., in the worst case of ρ odd number, $O(n^{1/2})$ PCG steps. Recalling that the solution of a system with coefficient matrix $T_n((x-x_0)^{2k})$ costs $O(n\log(n))$ arithmetic ops and $O(\log(n))$ parallel steps if we use Algorithm 4.1, we can conclude a general statement. If we want to solve a linear system of the form $T_n(f)\mathbf{x} = \mathbf{b}$ where f has a zero of order ρ then the preconditioning by means of $T_n((x-x_0)^{2k})$ ($2k$ being the even number which minimizes the distance from ρ) produces

a PCG method having a total cost of $O\left(n^{\frac{|2k-\rho|}{2}} n\log(n)\right)$ arithmetic ops and $O\left(n^{\frac{|2k-\rho|}{2}} \log(n)\right)$ parallel steps. Finally this idea can be used in Algorithm 4.4, for the solution of the preconditioning systems, in order to deal successfully with the case of zeros of any order and f with nondefinite sign.

We discuss some numerical tests. The aim of these experiments is twofold: we want to show, by numerical evidence, the correctness of the asymptotical spectral analysis about the preconditioned matrices performed in Section 3 and we want to verify the good features of the PCG methods related to this kind of preconditioners. Actually, as shown in the subsequent tables, the theoretical prediction are fully confirmed. In addition, we point out that, very recently, this technique has been proved to be "robust" in the sense that an approximate knowledge of the position of the points where f vanishes is enough to define a good preconditioner [72].

In the first example we deal with a linear system $T_n(f)\mathbf{x} = \mathbf{b}$ where the generating function is $f(x) = x^2|x|^{\frac{1}{10}}$; therefore, as $\rho = 2.1$ we choose $g(x) = 2 - 2\cos(x)$ which is associated to the tridiagonal matrix $T_n(g)$ and has a zero of order $2k = 2$. In the light of Theorem 3.1, the condition number of the preconditioned matrix is asymptotical to $n^{\frac{1}{10}}$ and, consequently, by virtue of the powerful convergence analysis of Axelsson and Lindskög [3], we expect that the number of iterations $N(\epsilon)$ of the PCG method, in order to reach the solution within the fixed accuracy ϵ, is proportional to $n^{\frac{1}{20}}$. Since $n^{\frac{1}{20}}$ grows very slow, we expect that $N(\epsilon)$ is "weakly" depending on n and is "practically" constant. Actually, as shown in Table 6, for $n = 128, 256, 512$, we find $N(\epsilon) = 20$, 22, 22 which confirms the results of the theoretical analysis.

In this second case the generating function is $f(x) = x^2|x|^{\frac{1}{2}}$ with $\rho = 2.5$ and $g(x) = 2-2\cos(x)$ with $2k = 2$. According to Theorem 3.1, the condition number of the preconditioned matrix is asymptotical to $n^{\frac{1}{2}}$ and therefore the number of iterations $N(\epsilon)$ of the PCG method, in order to reach the solution within the fixed accuracy ϵ, is proportional to $n^{\frac{1}{4}}$. As shown in Table 7, for $n = 128, 256, 512$, we find $N(\epsilon) = 29, 35, 40$; we observe the perfect agreement with the theoretical analysis: in fact, the quantity $N(\epsilon)$ varies practically as $n^{\frac{1}{4}}$ since $29 \cdot 2^{\frac{1}{4}} \approx 35$ and $35 \cdot 2^{\frac{1}{4}} \approx 41$ which is close to 40.

In the third example the generating function is $f(x) = x^2|x|^{\frac{1}{3}}$ with $\rho = 2.3$ and $g(x) = 2 - 2\cos(x)$ with $2k = 2$. By Theorem 3.1, the condition number of the preconditioned matrix is asymptotical to $n^{\frac{1}{3}}$ and we expect that the number of iterations $N(\epsilon)$ of the PCG method, in order to reach the solution within the fixed accuracy ϵ, is proportional to $n^{\frac{1}{6}}$. As shown in Table 8, for $n = 128, 256, 512$, we find $N(\epsilon) = 22, 25, 28$; we notice again the perfect agreement with the theoretical analysis: in fact, the quantity $N(\epsilon)$ grows practically as $n^{\frac{1}{6}}$ since $22 \cdot 2^{\frac{1}{6}} \approx 25$ and $25 \cdot 2^{\frac{1}{6}} \approx 27$ which is close to 28.

In the last case the generating function is $f(x) = x^4|x|^{\frac{1}{12}}$ with $\rho = 4.08$ and $g(x) = 2 - 2\cos(x)$ with $2k = 4$. In the light of Theorem 3.1, the condition number of the preconditioned matrix is asymptotical to $n^{\frac{1}{12}}$ and as a consequence the number of iterations $N(\epsilon)$ of the PCG method, in order to reach the solution within the fixed accuracy ϵ, is proportional to $n^{\frac{1}{24}}$. We recall that the original condition number grows asymptotically more than n^4 (compare Corollary 2.1 and Theorem 2.6) and this is confirmed by the dramatic slowness of the CG method (without preconditioning). In this case the acceleration due to the preconditioning technique is really evident: as shown in Table 9, for $n = 128, 256, 512$, we find $N(\epsilon) = 21, 22, 22$ so that we conclude a substantial independence with respect to the dimension n.

Finally we stress that all the computations have been done in MATLAB by using the zero vector as starting point and the vector of all ones as vector $\mathbf{b}$. In all the considered cases the stopping criterion

has been $\frac{\|\mathbf{r}_q\|_2}{\|\mathbf{r}_0\|_2} < 10^{-7}$ where $\mathbf{r}_j$ is the residual vector at the j-th iteration. Moreover, in all the tables the numbers in the first row are the dimensions of the linear systems, the numbers appearing in the second row are the numbers of iterations without preconditioning (preconditioner $= I_n$) and those appearing in the third row are the numbers of iterations when the preconditioner $T_n(g)$ is used.

Table 6: PCG, $f(x) = x^2|x|^{\frac{1}{10}}$, $\rho = 2.1$, $2k = 2$, Preconditioner $= T_n(g)$, $g(x) = 2 - 2\cos(x)$

n	128	256	512
I_n	86	187	397
$T_n(g)$	20	22	22

Table 7: PCG, $f(x) = x^2\sqrt{|x|}$, $\rho = 2.5$, $2k = 2$, Preconditioner $= T_n(g)$, $g(x) = 2 - 2\cos(x)$

n	128	256	512
I_n	110	266	643
$T_n(g)$	29	35	40

Table 8: PCG, $f(x) = x^2|x|^{\frac{1}{3}}$, $\rho = 2 + 1/3$, $2k = 2$, Preconditioner $= T_n(g)$, $g(x) = 2 - 2\cos(x)$

n	128	256	512
I_n	99	227	516
$T_n(g)$	22	25	28

4.7 Further results and generalization to indefinite, non Hermitian problems

The main message of the band Toeplitz approach is that we have to choose a trigonometric polynomial g which matches the zeros (with the right order) of the symbol f of the original coefficient matrix $T_n(f)$. The only constraint given in Theorem 3.1 is that g has to be nonnegative and not identically zero in order to ensure the positive definiteness of the preconditioner $T_n(g)$. Here we relax this assumption by allowing g to be indefinite (when f is indefinite) and even not real valued (when f is not real valued). The theory is more involved and not completely developed: for precise statements we refer to [44, 43]. Here in Subsections 4.7.1 and 4.7.2 we just discuss some numerical evidences to give an idea of the results. Finally, in Subsection 4.7.3, we show the robustness of the band Toeplitz approach. Indeed, if the zeros of f are not analytically known, it is possible to recover them with their order (see [72]) by using cheap numerical procedures: only a low precision is required in order to preserve the optimality of the band Toeplitz preconditioning [79].

4.7.1 Indefinite preconditioning for indefinite problems

Since the preconditioner and the coefficient matrices are essentially indefinite, we perform our tests by using both the PCG (for which there is no theoretical guarantee that a break down it is impossible) and the preconditioned GMRES [59]. However, as the numerics show, the PCG is effective and even

Table 9: PCG, $f(x) = x^4|x|^{\frac{1}{12}}$, $\rho = 4 + 1/12$, $2k = 4$, Preconditioner $= T_n(g)$, $g(x) = (2 - 2\cos(x))^2$

n	128	256	512
I_n	383	1562	4663
$T_n(g)$	21	22	22

Table 10: P[CG/GMRES], $f(x) = (0.1x^2(2 - 2\cos(2x) + 1)\sin(x) + 0.2(2 - 2\cos(2x))$, $g(x) = \sin(x)$, $\mu_{\text{sp}} = \lambda_{\max}/\lambda_{\min}$

size=n	$\lambda_{\min}$	$\lambda_{\max}$	μ_{sp}	#(it)	#(outliers)
16	0.5146	2.2336	4.3397	13/13	1
32	0.5146	2.2489	4.3696	15/15	1
64	0.5146	2.2526	4.3767	16/15	1
128	0.5146	2.2540	4.3794	16/15	1
256	0.5146	2.2543	4.3801	16/15	1
512	0.5146	2.2544	4.3803	16/15	1

optimally convergent.

Test 1 We consider the family of indefinite Toeplitz whose generating function is

$$f(x) = (\gamma_1 + \gamma_2 x^2(2 - 2\cos(2x))\sin(x) + \gamma_3(2 - 2\cos(2x)),$$

with indefinite preconditioner generated by $g(x) = \sin(x)$. We can choose the real parameters γ_1, γ_2, and γ_3 in such a way that f/g is strictly positive and bounded (γ_1 positive and large enough) and $f - g$ has nonnegative sign ($\gamma_3 - \gamma_2$ positive and large enough): the first choice ensures that the zeros of f are matched while the second ensures that the eigenvalues of the preconditioned matrices $G_n = T_n^{-1}(g)T_n(f)$ are real (see [44]). In particular, we expect that most of the eigenvalues of G_n belong to a positive bounded interval since f/g is positive and bounded.

For instance with $\gamma_1 = 1$, $\gamma_2 = 0.1$, $\gamma_3 = 0.2$ we deduce that the range of f/g is an interval whose extremes are the following

$$\inf f/g = 0.63, \ \sup f/g = 2.29.$$

In actuality, the numerics in Table 10 show that all the eigenvalues of the preconditioned matrix are real, they strictly belong to the interval $[0.63, 2.29]$ with one very stable outlier equal to 0.5146. Moreover both the preconditioned GMRES and CG are optimal in the sense that the number of iterations does not increase with the size of the algebraic problem.

Test 2 We consider the indefinite Toeplitz problem with indefinite preconditioner whose generating functions are

$$f(x) = (x^2 + 5)\sin(x), \quad g(x) = \sin(x),$$

with

$$\inf f/g = 5, \ \sup f/g = 14.8696.$$

The numerical results displayed in Table 11 show that all the eigenvalues of the preconditioned matrix are real, they strictly belong to the interval $[5, 14.8696]$ with no outliers. Moreover both the

Table 11: P[CG/GMRES], $f(x) = (x^2 + 5)\sin(x)$, $g(x) = \sin(x)$, $\mu_{\rm sp} = \lambda_{\max}/\lambda_{\min}$

size=n	$\lambda_{\min}$	$\lambda_{\max}$	$\mu_{\rm sp}$	#(it)	#(outliers)
16	5.1252	13.0190	2.5402	8/8	0
32	5.0346	13.8726	2.7554	12/12	0
64	5.0091	14.3511	2.8650	15/12	0
128	5.0023	14.6050	2.9196	13/13	0
256	5.0006	14.7359	2.9468	13/13	0
512	5.0002	14.8024	2.9604	14/13	0

preconditioned GMRES and CG are optimal in the sense that the number of iterations is practically constant with regard to the size of the algebraic problem.

4.7.2 Non Hermitian preconditioning for non Hermitian problems

In the following we consider 4 basic examples in which we cover different situations. The common point is that the generating functions have (essential) zeros so that the related Toeplitz sequences $\{T_n(f)\}$ have asymptotical unbounded inverses and consequently the classic iterative solvers, when convergent, are all sublinear (i.e. they require a number of iterations exploding to infinity as n tends to infinity). The iterative solver is the preconditioned GMRES with band Toeplitz preconditioning.
Test 1 We consider the non Hermitian Toeplitz problem with non Hermitian preconditioner whose generating functions are

$$f(x) = (1 - e^{\mathbf{i}x})(5 + x^2), \quad g(x) = 1 - e^{\mathbf{i}x}.$$

We use a non Hermitian band preconditioning for a dense non Hermitian problem where f/g is real valued, bounded and strictly positive. We observe a very favorable picture (see the first part of Table 12) since all the eigenvalues are real and belong to the the interior of the range of the function f/g (as predicted for the real eigenvalues by a theorem in [43]). We recall that the interior of the range of the function f/g is $(5, 5 + \pi^2) = (5, 14.8696)$ in very good agreement with the numbers reported in the first two columns of Table 12. Furthermore, the number of iterations is bounded by an absolute constant independent of n and this is a confirmation of the optimality of the proposed technique.
Test 2 We consider the non Hermitian Toeplitz problem with positive definite preconditioner whose generating functions are

$$f(x) = (2 - 2\cos(x))(1 + \mathbf{i}x), \quad g(x) = 2 - 2\cos(x).$$

We use a positive definite band preconditioning for a dense non Hermitian problem having real part coinciding with the preconditioner. Therefore

$$T_n^{-1}(g)T_n(f) = I_n + \mathbf{i}T_n^{-1}(g)T_n(g(x)x).$$

Thus, all the eigenvalues have real part equal to one. Moreover the imaginary part of the eigenvalues is distributed as the function x and is localized in $(-\pi, \pi)$ since $-\pi = \inf_I x$ and $\pi = \sup_I x$ (see items **a** and **d** of Theorem 3.1). All the theoretical forecasts are fully honored as displayed in the second part of Table 12: moreover the proposed preconditioning is optimal as in the first case since we observe a stabilization to a fixed number of the related iteration count.

Test 3 We consider once a time the non Hermitian Toeplitz problem with positive definite preconditioner whose generating functions are

$$f(x) = x^2(1 + \mathbf{i}x), \quad g(x) = 2 - 2\cos(x).$$

We use a positive definite band preconditioning for a dense non Hermitian problem having positive definite real part which is spectrally equivalent to the preconditioner since $2 - 2\cos(x) \sim x^2$. Therefore

$$T_n^{-1}(g)T_n(f) = T_n^{-1}(g)T_n(x^2) + \mathbf{i}T_n^{-1}(g)T_n(x^3).$$

Thus, from Theorem 3.1, we know that the all the eigenvalues have real part belonging to the interval $(1, \pi^2/4)$, $\pi^2/4 = 2.4674$, while the imaginary part of the eigenvalues belongs $(-\pi^3/4, \pi^3/4)$, $\pi^3/4 = 7.7516$, since $-\pi^3/4 = \inf_I x^3/(2 - 2\cos(x))$ and $\pi^3/4 = \sup_I x^3/(2 - 2\cos(x))$. These theoretical expectations are confirmed in the part of Table 12 where we also notice the optimality of the proposed preconditioned GMRES method.

Test 4 We consider a non Hermitian Toeplitz problem with indefinite preconditioner whose generating functions are

$$f(x) = \sin(x)(1 + \mathbf{i}(x^2 + 1)), \quad g(x) = \sin(x).$$

We use an indefinite band preconditioning for a dense non Hermitian problem having indefinite real part coinciding with the preconditioner. Therefore $T_n^{-1}(g)T_n(f) = I_n + \mathbf{i}T_n^{-1}(g)T_n(g(x)(x^2 + 1))$. Thus, by the results in [44], we have all the eigenvalues with real part equal to 1 and imaginary part distributed as the function $x^2 + 1$ ($\inf_I x^2 + 1 = 1$ and $\sup_I x^2 + 1 = \pi^2 + 1 = 10.8696$). In this case theory on indefinite preconditioning is the basic tool for devising good preconditioning strategies for non Hermitian problems whose real part is indefinite. Indeed, in this last example, the main interest is that the indefinite preconditioning is effective for the given non Hermitian problem as we can observe in the last part of Table 12.

4.7.3 Robustness of the band Toeplitz preconditioning

Here we just report one that shows the robustness of the band Toeplitz approach. We consider the Toeplitz problem with symbol $f(x) = (x^2 - 1)^2$: the exact band Toeplitz preconditioner has generating function $g(x) = (2\cos(1) - 2\cos(x))^2$. We assume that the zeros of f, which clearly coincide with ± 1, are not known exactly and we define an approximated preconditioner with generating function $\tilde{g}(x) = (2\cos(0.994) - 2\cos(x))^2$ whose zeros are ± 0.994. As reported in Table 13, the performances of the approximate preconditioner are similar to those of the exact one and are both much better with respect to the circulant *Frobenius optimal preconditioners* $\Phi_{n,\mathcal{C}}$ (see [24] and Subsection 5.2).

More generally, if the zeros of f are not analytically known, it is possible to recover them with their order (see [72]) by using cheap numerical procedures: only a low precision (bounded by n^{-1}) is required in order to preserve the optimality of the band Toeplitz preconditioning (see [72, 79]).

5 Matrix algebra preconditioning

We introduce some class of important (unitary) matrix algebras whose unitary transforms have a cost proportional to $n \log(n)$. The idea is to find preconditioners in these spaces: the convergence analysis is carried via the Korovkin theory (Subsection 5.1) and is postponed to Subsection 5.2.

Table 12: P[GMRES], the four tests in the non Hermitian case

size=n	$\min\{\mathrm{Re}(\lambda)\}$	$\max\{\mathrm{Re}(\lambda)\}$	$\min\{\mathrm{Im}(\lambda)\}$	$\max\{\mathrm{Im}(\lambda)\}$	#(it)
Test 1					
16	5.1252	14.0246	0	0	8
32	5.0346	14.4332	0	0	12
64	5.0091	14.4646	0	0	13
128	5.0023	14.4757	0	0	13
256	5.0006	14.4813	0	0	13
512	5.0001	14.4841	0	0	13
Test 2					
16	1	1	-2.7174	2.7174	16
32	1	1	-2.9022	2.9022	32
64	1	1	-3.0099	3.0099	46
128	1	1	-3.0702	3.0702	50
256	1	1	-3.1034	3.1034	52
512	1	1	-3.1213	3.1213	52
Test 3					
16	1.1890	2.0133	-2.6857	2.6857	16
32	1.0464	2.1829	-2.8826	2.8826	32
64	1.0243	2.2998	-2.9986	2.9986	41
128	1.0128	2.3726	-3.0643	3.0643	45
256	1.0067	2.4152	-3.1003	3.1003	46
512	1.0034	2.4391	-3.1197	3.1197	47
Test 4					
16	1	1	1.1252	9.0192	8
32	1	1	1.0346	9.8726	16
64	1	1	1.0091	10.3510	22
128	1	1	1.0023	10.6049	24
256	1	1	1.0006	10.7358	24
512	1	1	1.0001	10.8023	24

In general, given the class α_n of matrices arising from a given problem, when dealing with the PCG method we have to face two challenges:

A. choose a suitable class β_n of matrices "close" enough to α_n and whose elements are easy to invert.

B. devise a suitable projection operator $\mathcal{P}_n : \alpha_n \to \beta_n$ to obtain the best approximation $P_n \in \beta_n$ for any given $A_n \in \alpha_n$.

One of the promising ways to meet point **A** is to look for preconditioners within matrix algebras like the circulant class [27] and other special algebras of matrices [10, 12, 47]. This typically offers a better knowledge of how close we may choose the approximation and the possibility to use a uniform, and often efficient, algorithm to solve the preconditioned system.

When dealing with real problems point **B** becomes delicate since we need to reduce the "informative content" of the original matrix in order to have cheap invertibility in the approximation space β_n. In these cases some averaging schemes (see (40), (41), and e.g. [19]) have proved to be useful: they are related the the so-called Frobenius optimal approximation [24] that we describe in Subsection 5.2.

Here we focus our attention on the spaces α_n which are of interest for us: given U_n unitary matrix (i.e. $U_n^H U_n = I_n$), we consider the associated (unitary) algebra of matrices defined as

$$\mathcal{A}_n = \{X = U_n D U_n^H : \ D \text{ diagonal matrix}\}. \tag{31}$$

Table 13: PCG, $f(x) = (x^2-1)^2$, $g(x) = (2\cos(1)-2\cos(x))^2$, $\tilde{g}(x) = (2\cos(0.994)-2\cos(x))^2$: #(iterations) for different preconditioners

n	128	256	512
I_n	182	449	> 1000
$\Phi_{n,\mathcal{C}}$	22	28	34
$T_n(g)$	18	19	19
$T_n(\tilde{g})$	19	20	22

The diagonal matrix in the above definition has entries in the same field where the entries of U_n are defined: thus we have real algebras of symmetric matrices and complex algebras of matrices. A classical example of (unitary) complex matrix algebra is the algebra of circulants which has been introduced in Subsection 4.1. The transform U_n is the Fourier matrix F_n.

Starting from circulants many other circulant-like algebras can be defined; this is the case of the Hartley class [12]. Here we report the transforms U_n both for circulants and Hartley matrices:

$$\begin{aligned} U_n = F_n &= \left(\frac{1}{\sqrt{n}} e^{\mathbf{i} j x_k^{(n)}}\right), \quad j,k = 0,\ldots,n-1, \\ &\quad W_n = \left\{x_k^{(n)} = \tfrac{2k\pi}{n} : k = 0,\ldots,n-1\right\} \subset [0,2\pi], \\ U_n = H_n &= \left(\tfrac{1}{\sqrt{n}}\left[\sin(j x_k^{(n)}) + \cos(j x_k^{(n)})\right]\right), \quad j,k = 0,\ldots,n-1, \\ &\quad W_n = \left\{x_k^{(n)} = \tfrac{2k\pi}{n} : k = 0,\ldots,n-1\right\} \subset [0,2\pi]. \end{aligned}$$

The set W_n is called the set of grid points and plays an important role in the study of preconditioners in these algebras. For the class of the ω-circulants [27, 9] we may consider value ω on the unit complex. More precisely, if $\omega = e^{\mathbf{i}2\pi\psi}$ then the matrix U_n has the following representation

$$\begin{aligned} U_n = F_{n,\omega} &= \left(\tfrac{1}{\sqrt{n}} e^{\mathbf{i}(j+\psi) x_k^{(n)}}\right), \quad j,k = 0,\ldots,n-1, \\ F_n = F_{n,1}, &\quad W_n = \left\{x_k^{(n)} = \tfrac{2k\pi}{n} : k = 0,\ldots,n-1\right\} \subset [0,2\pi]. \end{aligned}$$

In all these settings the matrices U_n are unitary. The Hartley transform H_n is also real and so the algebra is a real one. For all these cases the matrix vector product with matrix U_n and a generic vector can be performed in $O(n\log(n))$ arithmetic operations (real in the Hartley case) and therefore, as proved for circulants in Subsection 4.1, all the matrix operations such as eigenvalue computation, solution of a linear system, matrix vector product etc. can be performed in $O(n\log(n))$ arithmetic operations when a matrix in one of these algebras is concerned.

The same remarks hold for the 8 cosine/sine algebras (all real algebras) whose transform U_n is explicitly reported.

Discrete cosine/sine transform matrices U_n.

	Discrete transform	Inverse transform
DCT-I	$C_n^I = \sqrt{\frac{2}{n-1}} \left[g(j,k) \cos \frac{kj\pi}{n-1} \right]_{k,j=0}^{n-1}$	$[C_n^I]^T = C_n^I$
DCT-II	$C_n^{II} = \sqrt{\frac{2}{n}} \left[\eta_k \cos \frac{k(2j+1)\pi}{2n} \right]_{k,j=0}^{n-1}$	$[C_n^{II}]^T = C_n^{III}$
DCT-III	$C_n^{III} = \sqrt{\frac{2}{n}} \left[\eta_j \cos \frac{(2k+1)j\pi}{2n} \right]_{k,j=0}^{n-1}$	$[C_n^{III}]^T = C_n^{II}$
DCT-IV	$C_n^{IV} = \sqrt{\frac{2}{n}} \left[\cos \frac{(2k+1)(2j+1)\pi}{4n} \right]_{k,j=0}^{n-1}$	$[C_n^{IV}]^T = C_n^{IV}$
DST-I	$S_n^I = \sqrt{\frac{2}{n+1}} \left[\sin \frac{kj}{n+1}\pi \right]_{k,j=1}^{n}$	$[S_n^I]^T = S_n^I$
DST-II	$S_n^{II} = \sqrt{\frac{2}{n}} \left[\eta_k \sin \frac{k(2j-1)}{2n}\pi \right]_{k,j=1}^{n}$	$[S_n^{II}]^T = S_n^{III}$
DST-III	$S_n^{III} = \sqrt{\frac{2}{n}} \left[\eta_j \sin \frac{(2k-1)j}{2n}\pi \right]_{k,j=1}^{n}$	$[S_n^{III}]^T = S_n^{II}$
DST-IV	$S_n^{IV} = \sqrt{\frac{2}{n}} \left[\sin \frac{(2k-1)(2j-1)}{4n}\pi \right]_{k,j=1}^{n}$	$[S_n^{IV}]^T = S_n^{IV}$

Here $\eta_0 = \eta_n = \frac{1}{2}$, $\eta_k = 1$ for $k = 1, 2, \ldots, n-1$ and $g(j,k) = \eta_k \eta_{n-1-k} \eta_j \eta_{n-1-j}$. The grid points of the above mentioned 8 cosine/sine transforms belong to $[0, \pi]$ and are defined as follows:

$$
\begin{array}{rl}
\mathrm{DCT-I}, & W_n = \left\{ x_k^{(n)} = \frac{k\pi}{n-1} : k = 0, \ldots, n-1 \right\}, \\
\mathrm{DCT-II}, & W_n = \left\{ x_k^{(n)} = \frac{k\pi}{n} : k = 0, \ldots, n-1 \right\}, \\
\mathrm{DST-I}, & W_n = \left\{ x_k^{(n)} = \frac{(k+1)\pi}{n+1} : k = 0, \ldots, n-1 \right\}, \\
\mathrm{DST-II}, & W_n = \left\{ x_k^{(n)} = \frac{(k+1)\pi}{n} : k = 0, \ldots, n-1 \right\}, \\
\mathrm{DCT-III},\ \mathrm{DCT-IV}, \\ \mathrm{DST-III},\ \mathrm{DST-IV}, & W_n = \left\{ x_k^{(n)} = \frac{(k+1/2)\pi}{n} : k = 0, \ldots, n-1 \right\}.
\end{array}
$$

We notice that all the sequences of unitary algebras considered so far have grid points W_n which are uniformly distributed in the reference interval. For a formal definition of quasi-uniform and uniform distribution see the following

Definition 5.1 *A sequence of grids $\{W_n = \{x_k^{(n)} : \ k = 0, \ldots, n-1\}\}$ belonging to an interval J is called quasi-uniform in J if*

$$
\lim_{n \to \infty} \sum_{k=0}^{n-1} \left| \frac{|J|}{n} - (x_k^{(n)} - x_{k-1}^{(n)}) \right| = 0, \tag{32}
$$

with $|J|$ being the width of J. If the previous relation in (32) holds with a decay of $O(n^{-1})$, then the mesh-sequence $\{W_n\}$ is called uniform.

Concerning requirement **A**, it is evident that any invertible matrix belonging to one these algebras is easy to invert ($O(n \log(n))$ ops required). What about the closeness of these spaces with respect to the Toeplitz matrices? The answer is positive and indeed if α_n is the space of Toeplitz matrices with generic symbol, then any of the trigonometric spaces $\mathcal{A}_n$ with complex transform can be used as β_n; if we restrict the attention to the Toeplitz matrices with real and even symbol (then $T_n(f)$ is real and symmetric), then as space β_n we can use any of the above mentioned algebras. We prove this closeness on the space of band Toeplitz matrices i.e. on the class of Toeplitz matrices with polynomial symbols.

Theorem 5.1 *Let $\mathcal{A}_n$ be any of the ω-circulant, Hartley, or of the 8 cosine/sine algebras. Setting*

$$\mathcal{A}_n(g) = U_n D_g U_n^H, \;\; D_g = \mathrm{diag}_{0\le k\le n-1}\left(g\left(x_k^{(n)}\right)\right), \;\; x_k^{(n)} \in W_n, \tag{33}$$

we have the following:

- *if p is a real valued even trigonometric polynomial, then there exists a constant c such than $\mathrm{rank}(T_n(p) - \mathcal{A}_n(p)) \le c$ with c being linear with respect to the degree of p;*
- *if p a (complex valued) trigonometric polynomial and if $\mathcal{A}_n$ is the circulant algebra or any of the ω-circulant algebras, then there exists a constant c such than $\mathrm{rank}(T_n(p) - \mathcal{A}_n(p)) \le c$ with c being linear with respect to the degree of p.*

Proof. We give the proof only for the circulant algebra: the other cases are similar but require some additional technical details concerning the structure of the remaining algebras. Let $p(x) = \sum_{-q_1\le k\le q_2} a_k e^{\mathbf{i}kx}$. Therefore for n large enough, we have

$$T_n(p) = \begin{pmatrix} a_0 & a_{-1} & \dots & a_{-q_1} & & & \\ a_1 & \ddots & \ddots & \ddots & \ddots & & \\ \vdots & \ddots & & & & \ddots & \\ a_{q_2} & \ddots & & & & \ddots & a_{-q_1} \\ & \ddots & & & & \ddots & \vdots \\ & & \ddots & \ddots & \ddots & \ddots & a_{-1} \\ & & & a_{q_2} & \dots & a_1 & a_0 \end{pmatrix}.$$

We choose the circulant matrix A_n (known in the literature as Strang preconditioner [87])

$$A_n = \begin{pmatrix} a_0 & a_{-1} & \dots & a_{-q_1} & & & a_{q_2} & \dots & a_1 \\ a_1 & \ddots & \ddots & \ddots & \ddots & & & \ddots & \vdots \\ \vdots & \ddots & & & & & & & a_{q_2} \\ a_{q_2} & \ddots & & & & & & & \\ & \ddots & & & & & & \ddots & \\ & & & & & & & \ddots & a_{-q_1} \\ a_{-q_1} & & & & & & & \ddots & \vdots \\ \vdots & \ddots & & & \ddots & \ddots & \ddots & \ddots & a_{-1} \\ a_{-1} & \dots & a_{-q_1} & & & a_{q_2} & \dots & a_1 & a_0 \end{pmatrix}.$$

It is clear that $\mathrm{rank}(T_n(p) - A_n) \le c$ with $c = q_1 + q_2$. Finally a simple computation based on the formula (18) shows that $A_n = \mathcal{A}_n(p)$. •

The formula (33) is very important for the approximation of Toeplitz matrices by using unitary matrix algebras. Indeed it can be proved that a necessary condition for $\mathcal{A}_n(g)$ to be a good preconditioner for $T_n(f)$ is that $g = \mathcal{S}_n(f)$ where $\mathcal{S}_n(f)$ converges to f as n tends to infinity; if f is strictly positive and bounded then the condition is also sufficient (see [68]). A very popular preconditioning technique is Strang approximation [87, 22]: it corresponds to $\mathcal{S}_n(f) = F_{\lceil n/2\rceil}(f)$ with

$F_q(f) = \sum_{-q \le k \le q} a_k e^{\mathbf{i}kx}$ being the Fourier sum of f (with Fourier coefficients a_k). The algebraic expression of the Strang preconditioner corresponds to take the matrix $T_n(f)$ and to copy only its $\lceil n/2 \rceil$ central diagonal by replacing the others in order to obtain a circulant structure. A further important alternative is the T. Chan preconditioner related to the Frobenius optimal approximation whose features are discussed in a more general setting in Subsection 5.2.

5.1 The classical Korovkin theory

Our aim is to approximate continuous functions over a compact domain $K \subset \mathbf{R}^d$, $d \ge 1$, by means of functions which are simpler from the computational viewpoint. A natural choice is the polynomial model since the evaluation of a generic polynomial implies only a finite number of sums and products. Concerning the notion of approximation, the idea is to replace the given continuous function f with a polynomial p which is close to it in the whole domain K. More precisely we endow the set $C(K)$ of all continuous functions on K with the sup distance $d(f,g) = \|f-g\|_{\infty,K}$ where $\|\cdot\|_{\infty,K}$ indicates the sup norm namely

$$\|h\|_{\infty,K} = \sup_{x \in K} |h(x)|, \quad h \in C(K).$$

In this way the pair $(C(K), d(\cdot,\cdot))$ is a Banach space i.e. any Cauchy sequence has limit in it. In view of the Weierstrass Theorem (on the existence of points of the compact set K where a continuous functions f attains its minimum and maximum), it is evident that the sup can be replaced by a max (K is a compact set of $\mathbf{R}^d$, $d \ge 1$) and therefore

$$\|h\|_{\infty,K} = \max_{x \in K} |h(x)|, \quad h \in C(K). \tag{34}$$

Now the problem is posed. We observed that the space of polynomials is a good idea from a computational point of view but is a good choice from an approximation viewpoint?

The Weierstrass Theorem on the polynomial approximation gives a complete answer to this question.

Theorem 5.2 (Weierstrass) *Let $f \in C(K)$ with $K \subset \mathbf{R}^d$, $d \ge 1$, compact set. For every $\epsilon > 0$ there exists a polynomial p_ϵ such that*

$$\|f - p_\epsilon\|_{\infty,K} \le \epsilon.$$

In the following with the help of the Korovkin theorem and with the help of the Bernstein polynomial, we give a constructive proof of the Weierstrass Theorem (and indeed we provide a general tool for proving Weierstrass Theorems in many different contexts).

The power of the Korovkin Theorem can be resumed as follows: its assumptions are simple and easy to verify, the claim is strong and the proof is elementary in the sense that it does not require advanced mathematical knowledge. More specifically, given a sequence of operators Φ_n from $C(K)$ into itself, it is enough to verify that they are definitely linear and positive and that converge on a finite number of simple functions (some polynomials of degree at most 2) in order to conclude that they converge on every continuous function f i.e.

$$d(\Phi_n(f), f) \to 0, \ \text{ as } n \to \infty.$$

In the following we report the definition of linear positive operator (LPO) and the notion of linear approximation process.

Definition 5.2 *Let $\mathcal{S}$ be a vector space of functions taking values in $\mathbf{K}$ (with $\mathbf{K}$ being $\mathbf{R}$ or $\mathbf{C}$) and let us consider an operator Φ from $\mathcal{S}$ to $\mathcal{S}$ satisfying the following pair of properties: Φ:*

1. *for every α and β in $\mathbf{K}$ for every f and g in $\mathcal{S}$, $\Phi(\alpha f + \beta g) = \alpha\Phi(f) + \beta\Phi(g)$ (linearity);*

2. *for every $f \geq 0$, $f \in \mathcal{S}$, $\Phi(f) \geq 0$ (positivity).*

An operator Φ fulfilling both the conditions is said linear and positive (LPO).
Analogously, if $\mathcal{S}$ is linear space (over $\mathbf{K}$) of matrices and Φ is an operator from $\mathcal{S}$ into $\mathcal{S}$ satisfying the following two properties,

1. *for every α and β in $\mathbf{K}$ for every f and g in $\mathcal{S}$, $\Phi(\alpha f + \beta g) = \alpha\Phi(f) + \beta\Phi(g)$ (linearity);*

2. *for every $A \in \mathcal{G}$ Hermitian and nonnegative definite, $\Phi(A)$ is Hermitian and nonnegative definite.*

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

Definition 5.3 *Let $(\mathcal{S}, d(\cdot,\cdot))$ be a Banach space of functions taking values in $\mathbf{K}$ (with $\mathbf{K}$ being $\mathbf{R}$ or $\mathbf{C}$) and let us consider a sequence $\{\Phi_n\}$ of operators from $\mathcal{S}$ into $\mathcal{S}$. The sequence is called sequence of operators of approximation or more briefly process of approximation if for every $f \in \mathcal{S}$ we have*

$$\lim_{n\to\infty} d(\Phi_n(f), f) = 0.$$

(In other words $\{\Phi_n\}$ converges pointwise to the identity operator).

We now give a proof of the Korovkin Theorem. With $\|\cdot\|_2$ we denote the Euclidean norm over $\mathbf{R}^d$ that is $\|x\|_2 = \left(\sum_{i=1}^d |x_i|^2\right)^{1/2}$ for any $x \in \mathbf{R}^d$.

Theorem 5.3 (Korovkin) *Let K be a compact set of $\mathbf{R}^d$ and let $C(K)$ be the Banach space of the continuous functions (real valued or complex valued) over K with the sup norm. We consider the set of functions $T = \{1, x_i, \|x\|_2^2 : \ i = 1, \ldots, d\}$, that we call the Korovkin set. Let $\{\Phi_n\}$ be a LPO sequence over $C(K)$. If for every $g \in T$*

$$\Phi_n(g) \ \text{converges uniformly to } g$$

then $\{\Phi_n\}$ is an approximation process i.e.

$$\Phi_n(f) \ \text{converges uniformly to } f, \ \forall f \in C(K).$$

The same statement holds if the uniform convergence is replaced by the pointwise convergence or the convergence on the domain K is replaced by the convergence on any sub domain J.

Proof We fix $f \in C(K)$, we fix $\epsilon > 0$ and we show that there exists $\bar{n}$ large enough such that, for all $n \geq \bar{n}$, we have $\|f - \Phi_n(f)\|_{\infty,K} \leq \epsilon$. Therefore we assume that x is a generic point of K and we consider the difference

$$f(x) - (\Phi_n(f(y)))(x)$$

where y is the "dummy" variable where the function f acts as argument of Φ_n. Now the constant function 1 belongs to the "Korovkin set" T and therefore $1 = (\Phi_n(1))(x) - (\epsilon_n(1))(x)$ where $\epsilon_n(1)$ converges uniformly to zero over K. From the linearity of the operators Φ_n, we deduce

$$\begin{aligned} f(x) - (\Phi_n(f(y)))(x) &= [(\Phi_n(1))(x) - (\epsilon_n(1))(x)]f(x) - (\Phi_n(f(y)))(x) \\ &= (\Phi_n(f(x) - f(y)))(x) - (\epsilon_n(1))(x)f(x). \end{aligned}$$

Therefore we can find a value n_1 such that for $n \geq n_1$

$$|(\epsilon_n(1))(x)f(x)| \leq \|\epsilon_n(1)\|_{\infty,K}\|f\|_{\infty,K} \leq \epsilon/4$$

and consequently, by exploiting the linearity and the positivity of Φ_n, we have

$$\begin{aligned} |f(x) - (\Phi_n(f(y)))(x)| &\leq |(\Phi_n(f(x) - f(y)))(x)| + \epsilon/4 \quad (35) \\ &\leq (\Phi_n(|f(x) - f(y)|))(x) + \epsilon/4. \quad (36) \end{aligned}$$

The remaining part of the proof is now a manipulation of the term $|f(x) - f(y)|$ where we look for a "clever" upperbound in order to properly exploit the positivity of the operators and the convergence assumptions on the Korovkin test.

First of all we notice that a continuous function over a compact set is also uniformly continuous and therefore, in correspondence to our fixed $\epsilon > 0$, there exists a value $\delta > 0$ for which if $\|x-y\|_2 \leq \delta$ then $|f(x) - f(y)| \leq \epsilon/4$. In association to the parameter δ and therefore in association to the behavior of $|f(x) - f(y)|$, we define the pair of sets

$$Q_\delta(x) = \{y \in K : \|x - y\|_2 \leq \delta\}, \;\; Q_\delta^C(x) = K \backslash Q_\delta,$$

where $|f(x)-f(y)|$ as function of y (x is fixed and therefore it acts as a parameter) is bounded by $\epsilon/4$ on $Q_\delta(x)$ and, as a consequence of the triangle inequality, is bounded by da $2\|f\|_{\infty,K}$ over $Q_\delta^C(x)$. Denoted by $\chi_{\mathcal{J}}$ the characteristic function of $\mathcal{J}$, we remark that $\forall y \in Q_\delta^C(x)$ we have $\|x - y\|_2 > \delta$ i.e.

$$1 \leq \|x - y\|_2^2/\delta^2.$$

Hence we deduce the following chain of inequalities:

$$\begin{aligned} |f(x) - f(y)| &\leq \epsilon/4\chi_{Q_\delta(x)}(y) + 2\|f\|_{\infty,K}\chi_{Q_\delta^C(x)}(y) \\ &\leq \epsilon/4\chi_{Q_\delta(x)}(y) + 2\|f\|_{\infty,K}\chi_{Q_\delta^C(x)}(y)\|x - y\|_2^2/\delta^2 \\ &\leq \epsilon/4 + 2\|f\|_{\infty,K}\|x - y\|_2^2/\delta^2. \end{aligned}$$

In spite of the use of discontinuous functions (the characteristic functions), we observe that the last term of the inequality chain is a continuous functions: as a consequence we can apply the operator Φ_n and its positivity allows us to conclude

$$\Phi_n(|f(x) - f(y)|) \leq \Phi_n\left(\epsilon/4 + 2\|f\|_{\infty,K}\|x - y\|_2^2/\delta^2\right).$$

Therefore by means of the linearity of Φ_n and setting $(\Delta_n(f))(x) = |f(x) - (\Phi_n(f(y)))(x)|$, from (35) we deduce the following further chain of relationships:

$$\begin{aligned} (\Delta_n(f))(x) &\leq (\Phi_n(|f(x) - f(y)|))(x) + \epsilon/4 \\ &\leq \left(\Phi_n\left(\epsilon/4 + 2\|f\|_{\infty,K}\|x - y\|_2^2/\delta^2\right)\right)(x) + \epsilon/4 \\ &= \epsilon/4(\Phi_n(1))(x) + 2\|f\|_{\infty,K}/\delta^2\left(\Phi_n\left(\|x - y\|_2^2\right)\right)(x) + \epsilon/4 \\ &= \epsilon/4(\Phi_n(1))(x) + 2\|f\|_{\infty,K}/\delta^2 \\ &\quad \left(\Phi_n\left(\sum_{i=1}^d x_i^2 - 2x_iy_i + y_i^2\right)\right)(x) + \epsilon/4 \\ &= \epsilon/4(\Phi_n(1))(x) + 2\|f\|_{\infty,K}/\delta^2 \\ &\quad \sum_{i=1}^d \left(\Phi_n\left(x_i^2 - 2x_iy_i + y_i^2\right)\right)(x) + \epsilon/4 \end{aligned}$$

$$
\begin{aligned}
&= \epsilon/4(\Phi_n(1))(x) + 2\|f\|_{\infty,K}/\delta^2 \\
&\quad \sum_{i=1}^{d}\Big[x_i^2(\Phi_n(1))(x) - 2x_i(\Phi_n(y_i))(x) + (\Phi_n(y_i^2))(x)\Big] + \epsilon/4 \\
&= \epsilon/4(\Phi_n(1))(x) + 2\|f\|_{\infty,K}/\delta^2 \\
&\quad \left[\sum_{i=1}^{d}\Big[x_i^2(\Phi_n(1))(x) - 2x_i(\Phi_n(y_i))(x)\Big] + (\Phi_n(\|y\|_2^2))(x)\right] \\
&\quad +\epsilon/4.
\end{aligned}
$$

We have reduced the original problem to linear combinations of products among continuous functions where the operator Φ_n is applied only over the functions of the Korovkin set. Therefore we can proceed with explicit computations: setting $\Phi_n(g) = g + \epsilon_n(g)$, with $g \in T$ and $\epsilon_n(g)$ converging to zero uniformly over K, we have

$$
\begin{aligned}
\Delta_n(f) &\leq \epsilon/4(1 + (\epsilon_n(1))(x)) + 2\|f\|_{\infty,K}/\delta^2 \\
&\quad \left[\sum_{i=1}^{d}\Big[x_i^2(\Phi_n(1))(x) - 2x_i(\Phi_n(y_i))(x)\Big] + (\Phi_n(\|y\|_2^2))(x)\right] + \epsilon/4 \\
&= \epsilon/4(1 + (\epsilon_n(1))(x)) + 2\|f\|_{\infty,K}/\delta^2 \\
&\quad \left[\sum_{i=1}^{d}\Big[x_i^2(\epsilon_n(1))(x) - 2x_i(\epsilon_n(y_i))(x)\Big] + (\epsilon_n(\|y\|_2^2))(x)\right] + \epsilon/4.
\end{aligned}
$$

Finally, by virtue of the uniform convergence to zero of the functions $\epsilon_n(g)$, we infer that there exists a value $\bar{n} \geq n_1$ such that $\forall n \geq \bar{n}$ we have $(\epsilon_n(1))(x)) \leq 1$ and

$$
2\|f\|_{\infty,K}/\delta^2 \left[\sum_{i=1}^{d}\Big[x_i^2(\epsilon_n(1))(x) - 2x_i(\epsilon_n(y_i))(x)\Big] + (\epsilon_n(\|y\|_2^2))(x)\right] \leq \epsilon/4.
$$

Finally, by combining all the partial results, we have proved the desired result that is, uniformly with respect to x,

$$
(\Delta_n(f))(x) = |f(x) - (\Phi_n(f(y)))(x)| \leq \epsilon, \quad \forall n \geq \bar{n}.
$$

The proof in the case of pointwise convergence or in the case of convergence on a sub domain of K follow exactly the same lines of the proof above. •

A periodic version of the latter result holds; the proof is virtually unchanged so that we leave it to the reader.

Theorem 5.4 (Korovkin) *Let $K = [-\pi, \pi]^p$ and let $C_p(K)$ be the Banach space of the (2π)-periodic continuous functions (real valued or complex valued) over K with the sup norm. We consider the set of functions $T = \{1, e^{\mathbf{i}jx_i} : \ i = 1, \ldots, d, j = \pm 1\}$, that we call the Korovkin set. Let $\{\Phi_n\}$ be a LPO sequence over $C_p(K)$. If for every $g \in T$*

$$
\Phi_n(g) \ \text{converges uniformly to } g
$$

then $\{\Phi_n\}$ is an approximation process i.e.

$$
\Phi_n(f) \ \text{converges uniformly to } f, \ \forall f \in C_p(K).
$$

The same statement holds if the uniform convergence is replaced by the pointwise convergence or the convergence on the domain K is replaced by the convergence on any sub domain J.

Finally we briefly report a bit of the quantitative Korovkin theory.

Theorem 5.5 *Under the assumptions of Theorem 5.3 (under the assumptions of Theorem 5.4), if* $\max_{g\in T}\|\Phi_n(g)-g\|_{\infty,K}=\theta_n$ *(*T *being the Korovkin set), then every polynomial* p *(then for every trigonometric polynomial* p*) of fixed degree (independent of* n*) we find*

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

Proof. We give only an idea of the proof for $d=1$, in the nonperiodic case. Let p be a standard polynomial of the real variable $x\in K$ of given degree. Then

$$p(x)-p(y)=p''(\eta(x,y))(x-y)^2$$

and therefore

$$\begin{aligned}
|p(x)-(\Phi_n(p(y)))(x)| &= |[(\Phi_n(1))(x)-(\epsilon_n(1))(x)]p(x)-(\Phi_n(p(y)))(x)| \\
&\le |(\Phi_n(p(x)-p(y)))(x)|+|(\epsilon_n(1))(x)p(x)| \\
&\le |(\Phi_n(p''(\eta(x,y))(x-y)^2))(x)|+\theta_n\|p\|_{\infty,K} \\
&\le (\Phi_n(|p''(\eta(x,y))|(x-y)^2))(x)+\theta_n\|p\|_{\infty,K} \\
&\le \|p''\|_{\infty,K}(\Phi_n((x-y)^2))(x)|+\theta_n\|p\|_{\infty,K} \\
&\le \|p''\|_{\infty,K}C\theta_n+\theta_n\|p\|_{\infty,K}
\end{aligned}$$

with C universal constant depending only the set K.

The multilevel nonperiodic case can be handled in a very similar way, while the proof in the periodic case is completely different (see [71]). •

These interesting results deserve some additional comments.

A)

In order to obtain a proof of the first Weierstrass Theorem 5.2 (in **B)** we propose a sequence $\{\Phi_n\}$ of LPOs satisfying the Korovkin test such that $\Phi_n(f)$ is polynomial for every $f\in C(K)$. However we cannot forget that the Korovkin Theorem 5.4 is completely general and does not impose any restriction on the functions $\Phi_n(f)$. In actuality, by following carefully the proof of the Theorem 5.4, we observe that the proof is unchanged even if $\Phi_n(f)$ is not continuous for some $f\in C(K)$ i.e. if we violate the assumption $\Phi_n: C(K)\to C(K)$. The only constraint is that it should makes sense to compute $\|\Phi_n(f)\|_{\infty,K}$ that is $\Phi_n(f)\in L^\infty(K)$.

B)

The Korovkin Theorem 5.4 can be used for proving the first Weierstrass Theorem 5.2 on the approximation of functions in $C(K)$ by polynomials: the tools are the Bernstein polynomials. In one dimension i.e. $d=1$ and for $K=[0,1]$, they are defined as

$$(B_n(f))(x)=\sum_{\nu=0}^{n}\left(\begin{array}{c} n \\ \nu \end{array}\right)x^\nu(1-x)^{n-\nu}f\left(\frac{\nu}{n}\right).$$

It is evident that $B_n(\cdot)$ is a LPO since it is a linear combination of nonnegative polynomials of degree n $x^\nu(1-x)^{n-\nu}$, via coefficients

$$\left(\begin{array}{c} n \\ \nu \end{array}\right)f\left(\frac{\nu}{n}\right)$$

which are all nonnegative if $f\ge 0$. Moreover it is easy to prove that $B_n(g)$ converges uniformly to g for $g(y)=y^j$, $j=0,1,2$ (the one-dimensional Korovkin set). Therefore, by the Korovkin

Theorem 5.4, $B_n(f)$ uniformly converges to f for every $f \in C(K)$. A d-dimensional generalization of the Bernstein polynomials allows to give a d-dimensional version of Weierstrass approximation Theorem.

C)

The Korovkin Theorem can be used for proving the second Weierstrass Theorem on the approximation of functions in $C_p(K)$ by trigonometric polynomials where K is a d-dimensional rectangle and $f \in C_p(K)$ means that it is continuous and periodic in every direction: a possible tool is represented by the Cesaro sums. In one dimension i.e. $d = 1$ and for $K = [-\pi, \pi]$, they are defined as

$$(C_n(f))(x) = \frac{1}{n}\sum_{\nu=0}^{n-1}(F_\nu(f))(x), \;\; (F_\nu(f))(x) = \sum_{j=\nu}^{\nu} a_j e^{\mathbf{i}jx}$$

with a_j being the Fourier coefficients of f. By an integral representation of $C_n(\cdot)$, it is possible to prove that $C_n(\cdot)$ is a LPO. Furthermore it is easy to prove that $C_n(g)$ converges uniformly to g for $g(y) = e^{\mathbf{i}jx}$, $j = 0, \pm 1$ (the one-dimensional periodic Korovkin set). Therefore, by the (periodic) Korovkin Theorem 5.4, it follows that $C_n(f)$ uniformly converges to f for every $f \in C_p(K)$.

A d-dimensional generalization of the Cesaro polynomials allows to give a d-dimensional version of the second Weierstrass approximation Theorem.

D)

In the Korovkin Theorems 5.3, 5.4 and 5.5, we considered the uniform convergence. The same kind of result holds if the we consider the point wise convergence or different convergence in norm (L^p convergence in L^p spaces). Also of interest is the following notion of convergence. Let $\{W_n\}$, $W_n \subset K$, be a sequence of subsets of K then we define

$$\|f\|_{\infty,W_n} = \sup_{x\in W_n} |f(x)|. \tag{37}$$

If in Theorem 5.4 and Theorem 5.5 we replace $\|\cdot\|_{\infty,K}$ by $\|\cdot\|_{\infty,W_n}$, then the theorems stand unchanged (the same is true of course also for Theorem 5.3). In our matrix generalization, in the following subsection (see Theorems 5.8 and 5.9), we use this type of convergence where $\{W_n\}$ is the sequence of grid points associated to the approximating matrix algebra sequence (*asymptotic discrete convergence*).

5.2 The Korovkin theorem for Toeplitz matrix sequences

Here we give a matrix version of the Korovkin theory and more specifically of Theorem 5.4. We consider the unilevel case of $d = 1$ while the general case is discussed in Subsections 6.1 and 6.2:

- The space $C_p(K)$ is replaced by the sequence space $\{\{T_n(f)\} : f \in C_p(K)\}$ with $K = [-\pi, \pi]$;
- the approximation space of the polynomials is replaced by $\{\mathcal{A}_n\}$ where $\mathcal{A}_n = \{X = U_n D U_n^H : D \text{ diagonal matrix}\}$ is a unitary algebra of matrices i.e. for any n the $n \times n$ transform U_n is a unitary matrix (see (31));
- the norm in the space of matrices is the Frobenius norm that is $\|X\|_F = \sqrt{\sum_{i,j=1}^n |X_{i,j}|^2}$;
- the approximation operators $\Phi_n(\cdot)$ are defined as follows: for every $n \times n$ matrix A we consider $\phi_n(\cdot) : M_n(\mathbf{C}) \to \mathcal{A}_n \subset M_n(\mathbf{C})$ described by the relation

$$\phi_n(A) = \arg\min_{X\in\mathcal{A}_n} \|A - X\|_F; \tag{38}$$

therefore for every $f \in C_p(K)$, $\Phi_n(f) : C_p(K) \to \mathcal{A}_n$ is defined as

$$\Phi_n(f) = \phi_n(T_n(f)). \tag{39}$$

We first observe that the Frobenius norm coincides with $\|\cdot\|_2$ where $\|\cdot\|_p$, $p \in [1, \infty]$ is the p-th Schatten norm (see (2)); moreover $\|\cdot\|_F$ is induced by the positive scalar product $(\cdot,\cdot)_F$ on $M_n(\mathbf{C})$ defined as $(A, B)_F$ =trace$(A^* \cdot B)$. Therefore the existence and the uniqueness of the minimum

$$\phi_n(A) = \arg\min_{X \in \mathcal{A}_n} \|A - X\|_F$$

follows from the fact that the space $(M_n(\mathbf{C}), (\cdot,\cdot)_F)$ is a Hilbert space and $\mathcal{A}_n$ is a closed convex subset since it is a finite dimensional vector space. In the case of $A = T_n(f)$ and in the case where $\mathcal{A}_n$ is the space of circulants, by expanding a generic circulant in its canonical basis, a simple calculation shows that

$$\Phi_n(f) = \begin{pmatrix} a_0 & a'_1 & a'_2 & \dots & a'_{n-3} & a'_{n-2} & a'_{n-1} \\ a'_{n-1} & a_0 & a'_1 & a'_2 & \dots & a'_{n-3} & a'_{n-2} \\ & \ddots & \ddots & \ddots & & & \\ \vdots & & \ddots & \ddots & \ddots & & \vdots \\ & & & \ddots & \ddots & \ddots & \\ a'_2 & \dots & a'_{n-3} & a'_{n-2} & a'_{n-1} & a_0 & a'_1 \\ a'_1 & a'_2 & \dots & a'_{n-3} & a'_{n-2} & a'_{n-1} & a_0 \end{pmatrix} \tag{40}$$

where a_j are the Fourier coefficients of f (i.e. the entries of $T_n(f)$) and

$$a'_j = [(n-j)a_{-j} + ja_{n-j}]/n, \ \forall j = 0, \dots, n-1. \tag{41}$$

By means of simple algebraic arguments, it is possible to prove the following Lemma (see e.g. [32]).

Lemma 5.1 *With $A, B \in M_n(\mathbf{C})$ and the previous definition of $\phi_n(\cdot)$, we have*

a. $\phi_n(A) = U_n \sigma(U_n^H A U_n) U_n^H$, *with $\sigma(X)$ being the diagonal matrix having $(X)_{i,i}$ as diagonal elements,*

b. $\phi_n(\alpha A + \beta B) = \alpha\phi_n(A) + \beta\phi_n(B)$ *and* $\alpha, \beta \in \mathbf{C}$,

c. $\phi_n(A^H) = (\phi_n(A))^H$,

d. $\|A - \phi_n(A)\|_F^2 = \|A\|_F^2 - \|\phi_n(A)\|_F^2$,

e. $\|\phi_n(A)\| \le \|A\|$.

Proof. Since U_n is unitary it follows that the Euclidean norm of a generic vector $\mathbf{v}$ is equal the Euclidean norm of $U_n\mathbf{v}$. Therefore

$$\begin{aligned} \phi_n(A) &= \arg\min_{X \in \mathcal{A}_n} \|A - X\|_F \\ &= \arg\min_{X \in \mathcal{A}_n} \|U_n^H(A - X)\|_F \\ &= \arg\min_{X \in \mathcal{A}_n} \|U_n^H(A - X)U_n\|_F \\ &= \arg\min_{X = U_n D U_n^H:\ D \text{ is diagonal}} \|U_n^H A U_n - D\|_F \end{aligned}$$

and consequently the optimal diagonal matrix is $\sigma(U_n^* A U_n)$ i.e.

$$\phi_n(A) = U_n \sigma(U_n^H A U_n) U_n^H.$$

This proves the first item. The second is a direct consequence of the first, while for the third it is sufficient to observe that $\sigma(U_n^H A^H U_n) = (\sigma(U_n^H A U_n))^H$ and therefore $\phi_n(A^H) = (\phi_n(A))^H$. Part **d** is simply the Pithagora law for general Hilbert spaces and the last item follows again from the first one since $\|\phi_n(A)\| \leq \|U_n\| \|\sigma(U_n^H A U_n)\| \|U_n^H\| = \|\sigma(U_n^H A U_n)\|$: finally,

$$\begin{aligned} \|\sigma(U_n^H A U_n)\| &= \max_i |\mathbf{u}_i^H A \mathbf{u}_i| \\ &\leq \max_{\|\mathbf{u}\|=\|\mathbf{v}\|=1} |\mathbf{u}^H A \mathbf{v}| = \sigma_n(A) = \|A\| \end{aligned}$$

with $\mathbf{u}_i$ i-th column of U_n and $\sigma_n(A)$ largest singular value of A (for the relation $\max_{\|\mathbf{u}\|=\|\mathbf{v}\|=1} |\mathbf{u}^H A \mathbf{v}| = \sigma_n(A) = \|A\|$ see e.g. [7]) . •

Also of interest is the following result (see e.g [32]).

Lemma 5.2 *If A is Hermitian ($A = A^*$), then the eigenvalues of $\phi_n(A)$ 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, $\phi_n(A)$ is positive definite as well.*

Proof. From item **c** of Lemma 5.1, it follows that $\phi_n(A)$ is Hermitian if A is and its eigenvalues $(\sigma(U_n^H A U_n)_{i,i}$ are of the form

$$\mathbf{u}_i^H A \mathbf{u}_i \tag{42}$$

with $\mathbf{u}_i$ i-th column of U_n. Moreover by the Schur normal form Theorem (see e.g. [7]) we have $\lambda_1(A) = \min_{\|\mathbf{v}\|_2=1} \mathbf{v}^H A \mathbf{v}$ and $\lambda_n(A) = \max_{\|\mathbf{v}\|_2=1} \mathbf{v}^H A \mathbf{v}$ and therefore

$$\lambda_1(A) = \min_{\|\mathbf{v}\|_2=1} \mathbf{v}^H A \mathbf{v} \leq \mathbf{u}_i^H A \mathbf{u}_i, \quad \lambda_n(A) = \max_{\|\mathbf{v}\|_2=1} \mathbf{v}^H A \mathbf{v} \geq \mathbf{u}_i^H A \mathbf{u}_i.$$

The second part is a consequence of the first. •

5.2.1 A Weierstrass matrix theory for Toeplitz matrices

In order to properly state the "matrix approximation results", we require a concept of "matrix convergence". We say that "$\{\Phi_n(f)\}$ (strongly) converges to $\{T_n(f)\}$" if $\{\Phi_n(f) - T_n(f)\}$ is properly clustered at zero in the sense of the singular values; the convergence is "weak" if the same difference sequence is weakly clustered at zero (see the notions of proper and weak clustering in Definition 8.1).

Theorem 5.6 *Let f be a continuous (2π)-periodic function on $I = [-\pi, \pi)$. Then, $\{\Phi_n(f)\}$ converges to $\{T_n(f)\}$ if $\{\Phi_n(p)\}$ converges to $\{T_n(p)\}$ for all the trigonometric polynomials p.*

Proof. Let p_k be the polynomial having degree k of best approximation of f in supremum norm [45]. For any $\epsilon > 0$, fix the integer M such that $\|f - p_M\|_\infty < \epsilon/3$. Then, by using (2) and items **b** and **d** of Lemma 5.1 we have $\|T_n(f) - T_n(p_M)\| = \|T_n(f - p_M)\| < \epsilon/3$, $\|\Phi_n(f) - \Phi_n(p_M)\| = \|\Phi_n(f - p_M)\| < \epsilon/3$. Therefore, from the identity

$$\begin{aligned} T_n(f) - \Phi_n(f) &= T_n(f) - T_n(p_M) - \\ &\quad -(\Phi_n(f) - \Phi_n(p_M)) + \\ &\quad +T_n(p_M) - \Phi_n(p_M) \end{aligned}$$

we have that, except for a term of norm bounded by $2\epsilon/3$, the difference $T_n(f)-\Phi_n(f)$ coincides with $T_n(p_M)-\Phi_n(p_M)$. From the hypothesis of convergence, we can split the matrix $T_n(p_M)-\Phi_n(p_M)$ 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 [7, 55] for singular values (or eigenvalues if f is real valued). •

Theorem 5.7 *Let f be a continuous periodic function. Then, $\{\Phi_n(f)\}$ weakly converges to $\{T_n(f)\}$ if $\{\Phi_n(p)\}$ weakly converges to $\{T_n(p)\}$ for all the trigonometric polynomials p.*

Proof. The proof is the same as the one of Theorem 5.6 with the exception of the last part where we split $\{T_n(p_M)-\Phi_n(p_M)\}$ 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 Theorem completes the proof. •

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

Corollary 5.1 *Under the assumption of the Theorem 5.6, if f has range whose convex hull does not contain the complex zero (if f is positive), then, according to Definition 8.1, the matrix sequence $\{\Phi_n^{-1}(f)T_n(f)\}$ is spectrally clustered to one in the sense of the singular values (in the sense of the eigenvalues if $f>0$); moreover $\{\Phi_n^{-1}(f)T_n(f)\}$ and $\{T_n^{-1}(f)\Phi_n(f)\}$ are spectrally bounded.*

Proof. The assumptions on the range of f implies that $\{T_n^{-1}(f)\}$ is uniformly bounded (see [16]); the continuity of f implies that $f\in L^\infty(I)$ and therefore by (2) we have that also $\{T_n^{-1}(f)\}$ is uniformly bounded. From the representation formula of $\Phi_n(f)$ in Lemma 5.1, it follows that also $\{\Phi_n(f)\}$ and $\{\Phi_n^{-1}(f)\}$ are spectrally bounded: as a consequence both $\{\Phi_n^{-1}(f)T_n(f)\}$ and $\{T_n^{-1}(f)\Phi_n(f)\}$ are spectrally bounded. For the part concerning the spectral clustering at the unity see the proof of the second item of Theorems 8.4 and 8.5. •

Finally, we observe that if the assumption of Corollary 5.1 is fulfilled, then we have a superlinear PCG method (see [3] and Appendix **A**).

5.2.2 The LPO sequences related to $\{\Phi_n(\cdot)\}$

The behavior of the eigenvalues of $\Phi_n(f)$ is studied in this section. If U_n is completely generic not very much can be said, but under the assumption that U_n is the unitary matrix related to one of the special trigonometric algebras previously introduced (circulants, ω-circulants, Hartley class, sine/cosine matrix algebras), a richer analysis can be carried out.

Let $x_j^{(n)}$ be the j-th grid point of one of the considered trigonometric algebras (see the beginning of Section 5), then we define the operators ψ_n (from the continuous (2π)-periodic functions on I on itself) in a implicit way as follows: calling $\mathbf{u}_j$ the j-th column of the unitary matrix U_n, we have

$$(\psi_n(f))(x_j^{(n)})=\mathbf{u}_j^H T_n(f)\mathbf{u}_j,\quad f\in C_p(I).$$

From the definition, it is obvious that $(\psi_n(f))(x_j^{(n)})$ is the j-th eigenvalue of the Frobenius optimal approximation $\Phi_n(f)$ of $T_n(f)$ according to the first item of Lemma 5.1.

The following lemma holds.

Lemma 5.3 *$\psi_n(\cdot)$ is a linear positive operator and $\Phi_n(\cdot)$ is also a LPO in the matrix sense (see Definition 5.2).*

Proof The linearity of both the operators follows from the second item of Lemma 5.1. Fix $f \geq 0$, then $T_n(f)$ is nonnegative definite by the first item of Theorem 2.2 and therefore every Rayleigh quotient is nonnegative and particularly the one giving rise to $(\psi_n(f))(x_j^{(n)})$. Moreover, the latter implies that also $\Phi_n(f)$ is nonnegative definite (see Lemma 5.2). •

Now we resort to the Korovkin Theorem to establish whether the eigenvalues of $\{\Phi_n(f)\}$ tend to $\{\{f(x_j^{(n)})\}\}$ for n going to infinity.

On the other hand, this fact is implied by the asymptotic discrete convergence of $\{\psi_n(g)\}$ to g (see (37)) for the three test functions reported in the Korovkin Theorem 5.4.

Theorem 5.8 *Let f be a continuous (2π)-periodic function and let $\{\Phi_n(f)\}$ the sequence of Frobenius optimal approximations of $T_n(f)$ in the sequence of algebras $\{\mathcal{A}_n\}$ with quasi-uniform grid sequence $\{W_n\}$ in $I = [-\pi, \pi)$. If $\psi_n(g) = g + \epsilon_n(g)$ with ϵ_n going uniformly to zero on the grid sequence $\{W_n\}$ in the sense of equation (37), then $\{\Phi_n(f)\}$ converges to $\{T_n(f)\}$ in the weak sense.*

Proof From identity **d** in Lemma 5.1, for any polynomial p we have

$$0 \leq \|T_n(p) - \Phi_n(p)\|_F^2 = \|T_n(p)\|_F^2 - \|\Phi_n(p)\|_F^2.$$

By Theorem 5.5, from the asymptotic discrete convergence of $\{\psi_n(g)\}$ to g on the test functions we obtain the same convergence property for any polynomial p of fixed degree, i.e., $\psi_n(p) = p + \epsilon_n(p)$ on the grids W_n and

$$\lim_{n\to\infty} \|\epsilon_n(p)\|_{\infty,W_n} = \lim_{n\to\infty} \sup_{x\in W_n} |(\epsilon_n(p))(x)| = 0. \tag{43}$$

Therefore

$$\|T_n(p) - \Phi_n(p)\|_F^2 = \|T_n(p)\|_F^2 - \sum_{j=0}^{n-1} |p(x_j^{(n)}) + \epsilon_n(p)(x_j^{(n)})|^2.$$

Now, from the definition of the Frobenius norm, we find that

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

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 (73) and Definition 9.1 with $F(z) = z^2$ if $z \in \mathcal{ER}(p)$). Then, by applying the quoted result, we find

$$\|T_n(p)\|_F^2 = n \cdot \frac{1}{2\pi} \int_{-\pi}^{\pi} |p|^2 + o(n). \tag{44}$$

In addition, by exploiting the convergence of $\{\psi_n(p)\}$ to p i.e. relation (43), we conclude that

$$\sum_{j=0}^{n-1} |p(x_j^{(n)}) + \epsilon_n(p)(x_j^{(n)})|^2 = \sum_{j=0}^{n-1} |p|^2(x_j^{(n)}) + o(n). \tag{45}$$

Thus, by virtue of the quasi-uniform distribution of grid points $W_n = \{x_k^{(n)}\}$, we infer

$$\sum_{j=0}^{n-1} |p(x_j^{(n)}) + \epsilon_n(p)(x_j^{(n)})|^2 = n \cdot \frac{1}{2\pi} \int_{-\pi}^{\pi} |p|^2 + o(n). \tag{46}$$

The combination of equations (44) and (46), in the light of the powerful Theorem 8.5, allows one to state the weak convergence of $\{\Phi_n(p)\}$ to $\{T_n(p)\}$. But, by noticing that this is the assumption of the second Weierstrass type Theorem 5.7, the theorem is proved. •

Theorem 5.9 *Under the same assumption of the previous Theorem 5.8, if* $\|\epsilon_n(g)\|_{\infty,W_n} = O(n^{-1})$ *for the three test functions* g *and if the grid points* $\{W_n\}$ *of the algebra are uniformly distributed in* $I = [-\pi, \pi)$ *in the sense of Definition 5.1, then* $\{\Phi_n(f)\}$ *converges to* $\{T_n(f)\}$ *in the strong sense.*

Proof We follow the same proof given in Theorem 5.8. In particular, in all the equations (44), (45) and (46) the terms $o(n)$ are replaced by terms of constant order. In equation (44), we notice that for all the polynomials p we have $\|T_n(f\|_F^2 - \frac{n}{2\pi}\int_{-\pi}^{\pi}|p|^2 = O(1)$ (see also [100]). For the relation (45), the hypothesis on $\epsilon_n(g)$ with g test function and Theorem 5.5 are used while, for equation (46), we need the uniform distribution instead of the quasi-uniform one. Finally, Theorem 8.5 and the first Weierstrass type Theorem 5.6 are invoked. •

Finally we just recall these Korovkin style results are very general and indeed the weak convergence of $\{T_n(f) - \Phi_n(f)\}$ can be proven for $f \in L^1(I)$ just by verifying the Korovkin test.

5.2.3 Verification of the Korovkin test

We provide a verification of the Korovkin test in the case of the circulant algebra whose grid sequence $\{W_n\}$ is uniform since it is perfectly equispaced ($W_n = \left\{x_k^{(n)} = \frac{2k\pi}{n} : k = 0, \ldots, n-1\right\}$).

Consider the assumptions of Theorems 5.8 and 5.9: we have only to prove that

$$\lim_{n\to\infty} \sup_{x\in W_n} |(\psi_n(g))(x) - g(x)| = 0, \quad \forall g \in T = \{1, e^{\pm \mathbf{i}x}\}$$

by estimating the convergence rate to zero. For $g = 1$, we have $T_n(g) = I_n$ and therefore $\Phi_n(g) = I_n$ so that $\psi_n(1) \equiv 1$ and there is nothing to prove; for $g(x) = e^{-\mathbf{i}x}$ we deduce that $T_n(g)$ is a Jordan block ($(T_n(g))_{j,k} = a_{j-k} = 1$ if $j - k = -1$ and zero otherwise): therefore by formula (40), we have

$$\Phi_n(g) = (1 - 1/n)Z_1 = (1 - 1/n)\begin{pmatrix} 0 & 1 & 0 & \ldots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & 0 \\ 0 & & & \ddots & 1 \\ 1 & 0 & \ldots & \ldots & 0 \end{pmatrix},$$

where Z_1 is the generator of the circulant algebra. The eigenvalues of Z_1 have been explicitly computed in Subsection 4.1.1 and consequently

$$(\psi_n(g))(x_j^{(n)}) = (1 - 1/n)e^{-\mathbf{i}x_j^{(n)}}$$

so that

$$\sup_{x\in W_n} |(\psi_n(g))(x) - g(x)| = n^{-1}.$$

Now $T_n(e^{\mathbf{i}x}\}) = T_n^H(e^{-\mathbf{i}x}\}$ and therefore by the third item of Lemma 5.1, we have $\Phi_n(e^{\mathbf{i}x}) = \Phi_n^H(e^{-\mathbf{i}x})$ so that the eigenvalues of $\Phi_n(e^{\mathbf{i}x})$ are the conjugated of those of $\Phi_n^H(e^{-\mathbf{i}x}\}$. Consequently for for $g(x) = e^{\mathbf{i}x}$ we have

$$(\psi_n(g))(x_j^{(n)}) = (1 - 1/n)e^{\mathbf{i}x_j^{(n)}}$$

n	$\tau_{n,\mathrm{nat}}$	$\mathcal{C}_{n,\mathrm{nat}}$	$\Phi_{n,\tau}$	$\Phi_{n,\mathcal{C}}$
32	3 (3)	4 (5)	3 (3)	4 (5)
64	3 (3)	4 (5)	3 (3)	4 (5)
128	3 (3)	4 (5)	3 (3)	4 (5)
256	3 (3)	4 (4)	3 (3)	4 (4)
512	3 (3)	4 (4)	3 (3)	4 (4)

Table 14: Number of PCG steps in the case of f_α with $\alpha = 5 + \pi^2/6$.

so that

$$\sup_{x \in W_n} |(\psi_n(g))(x) - g(x)| = n^{-1}.$$

Therefore by Theorem 5.9, $\{\Phi_n(f)\}$ converges to $\{T_n(f)\}$ in the strong sense i.e. $\{T_n(f) - \Phi_n(f)\}$ is properly clustered at zero: now if $T_n(f)$ is uniformly bounded with its inverse (=the convex hull of its range does not contain zero), then by Corollary 5.1, $\{\Phi_n^{-1}(f)T_n(f)\}$ is clustered at one and is bounded with its inverse: by Theorem 8.3, the combination of the latter three properties means that the considered PCG is optimal and the quality of the convergence is superlinear.

The same very elementary computations have been worked out (see [71, 32]) for all the circulant-like algebras (Hartley and ω-circulants) and for all the 8 cosine/sine matrix algebras at the beginning of Section 5: the conclusions are identical. All the related Frobenius optimal approximations are superlinear preconditioners for well-conditioned Toeplitz matrices $T_n(f)$ (for ill-conditioned Toeplitz sequences we lose this good approximation property as established in [94, 31] and as observed in Subsection 6.3.1).

Moreover, if a new sequence of algebras will appear in the literature, it is enough to identify its grid sequence $\{W_n\}$ and to perform the same elementary Korovkin test.

5.2.4 Numerical experiments

Here we discuss few numerical experiments where we consider two kind of test functions: $f_\alpha = x^2/2 + \alpha$ and $h_\alpha = ((x/\pi)^2 - 1)^2 + \alpha$. Moreover we consider two different types of data vectors $\mathbf{b}$. The first is made up by all ones. The second is a randomly generated one. In all the subsequent tables the numbers between parentheses are related to the number of PCG iterations when a random data vector $\mathbf{b}$ is considered. The stopping criterion is given by the relative two-norm of the residual less than 10^{-7}. All the experiments are done by using MATLAB.

For the choice of the preconditioners we considered two algebras, namely the τ class (trigonometric algebra related to the transform DST I [10]) and the circulants. In the PCG algorithm we use the *natural preconditioners* or Strang type $\mathcal{C}_{n,\mathrm{nat}}$ and $\tau_{n,\mathrm{nat}}$ whose related approximation process is given by the Fourier polynomial of degree $n/2$ and n respectively, and the *Frobenius optimal preconditioners* $\Phi_{n,\mathcal{C}}$ and $\Phi_{n,\tau}$ whose approximation processes have the same asymptotical behavior as the Cesaro sum.

Concerning the results displayed in Tables 14, 15 and 16 two remarks are needed. First we notice that the number of iterations for the τ preconditioners is generally slightly less than in the circulant case. This agrees with the results of [68] where it is proved that the border conditions are slightly heavier in the circulant case: more precisely, if p is an even real valued polynomial of fixed degree and if we consider the quantities $\mathrm{rank}(T_n(p) - \mathcal{A}_n(p))$ (see (33)) when $\mathcal{A}_n$ is either the circulant or the τ algebra, then we observe a smaller number in τ case. Moreover, for an even polynomial of fixed degree, $\mathcal{A}_n(p)$ coincides with $\tau_{n,\mathrm{nat}}$ in the case where $\mathcal{A}_n$ is τ algebra and $\mathcal{A}_n(p)$ coincides with $\mathcal{C}_{n,\mathrm{nat}}$

n	$\tau_{n,\mathrm{nat}}$	$\mathcal{C}_{n,\mathrm{nat}}$	$\Phi_{n,\tau}$	$\Phi_{n,\mathcal{C}}$
32	3 (3)	3 (4)	3 (3)	4 (5)
64	3 (3)	3 (4)	3 (3)	4 (5)
128	3 (3)	3 (4)	3 (3)	3 (4)
256	3 (3)	3 (4)	3 (3)	3 (4)
512	2 (3)	3 (4)	3 (3)	3 (4)

Table 15: Number of PCG steps in the case of h_α with $\alpha = 1$.

n	$\tau_{n,\mathrm{nat}}$	$\mathcal{C}_{n,\mathrm{nat}}$	$\Phi_{n,\tau}$	$\Phi_{n,\mathcal{C}}$
32	3 (3)	4 (5)	3 (4)	7 (10)
64	3 (3)	4 (5)	3 (4)	7 (10)
128	3 (3)	4 (5)	3 (4)	6 (9)
256	3 (3)	4 (5)	3 (4)	5 (8)
512	2 (3)	3 (4)	3 (3)	5 (6)

Table 16: Number of PCG steps in the case of h_α with $\alpha = 0.01$.

in the case where $\mathcal{A}_n$ is circulant algebra. Secondly if we observe the behavior of the preconditioners $\tau_{n,\mathrm{nat}}$ and $\Phi_{n,\tau}$ we conclude that their behavior is substantially identical in particular for bigger dimensions. Nevertheless they are characterized by two very different approximation processes, one substantially faster than the other. This insensitivity to the convergence rate of the approximation process fully agrees with the analysis provided in [68]. The same remark holds true if we analyze the behavior of the preconditioners $\mathcal{C}_{n,\mathrm{nat}}$ and $\Phi_{n,\mathcal{C}}$.

On the other hand, the convergence speed of the associated approximation process plays a crucial role when the generating function of the Toeplitz matrix has zeros (see [31]).

6 The multilevel case

We consider the solution of large linear systems where the coefficient matrices have a multilevel Toeplitz structure. We recall that this kind of matrices arise in different applications in several fields (see e.g. [17, 21, 37]) for which efficient strategies for the solution of very large systems are often required: among them, we recall Markov chains, some integral equations in imaging and the numerical solution by means of finite differences of certain PDEs.

In this section we study the generalization of the preconditioning techniques introduced in the preceding section in the multilevel case. We stress that what is generalizable in the two-level context is generalizable to the d-level context with $d > 2$: therefore for the sake of notational simplicity we provide an informal discussion in the two level case (often called BTTB=block Toeplitz with Toeplitz blocks) and then we state the results in full generality.

Let us consider the solution of a linear system

$$T_{n_1,n_2}\mathbf{x} = \mathbf{b}$$

in the case where T_{n_1,n_2} has a block Toeplitz structure, i.e.,

$$T_{n,m} = \begin{bmatrix} A_0 & A_{-1} & \cdots & A_{-(n_1-1)} \\ A_1 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & A_{-1} \\ A_{n_1-1} & \cdots & A_1 & A_0 \end{bmatrix} \tag{47}$$

with

$$A_j = \begin{bmatrix} a_{j,0} & a_{j,-1} & \cdots & a_{j,-(n_2-1)} \\ a_{j,1} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & a_{j,-1} \\ a_{j,n_2-1} & \cdots & a_{j,1} & a_{j,0} \end{bmatrix}. \tag{48}$$

If (j,k) indicates the block in the matrix T_{n_1,n_2} and (p,q) the position of the entry in the block, then $(T_{n_1,n_2})_{(j,k)(p,q)} = a_{k-j,q-p}$ for $j,k = 0,1,\ldots,n_1-1,\ p,q = 0,1,\ldots,n_2-1$.

As in the one dimensional setting, we consider the case where $\{T_{n_1,n_2}\}$, $n_1,n_2 \in \mathbf{N}$, is related to the Fourier coefficients $a_{k,q}$ of an assigned (2π)-periodic function $f : I^2 \to \mathbf{C}$, $I = [-\pi,\pi)$, periodically extended on $\mathbf{R}^2$, that is,

$$a_{k,q} = \frac{1}{4\pi^2}\int_{-\pi}^{\pi}\int_{-\pi}^{\pi} f(x_1,x_2)e^{-i(kx_1+qx_2)}dx_1dx_2, \quad \mathbf{i}^2 = -1.$$

It is a simple matter to verify that $T_{n_1,n_2}(f)$ is Hermitian if the function f is real valued and it real block symmetric with symmetric blocks (also called quadrantally symmetric) if the function f is real valued with $f(x_1,x_2) = f(|x_1|,|x_2|)$ for every $(x_1,x_2) \in I^2$.

Now we present an example of a two-level Toeplitz problem and we explain the extensions of PCG techniques in Section 4 to the case of a given two-level structure.

We consider the matrix $T_{n_1,n_2}(f)$ with $f(x_1,x_2) = x_1^2 + x_2^2$ which appears in the Sinc Galerkin discretization of the Poisson equation on a rectangle [52]; this matrix is dense and can be written as $T_{n_1}(h) \otimes I_{n_2} + I_{n_1} \otimes T_{n_2}(h)$ with $h(x) = x^2$ and where the Fourier coefficients of h are such that $a_0 = \frac{\pi^3}{2}$ and $a_k = \frac{2(-1)^k}{k^2}$, $k \in \mathbf{Z}$.

The function f is nonnegative with $\inf_{I^2} f = 0$ and $\sup_{I^2} f = 2\pi^2$: therefore the eigenvalues of $T_{n_1,n_2}(f)$ belong to the interval $(0,2\pi^2)$ (see item **a** in Theorem 6.1). Moreover, as n_1 and n_2 tend to infinity, the minimal eigenvalue tends to 0 and the maximal one tends to $2\pi^2$ by item **b** of Theorem 6.1: from this we know that the sequence $\{T_{n_1,n_2}(f)\}$ is asymptotically ill-conditioned. It is possible (in analogy to the scalar case) to have a precise estimate of the asymptotic growth of the spectral condition number: since f has a zero of order 2, by item **c** of Theorem 6.1, we deduce that the minimal eigenvalue goes to zero as ν^{-2} with $n_1 \sim n_2 \sim \nu$ so that the asymptotic ill-conditioning is of order ν^2.

The main idea of band Toeplitz preconditioning is to find a simpler function g (e.g. a trigonometric polynomial which generates a band Toeplitz sequence) which erases the zeros of f and such that f/g is bounded: as observed in Subsection 4.3, the smaller is the ratio $\left(\inf_{I^2} f/g\right)^{-1} \sup_{I^2} f/g$ the faster is the associated PCG converges.

In this case a good choice is $g(x_1,x_2) = 4\sin^2(x_1/2) + 4\sin^2(x_2/2)$ whose Toeplitz matrix is the 2D discrete Laplace operator

$$\begin{aligned} T_{n_1,n_2}(g) &= [-1,2,-1]_{n_1} \otimes I_{n_2} + I_{n_1} \otimes [-1,2,-1]_{n_2} \\ &= \begin{bmatrix} B & -I_{n_2} & & \\ -I_{n_2} & \ddots & \ddots & \\ & \ddots & \ddots & -I_{n_2} \\ & & -I_{n_2} & B \end{bmatrix}, \end{aligned}$$

with

$$B = T_{n_2}(4 - 2\cos(x)) = \begin{bmatrix} 4 & -1 & & \\ -1 & \ddots & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 4 \end{bmatrix}.$$

The function f/g has minimum equal to 1 and maximum given by $\pi^2/2$: therefore, by Theorem 6.2, we infer that all the eigenvalues lie in $(1, \pi^2/2)$, are equally distributed in this interval, and the extreme eigenvalues converge to 1 and to $\pi^2/2$ respectively.

Therefore, by Theorem 8.3, we deduce that the associated PCG is optimal and, by the equal distribution, the convergence estimates given in that theorem are tight.

We observe that matrix algebra preconditioners (see e.g. [21]) do not ensure a condition number bounded by a constant independent of n_1 and n_2 due to the negative results (see [84, 77, 53]) briefly discussed in Subsection 6.3.

In conclusion, the proposed preconditioner has the following features:

1) $T_{n_1,n_2}(g)$ is a band block Toeplitz matrix having band Toeplitz blocks.

2) The bandwidth of each block, as well as the block bandwidth of $T_{n_1,n_2}(g)$ is independent of n_1 and n_2.

3) $T_{n_1,n_2}(g)$ is positive definite.

4) $T_{n_1,n_2}^{-1/2}(g)T_{n_1,n_2}(f)T_{n_1,n_2}^{-1/2}(g)$ has a condition number independent of n_1 and n_2.

Every step of the PCG method requires the solution of a system of the type $T_{n_1,n_2}(g)\mathbf{y} = \mathbf{b}$, which can be obtained by a band solver (see e.g. [36]) adapted to block matrices: more in detail, we can use block Gaussian elimination performing $O(n_1 n_2^3)$ arithmetic ops or equivalently a scalar band Gaussian elimination with cost of $O(n_1 n_2^3)$ arithmetic ops. This technique does not use the band structure of the blocks; actually, during the Gaussian algorithm, we lose the band Toeplitz structure of the inner blocks. An alternative strategy to solve the former linear system uses special decompositions in suitable block matrix algebras [61]. However the more promising idea is the use of an algebraic multigrid method; in [35] (by generalizing some results in [39, 40]), it is shown that, in practice, the arithmetic cost of the solution of our double banded system is of order of $n_1 n_2$, that is, linear with respect to the dimension of the involved matrix. Moreover the parallel cost is of $O(\log(n_1 n_2))$ parallel steps. To conclude, since the product of $T_{n_1,n_2}(f)$ by a vector can be calculated by means of bivariate FFTs performing $O(n_1 n_2 \log(n_1 n_2))$ arithmetic ops and $O(\log(n_1 n_2))$ parallel steps (a generalization of the embedding technique shown in Subsection 4.1.1), we have that the global sequential and parallel costs of the considered PCG method is of $O(n_1 n_2 \log(n_1 n_2))$ arithmetic ops and $O(\log(n_1 n_2))$ parallel steps, respectively.

More generally, Algorithm 4.1 can be extended to the block case with $f(x_1, x_2) \geq 0$ a.e. on I^2:

Algorithm 6($f \geq 0$, zeros="algebraic curves")

Input. The zeros of f, $T_{n_1,n_2}(f)$, $\mathbf{b}$.

Step 1. Find the bivariate trigonometric polynomial $g(x_1, x_2)$ such that $0 < r < \frac{f}{g} < R < \infty$ a.e..

Step 2. Apply the PCG method to the system $T_{n_1,n_2}(f)\mathbf{x} = \mathbf{b}$ with $T_{n_1,n_2}(g)$ as preconditioner.

At step 2. there is a difficulty. In order to find g such that $0 < r < \frac{f}{g} < R < \infty$ a.e., it is necessary that $\mathcal{L} = \{(x, y) : f(x, y) = 0\}$ can be expressed as a finite collection of algebraic curves (with respect to trigonometric polynomials). Otherwise the problem has no solution. In the special case where $\mathcal{L}$ is a collection of isolated points and f is smooth enough, by using Taylor's series, it is possible to construct explicitly the polynomial g (see [61]). In other situations also an approximate solution is feasible: according to the boundary layer effect (see [79]), if the distance between $\mathcal{L}$ and its approximation is $O(n_1^{-1} + n_2^{-1})$ and the order of the zeros is correct, then we have no deterioration of the performances of the approximate preconditioner with respect to the theoretical one (see [79, 54]).

In the next subsections, we give more connections between the spectral properties of $T_n(f)$ and the function f and we discuss preconditioning strategies along the same lines followed for scalar Toeplitz matrices.

6.1 Generalizable results

Let f be a d variate (2π)-periodic real valued (Lebesgue) integrable function, defined over the hypercube I^d, with $I = [-\pi, \pi)$ and $d \geq 1$. From the Fourier coefficients of f

$$a_j = \frac{1}{(2\pi)^d} \int_{I^d} f(x) e^{-\mathbf{i}(j,x)}\, dx, \qquad \mathbf{i}^2 = -1, \quad j = (j_1, \ldots, j_d) \in \mathbf{Z}^d \tag{49}$$

with $x = (x_1, \ldots, x_d)$, $(j, x) = \sum_{k=1}^{d} j_k x_k$, $n = (n_1, \ldots, n_d)$ and $N(n) = n_1 \cdots n_d$, we can build the sequence of Toeplitz matrices $\{T_n(f)\}$, where $T_n(f) = \{a_{j-i}\}_{i,j=e^T}^{n} \in M_{N(n)}(\mathbf{C})$, $e^T = (1, \ldots, 1) \in \mathbf{N}^d$ is said to be the Toeplitz matrix of order n generated by f (see [95]). Furthermore, throughout these notes when we write $n \to \infty$ with $n = (n_1, \ldots, n_d)$ being a multi-index, we mean that $\min_{1 \leq j \leq d} n_j \to \infty$.

Very shortly, we can say that every spectral property of Toeplitz sequences $\{T_n(f)\}$, $n = (n_1, \ldots, n_d)$ and of preconditioned Toeplitz sequences $\{T_n^{-1}(g)T_n(f)\}$ with g nonnegative and not identically zero, stated in Sections 2 and 3 for the one level case can be generalized to the d-level setting with $d \geq 2$.

Concerning Toeplitz sequences the following result is true.

Theorem 6.1 *Let f be integrable and real valued over I^d, $I = [-\pi, \pi)$. Let $n = (n_1, \ldots, n_d)$, $N(n) = n_1 \cdots n_d$, and let us order the eigenvalues $\lambda_j^{(n)}$ of $T_n(f)$ in nondecreasing way. Let m_f and M_f be the essential infimum and the essential supremum of f ; the following relations hold.*

a. *$T_n(f)$ has eigenvalues in the open set (m_f, M_f) if $m_f < M_f$ and it coincides with $m_f I_{N(n)}$ if $m_f = M_f$.*

b. *The extreme eigenvalues of $T_n(f)$ are such that $\lim_{n\to\infty} \lambda_1^{(n)} = m_f$, $\lim_{n\to\infty} \lambda_n^{(n)} = M_f$, where $n \to \infty$ means that every n_j, $j = 1, \ldots, d$, diverges to infinity.*

c. *If $f - m_f$ has essential zeros that can be expressed as a finite collection of smooth k-dimensional manifold with $k \leq d - 1$ with maximal order α and if $n_j \sim \nu$, $j = 1, \ldots, d$, then $\lambda_1^{(n)} - m_f \sim \nu^{-\alpha}$.*

d. *Let $C_{\text{blimits}} = \{F : \mathbf{R} \to \mathbf{R},\ F$ continuos and with finite limits at $\pm\infty\}$; then for every $F \in C_{\text{blimits}}$ we have*

$$\lim_{n\to\infty} \frac{1}{N(n)} \sum_{j=1}^{N(n)} F\left(\lambda_j^{(n)}\right) = \frac{1}{(2\pi)^d} \int_{I^d} F\left(f(x)\right) dx. \tag{50}$$

Proof. All the statements can be proven as in the unilevel case (with minor changes). We explicitly give a proof of parts **a** and **b** by using a different technique: more precisely, for part **b** in place of using the Szegö distributional result we provide a direct proof (from [61]) in the case where f is continuous. We restrict the attention to the two-level case for notational simplicity, the d-level case being analogous.
Since f is real valued $T_{n_1,n_2}(f)$ is Hermitian and consequently its eigenvalues are real. To prove item **a** we consider the quadratic form $\mathbf{u}^H T_{n_1,n_2}(f)\mathbf{u}$, where $\mathbf{u} \in \mathbf{C}^{n_1 n_2}$ is normalized with respect to the Euclidean norm, i.e., $\mathbf{u}^H\mathbf{u} = 1$. By direct computation (see the beginning of Theorem 2.2), we arrive at the following relation:

$$\mathbf{u}^H T_{n_1,n_2}(f)\mathbf{u} = \frac{1}{4\pi^2}\int_{I^2} f(x_1,x_2)\left|\sum_{j=0}^{n_1-1}\sum_{p=0}^{n_2-1}(\mathbf{u}_j)_p e^{\mathbf{i}(jx_1+px_2)}\right|^2 dx_1dx_2,$$

where $\mathbf{u} = (\mathbf{u}_0,\ldots,\mathbf{u}_{n_1-1}),\ \mathbf{u}_j \in \mathbf{C}^{n_2}$. Therefore by the mean integral theorem we find that

$$\begin{aligned}\mathbf{u}^H T_{n_1,n_2}(f)\mathbf{u} &= f(\eta,\nu)\frac{1}{4\pi^2}\int_{I^2}\left|\sum_{j=0}^{n_1-1}\sum_{p=0}^{n_2-1}(\mathbf{u}_j)_p e^{\mathbf{i}(jx_1+qx_2)}\right|^2 dx_1dx_2\\ &= f(\eta,\nu)\mathbf{u}^H\mathbf{u} = f(\eta,\nu).\end{aligned}$$

Consequently $\mathbf{u}^H T_{n_1,n_2}(f)\mathbf{u} \in [m_f, M_f]$. Now if $m_f = M_f$, then $T_{n_1,n_2}(f) = m_f I_{n_1 n_2}$; otherwise the proof that $\lambda_j^{(n_1,n_2)} \in (m_f, M_f)$ can be done in the same way as in Theorem 2.2 since a nonzero polynomial

$$\left|\sum_{j=0}^{n_1-1}\sum_{p=0}^{n_2-1}(\mathbf{u}_j)_p e^{\mathbf{i}(jx_1+qx_2)}\right|^2$$

can vanish only on a zero measure set in the two variate case too.

Now we have to prove that the smallest eigenvalue tends to m_f and the greatest one tends to M_f. Since $T_{n_1,n_2}(f)$ is a leading submatrix of $T_{N_1,N_2}(f)$, $n_1 \le N_1, n_2 \le N_2$ then $\lambda_1^{(n_1,n_2)}$ is a nondecreasing sequence and consequently $\lim_{n_1,n_2\to\infty}\lambda_1^{(n_1,n_2)} = m \ge m_f$. By contradiction we suppose that $m > m_f$ and therefore, for any $\theta > 0$ such that $m_f + \theta < m$, the matrix $T_{n_1,n_2}(f) - (m_f+\theta)I$ is positive definite. In other words, for any n_1n_2–dimensional and unitary vector $\mathbf{u}$, setting $z_1 = e^{\mathbf{i}x_1},\ z_2 = e^{\mathbf{i}x_2}$, we have

$$0 < \mathbf{u}^H(T_{n_1,n_2}(f) - (m_f+\theta)\mathbf{u} = \frac{1}{4\pi^2}\int_{I^2}|p(z_1,z_2)|^2(f(x_1,x_2) - (m_f+\theta)dx_1dx_2$$

where $p(z_1,z_2)$ is the bivariate polynomial associated with $\mathbf{u}$.
Let $A_\theta = \{(x_1,x_2)\in I^2 :\ f(x_1,x_2) - (m_f+\theta) < 0\}$; we observe that $m(A_\theta) > 0$, otherwise, if $m(A_\theta) = 0$ then it follows that $m_f + \theta \le$essinf $f = m_f$ which is a contradiction. Now we consider $w^0 = (x_1^0, x_2^0)$ in the interior part of I^2 such that $m(A_\theta \cap J(w^0,\delta)) > 0$, $J(w^0,\delta)$ being the open ball of center w^0 and radius δ, and we define the continuous function $h(x_1,x_2) = h_\delta(x_1 - x_1^0)h_\delta(x_2 - x_2^0)$ where

$$h_\delta(s) = \begin{cases} 1 - \frac{|s|}{\delta} & \text{if } |s| \le \delta,\\ 0 & \text{otherwise}\end{cases}$$

By the approximation Theorem 1.9 (c) of [37], for any positive $\epsilon > 0$ there exist two nonnegative trigonometric polynomials $s(x_1), t(x_2)$ such that $\|h_\delta(x_1 - x_1^0) - s(x_1)\|_\infty < \epsilon$, $\|h_\delta(x_2 - x_2^0) -$

$t(x_2)\|_\infty < \epsilon$, where $\|h\|_\infty$ is defined as $\sup_x |h(x)|$. Moreover, by the representation Theorem 1.1.2 of [37], any nonnegative trigonometric polynomial can be seen as the square of the absolute value of a complex ordinary polynomial. Thus, $s(x_1) = |\alpha_1(e^{\mathbf{i}x_1})|^2$, $t(x_2) = |\alpha_2(e^{\mathbf{i}x_2})|^2\|$ and if $\mathbf{u}^*$ is the vector of $\mathcal{C}^{n_1^* n_2^*}$ related to $\alpha^* = \alpha_1(e^{\mathbf{i}x_1})\alpha_2(e^{\mathbf{i}x_2})$, setting $\theta(x_1, x_2) = f(x_1, x_2) - (m_f + \theta)$, then we have

$$\begin{aligned} 0 &< 4\pi^2 \mathbf{u}^{*H}(T_{n_1,n_2}(f) - (m_f + \theta))\mathbf{u}^* = \\ &\leq \epsilon C + \int_{A_\theta \cap J(w^0,\delta)} h(x_1, x_2)\theta(x_1, x_2)dx_1 dx_2 = \epsilon C + k(\theta), \end{aligned}$$

where C is a positive constant; whence it follows

$$\lim_{\epsilon \to 0} \epsilon C + k(\theta) = \int_Q h(x_1, x_2)\theta(x_1, x_2)dx_1 dx_2 < 0$$

which is a contradiction. •

Theorem 6.2 *Let f and g be integrable functions over I^d and let us suppose that g is nonnegative and not identically zero. Let $n = (n_1, \ldots, n_2)$, $N(n) = n_1 \cdots n_d$, and let us order the eigenvalues $\lambda_j^{(n)}$ of $G_n = T_n^{-1}(g)T_n(f)$ in nondecreasing way and let r and R be the essential infimum and the essential supremum of f/g ; the following relations hold.*

a. *G_n has eigenvalues in the open set (r, R) if $r < R$ and it coincides with $rI_{N(n)}$ if $r = R$.*

b. *If $m\{x \in I^d : g(x) = 0\} = 0$ then $\bigcup_{n=e}^{\infty} \bigcup_{j \leq N(n)} \lambda_j^{(n)}$ is dense in $\mathcal{ER}(f/g)$ where $\mathcal{ER}(f/g)$ is the essential range of f/g (we recall that $y \in \mathcal{ER}(h)$ if and only if for any $\epsilon > 0$ the Lebesgue measure of the set $\{x \in I^d : h(x) \in (y - \epsilon, y + \epsilon)\}$ is positive).*

c. *The extreme eigenvalues of G_n are such that $\lim_{n\to\infty} \lambda_1^{(n)} = r$, $\lim_{n\to\infty} \lambda_n^{(n)} = R$.*

d. *Let $C_{\text{blimits}} = \{F : \mathbf{R} \to \mathbf{R},\ F \text{ continuos and with finite limits at } \pm\infty\}$; then for every $F \in C_{\text{blimits}}$ we have*

$$\lim_{n\to\infty} \frac{1}{N(n)} \sum_{j=1}^{N(n)} F\left(\lambda_j^{(n)}\right) = \frac{1}{(2\pi)^d} \int_{I^d} F\left(f(x)/g(x)\right) dx. \tag{51}$$

The latter result gives a complete picture for the multilevel band Toeplitz preconditioning. Now we briefly consider the multilevel generalization of the Frobenius optimal approximation in matrix algebras. We consider again (unitary) matrix algebras $\mathcal{A}_N$ as in (31) with the right dimension $N = N(n)$, $n = (n_1, \ldots, n_d)$. A special case of interest is the case of multilevel matrix algebras where the unitary matrix U_n is of the form $U_{n_1}^{(1)} \otimes \cdots \otimes U_{n_1}^{(d)}$ where typically $U_\nu^{(j)}$ is one of the unitary transforms related to $\nu \log(\nu)$ algorithms considered at the beginning of in Section 5. As an example, for $d = 2$, $U_{n_1}^{(1)} = F_{n_1}$, and $U_{n_2}^{(2)} = H_{n_2}$, we have a two-level matrix algebra whose external structure is circulant and whose internal structure is Hartley.

The grids $W_N = \{x_1^{(N)}, \ldots, x_N^{(N)}\}$ now have size $N = N(n)$ and, in the case of multilevel algebras, are of the form

$$W_n = W_{n_1}^{(1)} \times \cdots \times W_{n_d}^{(d)}.$$

We say a grid sequence $\{W_N\}$ is quasi-uniformly distributed on $I = [-\pi, \pi)^d$ if for every trigonometric polynomial p

$$\lim_{n\to\infty} \frac{1}{N} \sum_{j=1}^{N} p\left(x_j^{(N)}\right) = \frac{1}{(2\pi)^d} \int_{I^d} p(x)dx.$$

The uniformity amounts to ask that the error

$$\left| \frac{1}{N} \sum_{j=1}^{N} p\left(x_j^{(N)}\right) - \frac{1}{(2\pi)^d} \int_{I^d} p(x)dx \right| = O(N^{-1}).$$

We notice that for $d = 1$ the new concepts coincide with those in Definition 5.1.

The good news is that the quasi-uniformity of every $\{W_\nu^{(j)}\}$ implies the quasi-uniformity of $\{W_n\}$, $n = (n_1, \ldots, n_d)$, and the uniformity of every $\{W_\nu^{(j)}\}$ implies the uniformity of $\{W_n\}$, $n = (n_1, \ldots, n_d)$.

With these changes, all the definitions and results stated in Subsection 5.2 have a natural generalization. More precisely, the definition of the Frobenius optimal approximation is the same as in (38) and (39) and the statements and proof of Lemmas 5.1, 5.2, 5.3, Corollary 5.1, Theorems 5.6, 5.7, 5.8, and 5.9 are practically unchanged. Moreover, the first item of Lemmas 5.1 can be completed as follows: if $\mathcal{A}_n$ is a multilevel algebra and if the symbol is separable, i.e., $f(x) = f_1(x_1)f_2(x_2)\cdots f_d(x_d)$ then

$$\Phi_n(f) = \Phi_{n_1}(f_1) \otimes \Phi_{n_2}(f_2) \otimes \cdots \otimes \Phi_{n_d}(f_d); \tag{52}$$

we observe that the latter property can be proved by using the representation formula contained in the part **a** of Lemma 5.1 and moreover that it is very useful when dealing with the multilevel Korovkin test since the functions in the Korovkin set T are all separable. The problem with the Korovkin theory is the verifications of the assumptions of the d-level version of Theorem 5.9 as discussed in the next subsection.

6.2 Not generalizable results

We report explicitly the multilevel version of Theorems 5.8 and 5.9, since we have to make some reasoning regarding its assumptions.

Theorem 6.3 *Let f be a continuous (2π)-periodic function, $n = (n_1, \ldots, n_d)$, $N = N(n) = n_1 \cdots n_d$, and let $\{\Phi_n(f)\}$ the sequence of Frobenius optimal approximations of $T_n(f)$ in the sequence of algebras $\{\mathcal{A}_n\}$ with quasi-uniform grid sequence $\{W_N\}$ in $I^d = [-\pi, \pi)^d$. If $\psi_n(g) = g + \epsilon_n(g)$ with ϵ_n going uniformly to zero on the grid sequence $\{W_n\}$ in the sense of equation (37), then $\{\Phi_n(f)\}$ converges to $\{T_n(f)\}$ in the weak sense.*

Theorem 6.4 *Under the same assumption of the previous Theorem 6.3, if $\|\epsilon_n(g)\|_{\infty,W_N} = O(N^{-1})$, $N = N(n)$, for the three test functions g and if the grid points $\{W_N\}$ of the algebra are uniformly distributed in $I^d = [-\pi, \pi)^d$ in the sense of Definition 5.1, then $\{\Phi_n(f)\}$ converges to $\{T_n(f)\}$ in the strong sense.*

Essentially, we observe that the uniformity of the grid sequence $\{W_N\}$ is guaranteed when we consider multilevel trigonometric algebras. Therefore the only thing which is not generalizable to the d-level version is *the convergence result on the Korovkin test*.

More precisely we cannot expect $\|\epsilon_n(g)\|_{\infty,W_N} = O(N^{-1})$, $N = N(n)$, and indeed for all the known matrix algebras of trigonometric and circulant-like type we find

$$\|\epsilon_n(g)\|_{\infty,W_n} = O\left(\sum_{j=1}^{d} n_j^{-1}\right),$$

which is enough for the weak convergence of $\{\Phi_n(f)\}$ to $\{T_n(f)\}$ but is is not enough for the strong convergence.

For having an evidence of this fact consider the case of $d = 2$ and the test function $g(x) = e^{\mathbf{i}x_1}$ whose Toeplitz matrix is given by

$$T_n(g) = Z_1 \otimes I_{n_2}$$

with Z_1 the generator of size n_1 of the circulant algebra. By (52) we deduce $\Phi_n(g) = \Phi_{n_1}(e^{\mathbf{i}x_1}) \otimes \Phi_{n_2}(1)$ and therefore, by exploiting the computation in Subsection 5.2.3, we have

$$\Phi_n(g) = (1 - 1/n_1)Z_1 \otimes I_{n_2}$$

whose eigenvalues are (see again Subsection 5.2.3)

$$(\psi_n(g))(x_j^{(n)}) = (1 - 1/n_1)e^{-\mathbf{i}x_j^{(n)}}, \quad \text{multiplicity } n_2$$

so that

$$\sup_{x \in W_n} |(\psi_n(g))(x) - g(x)| = n_1^{-1}$$

which is not $O(N^{-1})$ since $N = n_1 n_2$. In conclusion, all the assumptions of Theorem 6.3 are fulfilled while concerning Theorem 6.4 the hypothesis regarding the Korovkin test is not satisfied.

6.3 Advanced questions

We start by commenting the negative result contained in the last subsection: by using the Korovkin theory it is not possible to prove the proper clustering at the unity of the preconditioned matrices when using the Frobenius optimal approximation. Moreover, by making a more accurate analysis we observe that the number of outliers of $\{T_n(f) - \Phi_n(f)\}$ can be bounded by $cN(n)(\sum_{j=1}^{d} n_j^{-1})$ with c pure positive constant depending on the function f and on the sequence of algebras. This fact is confirmed in the numerical experiments where a number of outliers growing as $N(n)\left(\sum_{j=1}^{d} n_j^{-1}\right)$ can be observed.

The bad news is that this result, which is not satisfactory for $d \geq 2$, is the best we can obtain by using unitary algebras as stated in the following proposition (for these results see [84, 77, 85]).

Proposition 6.1 *If $d > 1$, $n = (n_1, \cdots, n_d)$ and $N = n_1 \cdots n_d$, then for every sequence of algebras $\mathcal{A} = \{\mathcal{A}_N\}$ with unitary transforms there exist infinitely many linearly independent d-variate trigonometric polynomials (and nonpolynomial functions) f such that for every $\{P_N\}$ with $P_N \in \mathcal{A}_N$ we observe that $\{T_n(f) - P_N\}$ is not properly clustered at zero in the sense of the singular values. Moreover the number of outliers grows at least as $cN(n)(\sum_{j=1}^{d} n_j^{-1})$ with $c > 0$ depending on f and on $\mathcal{A}$.*

As a consequence, we cannot have a proper clustering at the unity of $\{P_N^{-1}T_n(f)\}$ and therefore $\{P_N\}$ cannot be a superlinear preconditioning sequence for $\{T_n(f)\}$. Moreover, the results provided

by the Frobenius optimal approach are the best we can obtain for well conditioned Toeplitz sequence when using matrix algebras preconditioners.

This fact is enforced by recent results (see [53]) where it is proven that the search for essentially spectrally equivalent (up to a constant number of diverging eigenvalues) preconditioners cannot be successful in general (at least in the multilevel circulant and multilevel τ cases) when considering simple nonnegative polynomial symbols f with zeros. Here $T_n(f)$ is positive definite and therefore we restrict our attention to positive definite preconditioners.

More precisely the following statement is true: if $\{P_N^{-1}T_n(f)\}$ is spectrally bounded from below by a positive constant independent of n, then infinitely many eigenvalues of $\{P_N^{-1}T_n(f)\}$ tends to infinity as $n \to \infty$; conversely, if $\{P_N^{-1}T_n(f)\}$ is spectrally bounded from above by a positive constant independent of n, then infinitely many eigenvalues of $\{P_N^{-1}T_n(f)\}$ tends to zero as $n \to \infty$.

By taking into account the analysis in Subsection 6.1, we deduce that the aforementioned negative results represent an invitation for the research community to spend more attention on multilevel band Toeplitz preconditioning and on multigrid/multilevel methods which have been proven to be optimal even in the ill-conditioned multilevel case.

6.3.1 A two-level numerical evidence

We consider Example 5 in [50] where the two-level Toeplitz matrix $T_{n_1,n_2}(f)$ is characterized by $n_1 = n_2 = \nu$ and by the two-dimensional mask of the nonzero Fourier coefficients

$$[a_{j,k}]_{-2\le j,k\le 2} = (-1)\begin{bmatrix} 0.01 & 0.02 & 0.04 & 0.02 & 0.01 \\ 0.02 & 0.04 & 0.12 & 0.04 & 0.02 \\ 0.04 & 0.12 & -1 & 0.12 & 0.04 \\ 0.02 & 0.04 & 0.12 & 0.04 & 0.02 \\ 0.01 & 0.02 & 0.04 & 0.02 & 0.01 \end{bmatrix}. \tag{53}$$

We recall that masks of this nature arise in the discretization of constant-coefficients elliptic PDEs. When using the two-level circulant algebra, we observe that the Strang two-level preconditioner $\mathcal{C}_{\nu,\nu}$ is singular for this problem (see e.g. [20]): therefore in [50], in order to perform the preconditioning properly, $\mathcal{C}_{\nu,\nu}$ is slightly modified by replacing the unique zero eigenvalue with the smallest nonzero eigenvalue of $\mathcal{C}_{\nu,\nu}$.
In [50, 19], both the the modified Strang preconditioner and the Frobenius optimal approximation have been used and in both the cases a weak clustering at the unity for the preconditioned matrices is observed (according to Theorem 6.3). However, the number of outliers diverges to infinity as ν when ν tends to infinity (according to Proposition 6.1). Moreover, for the Frobenius optimal preconditioner, the condition numbers of $T_{\nu,\nu}(f)$ and $\Phi_{\nu,\nu}^{-1}(f)T_{\nu,\nu}(f)$ increase at the rate of $O(\nu^2)$ and $O(\nu)$, respectively, with the resulting cost of $O(\nu^{5/2}\log(\nu))$ arithmetic ops for the PCG with Frobenius optimal preconditioning. Similar results can be observed when the incomplete Cholesky factorization (see [26, 38, 48]) is used.

We recall that, because of the doubly band structure of $T_{\nu,\nu}(f)$, we require $O(\nu^4)$ arithmetic ops if we use of band solver.

Now we describe how to construct the generating function g of the preconditioner of Toeplitz band type according to the results in Section 6.1.

By using the information in (53) we find that

$$T_{\nu,\nu}(f) = I_\nu \otimes B + H \otimes C + K \otimes D,$$

where

$$B = \text{pentadiag}_\nu[-0.04, -0.12, 1, -0.12, -0.04],$$

$$
\begin{aligned}
C &= \text{pentadiag}_\nu[-0.02, -0.04, -0.12, -0.04, -0.02],\\
D &= \text{pentadiag}_\nu[-0.01, -0.02, -0.04, -0.02, -0.01],\\
H &= \text{tridiag}_\nu[1, 0, 1],\\
K &= \text{pentadiag}_\nu[1, 0, 0, 0, 1].
\end{aligned}
$$

The related bivariate generating function is

$$
\begin{aligned}
f(x_1, x_2) &= 1 - 0.24\cos(x_1) - 0.08\cos(2x_1) - 2\cos(x_2)[0.12\\
&\quad +0.08\cos(x_1) + 0.04\cos(2x_1)] - 2\cos(2x_2)[0.04\\
&\quad +0.04\cos(x_1) + 0.02\cos(2x_1)]
\end{aligned}
$$

which is nonnegative, vanishes in $(x_1, x_2) = (0, 0)$ and is strictly positive elsewhere. The Hessian of f calculated in $(0, 0)$ is a diagonal positive definite matrix and, therefore, in view of Theorem 6.2 we propose two preconditioners generated by two nonnegative functions:

$$
\begin{aligned}
g^{(1)}(x_1, x_2) &= 4 - \cos(x_1) - \cos(x_2),\\
g^{(2)}(x_1, x_2) &= 1.28(2 - 2\cos(x_1)) + 0.72(2 - 2\cos(x_2));
\end{aligned}
$$

so the proposed preconditioners are:

$$
\begin{aligned}
P_{(1)} = T_{\nu,\nu}(g^{(1)}) &= 4I_{\nu^2} - I_\nu \otimes H - H \otimes I_\nu,\\
P_{(2)} = T_{\nu,\nu}(g^{(2)}) &= 4I_{\nu^2} - 0.72(I_\nu \otimes H) - 1.28(H \otimes I_\nu).
\end{aligned}
$$

We have

$$
\begin{aligned}
\alpha_1 = \inf f/p^{(1)} = 0.16, \quad \beta_1 = \sup f/p^{(1)} = 0.64,\\
\alpha_2 = \inf f/p^{(2)} = 0.16, \quad \beta_2 = \sup f/p^{(2)} = 0.88
\end{aligned}
$$

and consequently from Theorem 6.2 we expect that the Euclidean condition numbers of $P_{(1)}^{-1/2} T_{\nu,\nu}(f) P_{(1)}^{-1/2}$ and $P_{(2)}^{-1/2} T_{\nu,\nu}(f) P_{(2)}^{-1/2}$ are bounded by $\beta_1/\alpha_1 = 4$ and $\beta_2/\alpha_2 = 5.6$, respectively.
The following Tables 17 and 18 show the perfect agreement with the theoretical results:

Table 17: Asymptotic eigenvalue behavior with preconditioner $P_{(1)}$

ν ($N(n) = \nu^2$)	$\lambda_{\min}(P_{(1)}^{-1} T_{\nu,\nu}(f))$	$\lambda_{\max}(P_{(1)}^{-1} T_{\nu,\nu}(f))$
5 (25)	0.170	0.525
10 (100)	0.163	0.598
15 (225)	0.161	0.618
20 (400)	0.160	0.627

In view of the results of [3] the theoretical convergence rates of the two preconditioned algorithms are independent of the dimension. Actually, for $n = 20$ (dimension$= 400$) we solve the system $T_{\nu,\nu}(f)\mathbf{x} = \mathbf{b}$, where $\mathbf{b}$ is a random vector, by using the preconditioners $P_{(1)}$ and $P_{(2)}$ in the conjugate gradient algorithm; the reduction of the 2–norm of the error shown by Table 19 confirms perfectly the theoretical convergence rates:

Therefore, with the two-level Toeplitz preconditioning, we obtain a uniformly bounded condition number with respect to ν. Moreover, the two preconditioners are not only two-level Toeplitz but they

Table 18: Asymptotic eigenvalue behavior with preconditioner $P_{(2)}$

$n\,(N(n)=\nu^2)$	$\lambda_{\min}(P_{(2)}^{-1}T_{\nu,\nu}(f))$	$\lambda_{\max}(P_{(2)}^{-1}T_{\nu,\nu}(f))$
5 (25)	0.170	0.547
10 (100)	0.163	0.691
15 (225)	0.161	0.755
20 (400)	0.160	0.786

Table 19: PCG, convergence history with preconditioners $P_{(1)}$ and $P_{(2)}$

step	$P_{(1)}$	$P_{(2)}$
2	$5.304435E-01$	$4.233243E-01$
4	$5.162726E-02$	$1.068652E-01$
6	$7.175051E-03$	$9.982540E-03$
8	$8.065245E-04$	$1.661406E-03$
10	$9.608787E-05$	$2.653328E-04$
12	$9.731413E-06$	$3.157243E-05$
14	$9.384791E-07$	$4.131877E-06$
16	$9.152440E-08$	$3.490431E-07$
18	$1.191590E-08$	$7.568337E-08$
20	$1.337295E-09$	$9.549909E-09$

also belong to the two-level τ algebra: as a consequence, the cost of solving a generic linear system with the preconditioners $P_{(1)}$ and $P_{(2)}$ is of order $O(\nu^2\log(\nu))$ and therefore the total cost of solving a system with coefficient matrix $T_{\nu,\nu}(f)$ is also of order $O(\nu^2\log(\nu))$. We mention that for this two-level doubly banded systems with nonnegative generating function an alternative strategy is the use of multigrid methods (see [39, 35]) that require $O(\nu^2)$ ops.

7 Conclusions

Essentially, there is one basic message: the matrix theoretic problem of preconditioning has been translated in the case of structured matrices of shift-invariant type into a function theory problem (here we have mainly considered Toeplitz matrices but the same holds for several classes of structured and "locally" structured matrices). Therefore some approximation theory results have been successfully adapted and used in our context.

Furthermore, the contents of the previous sections can be viewed as a link between two classical topics (structured linear algebra and approximation theory) in which, by interchanging and interlacing the point of view, we can simply find new results in both the directions.

References

[1] A. Arico', M. Donatelli, S. Serra Capizzano. "Multigrid optimal convergence for certain (multi-level) structured linear systems", *SIAM J. Matrix Anal. Appl.*, in press.

[2] O. Axelsson, V. Barker. *Finite Element Solution of Boundary Value Problems, Theory and Computation.* Academic Press Inc., New York, 1984

[3] O. Axelsson, G. Lindskog. “On the rate of convergence of the preconditioned conjugate gradient method”, *Numer. Math.*, 52 (1986), pp. 499-523.

[4] O. Axelsson, M. Neytcheva. “The algebraic multilevel iteration methods - theory and applications”, *Proc. of the 2nd Int. Coll. on Numerical Analysis*, D. Bainov Ed., Plovdiv (Bulgaria), august 1993, pp. 13–23.

[5] B. Beckermann, A. Kuijlaars. “Superlinear convergence of conjugate gradients”, *SIAM J. Numer. Anal.*, 39-1 (2001), pp. 700–329.

[6] B. Beckermann, S. Serra Capizzano. “On the asymptotic spectrum of Finite Elements matrices and CG convergence”, *manuscript*, (2002).

[7] R. Bhatia. *Matrix Analysis*. Springer-Verlag, New York, NY, 1997.

[8] D. Bini. “Matrix structure in parallel matrix computation”, *Calcolo*, 25 (1988), pp. 37–51.

[9] D. Bini. “Parallel solution of certain Toeplitz linear systems”, *SIAM J. Comput.*, 13 (1984), pp. 268–276.

[10] D. Bini, M. Capovani. “Spectral and computational properties of band symmetric matrices”, *Linear Algebra Appl.*, 52 (1983), pp. 99–125.

[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), july 1990, pp. 220–223.

[12] D. Bini, P. Favati. “On a matrix algebra related to the discrete Hartley transform”, *SIAM J. Matrix Anal. Appl.*, 14 (1993), pp. 500–507.

[13] D. Bini, V. Pan. *Numerical and Algebraic Computations with Matrices and Polynomials*. Birkäuser, Boston, 1994.

[14] A. Böttcher, S. Grudsky. “On the condition numbers of large semi-definite Toeplitz matrices”, *Linear Algebra Appl.*, 279 (1998), pp. 285–301.

[15] A. Böttcher, S. Grudsky. “Condition numbers of Toeplitz-like matrices”. *Talk at AMS-IMS-SIAM Summer Research Conference*, Boulder, Colorado, july 1999.

[16] A. Böttcher, B. Silbermann. *Introduction to Large Truncated Toeplitz Matrices*. Springer-Verlag, New York, NY, 1999.

[17] J. Bunch. “Stability of methods for solving Toeplitz systems of equations”, *SIAM J. Sci. Stat. Comp.*, 6 (1985), pp. 349–364.

[18] R.H. Chan. “Toeplitz preconditioners for Toeplitz systems with nonnegative generating functions”, *IMA J. Numer. Anal.*, 11 (1991), pp. 333–345.

[19] R.H Chan, T.F. Chan. “Circulant preconditioners for elliptic problems”, *Numer. Lin. Algebra Appl.*, 1 (1992), pp. 77–101.

[20] R.H. Chan, X.Q. Jin. “A family of block preconditioners for block systems”, *SIAM J. Sci. Stat. Comp.,* 13 (1992), pp. 1218–1235.

[21] R.H. Chan, M. Ng. “Conjugate gradient methods for Toeplitz systems”, *SIAM Rev.*, 38 (1996), pp. 427–482.

[22] R.H. Chan, G. Strang. “Toeplitz equations by conjugate gradient with circulant preconditioner”, *SIAM J. Sci. Stat. Comp.*, 10 (1989), pp. 104–119.

[23] R.H. Chan, P. Tang. “Fast band-Toeplitz preconditioners for Hermitian Toeplitz systems”, *SIAM J. Sci. Comp.*, 15 (1994), pp. 164–171.

[24] T.F. Chan. “An optimal circulant preconditioner for Toeplitz systems”, *SIAM J. Sci. Stat. Comp.*, 9 (1988), pp. 766–771.

[25] T.F. Chan, P.C. Hansen. “A look–ahead Levinson algorithm for indefinite Toeplitz systems”, *SIAM J. Matrix Anal. Appl.*, 13 (1992), pp. 491–506.

[26] P. Concus, G. Golub, G. Meurant. “ Block preconditioning for the conjugate gradient method”, *SIAM J. Sci. Stat. Comp.*, 6 (1985), pp. 220–252.

[27] P. Davis. *Circulant Matrices.* J. Wiley and sons, New York, 1979.

[28] C. De Lellis, P. Tilli. “On the spectral distribution of certain sequences of band matrices”, *Proc: Structured Matrices: Recent Developments in Theory and Computation*, NOVA Science Publishers, 2000.

[29] R. De Vore, G. Lorentz. *Constructive Approximation*. Springer-Verlag, Berlin, 1991.

[30] F. Di Benedetto, G. Fiorentino, S. Serra Capizzano. “C.G. Preconditioning for Toeplitz Matrices”, *Comput. Math. Appl.*, 25-6 (1993), pp. 33–45.

[31] F. Di Benedetto, S. Serra Capizzano. “A unifying approach to abstract matrix algebra preconditioning”, *Numer. Math.*, 82-1 (1999), pp. 57–90.

[32] F. Di Benedetto, S. Serra Capizzano. “Optimal multilevel matrix algebra operators”, *Linear Multilin. Algebra*, 48 (2000), pp. 35–66.

[33] G. Fiorentino, S. Serra Capizzano. “Multigrid methods for Toeplitz matrices”, *Calcolo*, 28-(3-4) (1991), pp. 283–305.

[34] G. Fiorentino, S. Serra Capizzano. “Fast parallel solvers for elliptic problems”, *Comput. Math. Appl.*, 32-2 (1996), pp. 61–68.

[35] G. Fiorentino, S. Serra Capizzano. “Multigrid methods for symmetric positive definite block Toeplitz matrices with nonnegative generating functions”, *SIAM J. Sci. Comp.*, 17-5 (1996), pp. 1068–1081.

[36] G. Golub, C. Van Loan. *Matrix Computations.* The Johns Hopkins University Press, Baltimore, 1983.

[37] U. Grenander, G. Szegö. *Toeplitz Forms and Their Applications.* Second Edition, Chelsea, New York, 1984.

[38] I. Gustafsson. “Stability and rate of convergence of modified incomplete Cholesky factorization methods”, *TR nr. 79.02R, Chalmers University of Technology, Sweden,* (1979).

[39] W. Hackbusch. *Multigrid Methods and Applications.* Springer Verlag, Berlin, 1985.

[40] W. Hackbusch. "Multigrid methods II", *Lecture notes in mathematics*, 1228 (1987).

[41] M. Hestenes, E. Stiefel. "Methods of conjugate gradients for solving linear systems", *J. Res. Nat. Bur. Stand.*, 49 (1952), pp. 409–436.

[42] T. Huckle, S. Serra Capizzano. "The spectra of preconditioned Toeplitz matrix sequences can have gaps", *SIAM J. Matrix Anal. Appl.*, in press.

[43] T. Huckle, S. Serra Capizzano, C. Tablino Possio. "Preconditioning strategies for non Hermitian Toeplitz linear systems", *Numer. Linear Algebra Appl.*, in press.

[44] T. Huckle, S. Serra Capizzano, C. Tablino Possio. "Preconditioning strategies for Hermitian indefinite Toeplitz linear systems", *SIAM J. Sci. Comp.*, in press.

[45] D. Jackson, *The Theory of Approximation*. American Mathematical Society, New York, 1930.

[46] M. Kac, W.L. Murdock, G.Szegö. "On the eigenvalues of certain Hermitian forms", *J. Rational Mech. Anal.*, 2 (1953), pp. 767–800.

[47] T. Kailath, V. Olshevsky. "Displacement structure approach to discrete-trigonometric-transform based preconditioners of G. Strang type and T. Chan type", *Proc. "Workshop on Toeplitz matrices"* Cortona (Italy), September 1996. *Calcolo*, 33 (1996), pp. 191–208.

[48] D. Kershaw. "The incomplete Cholesky-conjugate gradient method for the iterative solution of systems of linear equations", *J. Comput. Phys.*, 26 (1978), pp. 43–65.

[49] P.P. Korovkin. *Linear Operators and Approximation Theory* (English translation). Hindustan Publishing Co., Delhi, 1960.

[50] T.K. Ku, C.C.J. Kuo. "On the spectrum of a family of preconditioned Toeplitz matrices", *SIAM J. Sci. Stat. Comp.,* 13 (1992), pp. 948–966.

[51] A. Kuijlaars, S. Serra Capizzano. "Asymptotic zero distribution of orthogonal polynomials with discontinuously varying recurrence coefficients", *J. Approx. Th.*, 113 (2001), pp. 142–155.

[52] J. Lund, K. Bowers. *Sinc Methods for Quadrature and Differential Equations*. SIAM, Philadelphia, 1992.

[53] D. Noutsos, P. Vassalos, S. Serra Capizzano. "Matrix algebra preconditioners for multilevel Toeplitz systems do not insure optimal convergence rate", *Theoret. Comp. Sci.*, in press.

[54] D. Noutsos, P. Vassalos, S. Serra Capizzano. "A preconditioning proposal for ill-conditioned Hermitian two-level Toeplitz systems, *Numer. Lin. Algebra Appl.*, in press.

[55] B.N. Parlett. *The Symmetric Eigenvalue Problem*. Prentice Hall, Englewood Cliffs, 1980.

[56] S.V. Parter. "Extreme eigenvalues of Toeplitz forms and applications to elliptic difference equations", *Trans. Amer. Math. Soc."*, 99 (1966), pp. 153–192.

[57] M. Pourahmadi. "Remarks on extreme eigenvalues of Toeplitz matrices", *Internat. J. Math. & Math. Sci.*, 11 (1988), pp. 23–26.

[58] E.J. Remes, "Sur le calcul effectif des polynomes d'approximation de Tchebichef" *Compt. Rend. Acad. Sci. Paris*, 199 (1934), pp. 337–340.

[59] Y. Saad, M.H. Schultz. "GMRES: a generalized minimal residual algorithm for solving nonsymmetric linear systems", *SIAM J. Sci. Stat. Comp.*, 7 (1986), pp. 856–869.

[60] S. Serra Capizzano. "Multi-iterative methods", *Comput. Math. Appl.*, 26-4 (1993), pp. 65–87.

[61] S. Serra Capizzano. "Preconditioning strategies for asymptotically ill-conditioned block Toeplitz systems", *BIT*, 34-4 (1994), pp. 579–594.

[62] S. Serra Capizzano. "New PCG based algorithms for the solution of Hermitian Toeplitz systems", *Calcolo*, 32 (1995), pp. 153–176.

[63] S. Serra Capizzano. "On the extreme spectral properties of symmetric Toeplitz matrices generated by L^1 functions with several global minima/maxima", *BIT*, 36-1 (1996), pp. 135–142.

[64] S. Serra Capizzano. "Preconditioning strategies for Hermitian Toeplitz systems with nondefinite generating functions", *SIAM J. Matrix Anal. Appl.*, 17-4 (1996), pp. 1007–1019.

[65] S. Serra Capizzano. "Optimal, quasi–optimal and superlinear preconditioners for asymptotically ill–conditioned positive definite Toeplitz systems", *Math. Comput.*, 66-218 (1997), pp. 651–665.

[66] S. Serra Capizzano. "On the extreme eigenvalues of Hermitian (block) Toeplitz matrices", *Linear Algebra Appl.*, 270 (1998), pp. 109–129.

[67] S. Serra Capizzano. "An ergodic theorem for classes of preconditioned matrices", *Linear Algebra Appl.*, 282 (1998), pp. 161–183.

[68] S. Serra Capizzano. "Toeplitz preconditioners constructed from linear approximation processes", *SIAM J. Matrix Anal. Appl.*, 20-2 (1998), pp. 446–465.

[69] S. Serra Capizzano. "Superlinear PCG methods for symmetric Toeplitz systems", *Math. Comput.*, 68 (1999), pp. 793–803.

[70] S. Serra Capizzano. "The rate of convergence of Toeplitz based PCG methods for second order nonlinear boundary value problems", *Numer. Math.*, 81-3 (1999), pp. 461–495.

[71] S. Serra Capizzano. "A Korovkin-type Theory for finite Toeplitz operators via matrix algebras", *Numer. Math.*, 82-1 (1999), pp. 117–142.

[72] S. Serra Capizzano. "How to choose the best iterative strategy for symmetric Toeplitz systems", *SIAM J. Numer. Anal.*, 36-4 (1999), pp. 1078–1103.

[73] S. Serra Capizzano. "A note on the asymptotic spectra of finite difference discretizations of second order elliptic Partial Differential Equations", *Asian J. Math.*, 4 (2000), pp. 499–514.

[74] S. Serra Capizzano. "Distribution results on the algebra generated by Toeplitz sequences: a finite dimensional approach", *Linear Algebra Appl.*, 328-(1-3) (2001), pp. 121–130.

[75] S. Serra Capizzano. "Convergence analysis of Two-Grid methods for elliptic Toeplitz and PDEs matrix-sequences", *Numer. Math.*, 92-3 (2002), pp. 433–465.

[76] S. Serra Capizzano. “Test functions, growth conditions and Toeplitz matrices”, *Rend. Circ. Mat. Palermo* Serie II, N. 68 (2002), pp. 791–795.

[77] S. Serra Capizzano. “Matrix algebra preconditioners for multilevel Toeplitz matrices are not superlinear”, *Linear Algebra Appl.*, 343/344 (2002), pp. 303–319.

[78] S. Serra Capizzano. “Generalized Locally Toeplitz sequences: spectral analysis and applications to discretized Partial Differential equations”, *Linear Algebra Appl.*, 366-1 (2003), pp. 371–402.

[79] S. Serra Capizzano. “Practical band Toeplitz preconditioning and boundary layer effects”, *Numer. Alg.*, 34 (2003), pp. 427–440.

[80] S. Serra Capizzano, C. Tablino Possio. “Spectral and structural analysis of high precision Finite Difference matrices for Elliptic Operators”, *Linear Algebra Appl.*, 293 (1999), pp. 85–131.

[81] S. Serra Capizzano, C. Tablino Possio. “Analysis of preconditioning strategies for collocation linear systems”, *Linear Algebra Appl.*, 369 (2003), pp. 41–75.

[82] S. Serra Capizzano, P. Tilli. “Extreme singular values and eigenvalues of non Hermitian Toeplitz matrices”, *J. Comput. Appl. Math.*, 108-(1-2) (1999), pp. 113–130.

[83] S. Serra Capizzano, P. Tilli. “On unitarily invariant norms of matrix valued linear positive operators”, *J. Ineq. Appl.*, 7-3 (2002), pp. 309–330.

[84] S. Serra Capizzano, E.E. Tyrtyshnikov. “Any circulant-like preconditioner for multilevel matrices is not superlinear”, *SIAM J. Matrix Anal. Appl.*, 21-2 (1999), pp. 431–439.

[85] S. Serra Capizzano, E.E. Tyrtyshnikov. “How to prove that a preconditioner can not be superlinear”, *Math. Comput.* 72 (2003), pp. 1305–1316.

[86] J. Stoer, R. Bulirsh. *Introduction to Numerical Analysis.* 3rd Ed., Text in Appl. Math. 12, Springer, New York, 2002.

[87] G. Strang. “A proposal for Toeplitz matrix calculation”. *Stud. Appl. Math.*, 74 (1986), pp. 171–176.

[88] P. Tang. “A fast algorithm for linear complex Chebyshev approximation”, *Math. Comput.*, 51 (1988), pp. 721–739.

[89] P. Tilli. “Singular values and eigenvalues of non-Hermitian block Toeplitz matrices, *Linear Algebra Appl.*, 272 (1998), pp. 59–89.

[90] P. Tilli. “Locally Toeplitz sequences: spectral properties and applications”, *Linear Algebra Appl.*, 278 (1998), pp. 91–120.

[91] P. Tilli. “A note on the spectral distribution of Toeplitz matrices, *Linear Multilin. Algebra*, 45 (1998), pp. 147–159.

[92] P. Tilli. “Some results on complex Toeplitz eigenvalues”, *J. Math. Anal. Appl.*, 239 (1999), pp. 390–401.

[93] E.E. Tyrtyshnikov. “Influence of matrix operations on the distribution of eigenvalues and singular values of Toeplitz matrices”. *Linear Algebra Appl.*, 207 (1994), pp. 225–249.

[94] E.E. Tyrtyshnikov, "Circulant preconditioners with unbounded inverses", *Linear Algebra Appl.*, 216 (1995), pp. 1–23.

[95] E.E. Tyrtyshnikov. "A unifying approach to some old and new theorems on distribution and clustering", *Linear Algebra Appl.*, 232 (1996), pp. 1–43.

[96] E.E. Tyrtyshnikov, N. Zamarashkin. "Spectra of multilevel Toeplitz matrices: advanced theory via simple matrix relationships", *Linear Algebra Appl.*, 270 (1998), pp. 15–27.

[97] C. Van Loan. *Computational frameworks for the Fast Fourier Transform.* SIAM publication, Philadelphia, 1992.

[98] R. S. Varga. *Matrix Iterative Analysis.* Prentice Hall, Englewood Cliffs, 1962.

[99] H. Widom. "On the eigenvalues of certain Hermitian operators", *Trans. Amer. Math. Soc.*, 88 (1958), pp. 491–522.

[100] H. Widom. "On the singular values of Toeplitz matrices", *Zeit. Anal. Anw.*, 8 (1989), pp. 221–229.

[101] H. Widom. *Toeplitz Matrices.* In Studies in real and Complex analysis, I. Hirshman Jr Ed., Math. Ass. Am., 1965.

[102] H. Widom. "Asymptotic behavior of block Toeplitz matrices and determinants II", *Adv. Math.*, 21 (1976), pp. 1-29.

8 Appendix A: convergence results for (P)CG methods

We start by recalling that for a sequence of positive definite matrices $\{A_n\}$, typically arising as finite dimensional approximations of a given operator in infinite dimensions, the notion of sequence of difficult problems can be written as follows: we assume that the maximal eigenvalue tends to a positive constant and the minimal eigenvalue $\lambda_1^{(n)}$ of A_n tends to zero as n tends to infinity.

As observed in Section 2, in the Toeplitz case this situation occurs if the symbol is real valued, nonnegative with one or more essential zeros and with strictly positive essential supremum. A further class of difficult problems is represented by the Finite Differences (or Finite Elements) approximations of elliptic differential operators with homogeneous boundary conditions. For the sake of simplicity let us consider the class of variable coefficient operators

$$\frac{d^k}{dx^k}\left(a(x)\frac{d^k}{dx^k}u\right) \tag{54}$$

with strictly positive a. The conditioning of $A_n = A_n(a)$ obtained by equispaced centered Finite Difference formulae grows asymptotically as n^{2k} (see [70]). The conditioning can be even worse if a vanishes in some isolated points (semielliptic operators). A proof of these statements is very easy because the operators $A_n(\cdot)$ are linear and positive and $A_n(1)$ coincides with $T_n(f_k)$ where f_k is a polynomial having a unique zero at $x = 0$ of order $2k$ (the maximal order of the derivatives): therefore by linearity and positivity we have

$$(\inf a)T_n(f_k) = (\inf a)A_n(1) \leq A_n(a) \leq (\sup a)A_n(1) = (\sup a)T_n(f_k)$$

and the conditioning of $T_n(f_k)$ is known to grow as n^{2k} thanks to Corollary 2.1. We recall that for $k \geq 1$ and FD formulae of precision order 2 and minimal bandwidth, the polynomial $f_k(s)$ coincides with $(2 - 2\cos(s))^k$: for $k = 1$ the matrix $A_n(1) = T_n(f_1)$ is the one displayed in equation (5) while for $k = 2$ the matrix $A_n(1) = T_n(f_2)$ is the one displayed in equation (77).

To have an idea of the difficulties due to the ill-conditioning, we consider a class of matrices A_n with condition numbers asymptotic to $g(n)$ and $g(n)$ diverging as n: we observe that classical not specialized iterative methods (Jacobi, Gauss-Seidel, Conjugate Gradient [41]) require a number of steps (to reach a given accuracy) which is a diverging function as n somehow related to $g(n)$. As a simple but significant example, we consider equispaced centered Finite Difference formulae of precision 2 with minimal bandwidth applied to the operators in (54) with $a(x) = 1$ and $k \in \{1, 2\}$: the number of steps to reach the desired precision ϵ is displayed for some classical iterative solvers:

	$\frac{d^2}{dx^2}$	$\frac{d^4}{dx^4}$
Jacobi	$O(n^2)$	$O(n^4)$
Optimal damped Jacobi	$O(n^2)$	$O(n^4)$
Gauss-Seidel	$O(n^2)$	$O(n^4)$
Gauss-Seidel with optimal parameter	$O(n)$	$O(n^2)$
Conjugate Gradient	$O(n)$	$O(n)$

It is evident that these results are computationally not satisfactory: therefore for large dimensions we need "adaptive" methods that can exploit the structure of the special problem at hand. This is the case of the PCG methods and of the multigrid methods: in the first case the preconditioner P_n represents the crucial parameter for accelerating the convergence; in the second case the pair (smoother, projector) plays the same role.

8.1 A.1. Optimality of iterative solvers

We are interested in optimal methods. A definition has been proposed in [4]: a PCG method with preconditioner P_n is called optimal (in the sense of the convergence rate) if

o1 the spectrum of $P_n^{-1}A_n$ lies between two constants independent of n or equivalently $\{P^{-1/2}A_nP^{-1/2}\}$ is asymptotically well conditioned.

Analogously we say that a PCG method with preconditioner P_n is optimal (in the sense of the cost per iteration) if

o2 the cost of the solution of a generic linear system with coefficient matrix P_n is of the same order as the number of parameters characterizing the class $\{A_n\}$.

These definitions make sense only in a sequential model of computation and can be generalized to a generic iterative method. Moreover the second definition is not satisfactory because in some cases the number of parameters does not decide the complexity of the matrix vector product with matrix A_n and therefore requirement **o2** becomes too restrictive.

We modify and generalize the previous definitions as follows: we write that a generic iterative solver is optimal (in the sense of the convergence rate) if

opt1 the number of steps $N(\epsilon)$ for reaching the solution of a system within a preassigned accuracy ϵ and with coefficient matrix A_n can be bounded from above by a constant not depending on n.

The given iterative solver is optimal (in the sense of the cost per iteration) if

opt2 the cost per iteration is at most proportional to the cost of the product of A_n by a generic vector.

We observe that **opt1** and **opt2** coincides with Definition 4.1 and that **o1** is a special case of **opt1** in the case of the PCG method. On the other hand, as previously observed, requirement **o2** is too restrictive: it is a reasonable request in the case of the banded systems appearing in the discretization of partial differential equations but is not realistic for Toeplitz structures and especially for banded Toeplitz structures which are characterized by a constant number of parameters. Indeed, since a generic iterative solver uses at each step matrix vector products, it follows that a cost per iteration cannot be less than this and therefore requirement **opt2** is the right one. Finally we remark that in the case of generic banded system the new definition **opt2** reduces to the one of Axelsson and Neytcheva.

8.2 A.2. (Preconditioned) Conjugate Gradient method

Given a positive definite matrix A_n we consider the conjugate gradient (CG) method: from a theoretical viewpoint this is a direct method that constructs a finite sequence of m vectors with $m \leq n$ such that the last one coincides with the solution of $A_n\mathbf{x} = \mathbf{b}$. In practice due to unavoidable rounding errors, it behaves and it is used as an iterative solver. This apparently conflicting duality is unified in refined results due to Axelsson and Lindskög.

The idea behind the gradient algorithm is to transform the unique solution of the linear system $A_n\mathbf{x} = \mathbf{b}$ into the unique minimum point of the quadratic functional

$$\Phi_n(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T A_n\mathbf{x} - \mathbf{b}^T\mathbf{x}. \tag{55}$$

One step of a gradient method works as follows: at the first step we choose a initial guess (otherwise we start from the present iterate), we compute a direction vector along which the functional locally decreases and we compute a real positive constant which represents the length of the optimal correction. The new vector is computed by summing the previous one with the direction vector times the computed constant. If the direction vector is the residual $\mathbf{r}(\mathbf{x}) = \mathbf{b} - A_n\mathbf{x}$ i.e. the direction of the steepest descent then the method is exactly the *steepest descent* method. In spite of the local optimality of this method, if the matrix A_n is very ill-conditioned the obtained sequence is very oscillating and the convergence is slow: here for slow we mean that the reduction of the error at every step in a suitable norm is equal to $1 - \epsilon_n$ where ϵ_n is asymptotic to the inverse of the spectral conditioning of A_n: for proving this, use the Kantorovich inequality i.e.

$$\frac{\mathbf{x}^H\mathbf{x}}{(\mathbf{x}^H A_n\mathbf{x})(\mathbf{x}^H A_n^{-1}\mathbf{x})} \geq \frac{4\lambda_{\max}(A_n)\lambda_{\min}(A_n)}{(\lambda_{\max}(A_n) + \lambda_{\min}(A_n))^2}, \quad \forall\, \mathbf{x} \in \mathbf{C}^n.$$

In the case of a Toeplitz matrix with nonnegative symbol f having a zero of order 4 (e.g. the Finite Differences discretization of the fourth derivative), we have to expect $O(n^4)$ iterations in order to decrease the error by a given factor.

To improve things (in particular to avoid oscillations) the direction vector is chosen to be A_n-orthogonal to the previous one ($\mathbf{x}$ and $\mathbf{y}$ are A_n-orthogonal if $\mathbf{x}^H A_n\mathbf{y} = 0$): the resulting method is the conjugate gradient (CG) method. For completeness the structure of the CG algorithm is reported below.

$CG(\mathbf{x}_0)$:
Step 1. $k := 0$, $\mathbf{r}_0 := \mathbf{b} - A_n\mathbf{u}_0$, Step 2. if $\|\mathbf{r}_k\|_2 < \epsilon\|b\|_2$ stop, Step 3. $\beta_k := \frac{\mathbf{r}_k^T\mathbf{r}_k}{\mathbf{r}_{k-1}^T\mathbf{r}_{k-1}}$ $(\beta_0 = 0$ if $k = 0)$, $\mathbf{p}_k := \mathbf{r}_k + \beta_k\mathbf{p}_{k-1}$, $\alpha_k := \frac{\mathbf{r}_k^T\mathbf{r}_k}{\mathbf{p}_k^T A_n\mathbf{p}_k}$, $\mathbf{x}_{k+1} := \mathbf{x}_k + \alpha_k\mathbf{p}_k$, $\mathbf{r}_{k+1} := \mathbf{r}_k - \alpha_k A_n\mathbf{p}_k$, $k = k + 1$, go to step 2.

The following two theorems (see e.g. [36]) give convergence results looking at the method first as a direct one and then as an iterative one.

Theorem 8.1 *If A_n has s distinct eigenvalues then the number of steps necessary to compute the solution in exact arithmetic is finite and it is bounded by s: more in detail, there exists $m \leq s$ such that $\mathbf{r}_m = 0$.*

Dim. It directly follows from the orthogonality of the residuals $\mathbf{r}_k$ whose proof can be given by induction. •

Theorem 8.2 *The k-th error in A_n norm is reduced with respect to $(k-1)$-th error in A_n norm by a factor bounded by $1 - \epsilon_n$ where $\epsilon_n \sim (\mu(A_n))^{-1/2}$ with $\mu(A_n)$ denoting the spectral condition number of A_n.*

The improvement with respect to the steepest descent method is substantial: for the steepest descent the quantity $(\mu(A_n))^{-1/2}$ in Theorem 8.2 has to be replaced by $(\mu(A_n))^{-1}$. However, the dependency on the condition number is still heavy: this drawback can be substantially overcome since the CG method can be enriched by using preconditioners. The original problem is transformed in a new one but better conditioned. We look for a nonsingular matrix C_n and we apply the CG algorithm to a functional as in (55) where the matrix A_n is replaced by $C_n^{-T}A_nC_n^{-1}$. Since the conditioning of the new symmetric positive definite matrix is decided by its eigenvalues and the latter is similar to $P_n^{-1}A_n$ with $P_n = C_n^TC_n$, it is of interest to study the spectral properties of the sequence $\{P_n^{-1}A_n\}$ as done in Section 3 in the Toeplitz context. It is not necessary to form the matrix C_n explicitly as stressed in the subsequent algorithm unless it is useful for solving a generic system with matrix P_n. This situation occurs when P_n is an incomplete Cholesky factorization of A_n: in that case the triangular factors C_n and C_n^T are of computational interest.

$PCG(\mathbf{x}_0, P_n)$:
Step 1. $k := 0$, $\mathbf{r}_0 := \mathbf{b} - A_n\mathbf{u}_0$, Step 2. if $\|\mathbf{r}_k\|_2 < \epsilon\|b\|_2$ stop, Step 3. compute $\mathbf{z}_k$ such that $P_n\mathbf{z}_k = \mathbf{r}_k$, $\beta_k := \frac{\mathbf{z}_k^T\mathbf{r}_k}{\mathbf{z}_{k-1}^T\mathbf{r}_{k-1}}$ $(\beta_0 := 0$ if $k = 0)$, $\mathbf{p}_k = \mathbf{z}_k + \beta_k\mathbf{p}_{k-1}$, $\alpha_k := \frac{\mathbf{z}_k^T\mathbf{r}_k}{\mathbf{p}_k^T A_n\mathbf{p}_k}$, $\mathbf{x}_{k+1} := \mathbf{x}_k + \alpha_k\mathbf{p}_k$, $\mathbf{r}_{k+1} := \mathbf{r}_k - \alpha_k A_n\mathbf{p}_k$, $k = k + 1$, go to step 2.

Since the PCG is exactly the same CG applied to an equivalent functional with A_n replaced by $C_n^{-T} A_n C_n^{-1}$ it follows that the previous Theorems 8.1 and 8.2 hold unchanged except for the replacement of A_n by $C_n^{-T} A_n C_n^{-1}$. Finally we observe that in both the results the important information concerns the spectrum and therefore the convergence analysis of the PCG method can be carried out by studying the spectral properties of $P_n^{-1} A_n$ which is similar to $C_n^{-T} A_n C_n^{-1}$.

Therefore a good preconditioner should be spectrally "close" to the matrix A_n and "easily" invertible. Here, according to definition **opt2**, easily invertible means that the solution of a generic linear system with matrix P_n must have a cost at most proportional to the matrix vector product with matrix A_n.

Looking carefully to the two requirements it is not difficult to see that they are somehow conflicting and therefore the search for a good preconditioning strategy is a refined balancing analysis. For many classes of matrices (including those arising in a Partial Differential Equation context), a popular technique is based on the incomplete Cholesky factorization [2, 26, 38, 48]. These preconditioning techniques fulfill requirement **opt2** but in general do not fulfill requirement **opt1** (and therefore **o1**). For structured matrices of Toeplitz type, matrix algebra preconditioners satisfy requirement **opt2** and requirement **opt1** only in the unilevel context while the band Toeplitz matrices have a uniform behavior (see Section 6): the problem in the multilevel setting is the requirement **opt2** for which we have to use multigrid techniques.

We now report an important result [3] that combines both the direct and the iterative nature of the PCG algorithm.

Theorem 8.3 *Let A_n and P_n be positive definite matrices and let*

$$k^*(a,b,\epsilon) = \left\lceil \log\left[2\epsilon^{-1}\right] / \log\left[\sigma^{-1}\right] \right\rceil$$

be a function where a, b, ϵ are positive and with $\sigma = \dfrac{\sqrt{b}-\sqrt{a}}{\sqrt{b}+\sqrt{a}}$. Let us suppose that the spectrum Σ_n of $P_n^{-1} A_n$ behaves according to one of the following possibilities:

1. *$\Sigma_n \subset [a,b]$ except q eigenvalues all bigger than b;*
2. *$\Sigma_n \subset [a,b]$ except q eigenvalues all smaller than a;*
3. *$\Sigma_n \subset [a,b]$ except $2q$ eigenvalues: q of them bigger than b and the remaining smaller than a.*

Then, in order to reach the solution of a system with matrix A_n and within a preassigned accuracy ϵ, the PCG method with preconditioner P_n requires $N(\epsilon)$ iterations where we have:

- $N(\epsilon) = q + k^*(a,b,\epsilon)$;
- $N(\epsilon) = q + k^*(a,b,\epsilon^*)$ *with* $\epsilon^* = \epsilon \cdot \prod_{j=1}^{q} \lambda_j^- / b$;
- $N(\epsilon) = 2q + k^*(a,b,\epsilon^{**})$ *with* $\epsilon^{**} = \epsilon \cdot \prod_{j=1}^{q} 4\lambda_j^- / \lambda_j^+ \left(1 - \lambda_j^- / \lambda_j^+\right)^{-2}$.

If $\mu(A_n)$ denotes the spectral condition number and if q is constant with respect to n, then in the second and third case the quantities $N(\epsilon)$ are asymptotic to $\log(\mu(A_n))$ with constant depending on q and ϵ and, in the first case, $N(\epsilon)$ is uniformly bounded from above by a constant depending on q and ϵ.

We observe that the number of iterations is bounded by a constant independent of n in the first case (**opt1** and **o1** satisfied), while in the cases 2 and 3 there exists a possible dependency on n as the logarithm of the conditioning. Furthermore if $a = b$ (strong clustering at a) then not only **opt2** and **o2** are satisfied and the quality of the convergence is superlinear (see e.g. [11]).

From the latter statement we deduce that the strong clustering is a very interesting property: for the sake of completeness we recall the definitions of proper and weak clustering (for singular values and eigenvalues).

Definition 8.1 *A sequence of matrices $\{A_n\}$ with A_n of size d_n, $d_n < d_{n+1}$ for all n, is* properly clustered at $a \geq 0$ in the singular value sense *if for every $\epsilon > 0$ there exists c_ϵ (independent of n) such that $\#\{j : \sigma_j^{(n)} \notin (a-\epsilon, a+\epsilon)\} \leq c_\epsilon$ uniformly with respect to n and with $\sigma_j^{(n)}$ singular values of A_n.*
$\{A_n\}$ is weakly clustered at $a \geq 0$ in the singular value sense *if for every $\epsilon > 0$ it follows $\#\{j : \sigma_j^{(n)} \notin (a-\epsilon, a+\epsilon)\} = o(d_n)$.*
Analogously a sequence of matrices $\{A_n\}$ is properly clustered at a in the eigenvalue sense *if for every $\epsilon > 0$ there exists c_ϵ (independent of n) such that $\#\{j : |\lambda_j^{(n)} - a| > \epsilon\} \leq c_\epsilon$ uniformly with respect to n and with $\lambda_i^{(n)}$ eigenvalues of A_n.*
$\{A_n\}$ is weakly clustered at $a \geq 0$ in the eigenvalue sense *if for every $\epsilon > 0$ we have $\#\{j : |\lambda_j^{(n)} - a| > \epsilon\} = o(d_n)$.*

In the following we prove a sufficient condition for clustering.

Theorem 8.4 *Let $\{A_n\}$ and $\{P_n\}$ be two sequences of Hermitian matrices satisfying the condition $\|A_n - P_n\|_p \leq c_p$ for a certain $p \in [1, \infty)$ with c_p independent of n. Then the following is true:*

- *The sequence $\{A_n - P_n\}$ is properly clustered at zero in the sense of the eigenvalues;*
- *if P_n are positive definite for every n with uniformly bounded inverse (in spectral norm), then the sequence $\{P_n^{-1} A_n\}$ is properly clustered at one in the sense of the eigenvalues.*

Moreover, if $\|A_n - P_n\|_p^p = o(d_n)$ with d_n size of A_n, then the following is true:

- *The sequence $\{A_n - P_n\}$ is weakly clustered at zero in the sense of the eigenvalues;*
- *if P_n are positive definite for every n and are sparsely vanishing (i.e. their inverses are sparsely unbounded in the sense of (71)), then the sequence $\{P_n^{-1} A_n\}$ is weakly clustered at one in the sense of the eigenvalues.*

Proof. We prove the first two items concerning the proper clustering. Fix $\epsilon > 0$ and consider

$$k_{\epsilon,n} = \#\{j : \left|\lambda_j^{(n)}\right| > \epsilon\}$$

with $\lambda_j^{(n)}$ eigenvalues of $A_n - P_n$. From the assumption we have $c_p^p \geq \|A_n - P_n\|_p^p = \sum_i \left|\lambda_i^{(n)}\right|^p$ and therefore

$$\begin{aligned} c_p^p &\geq \|A_n - P_n\|_p^p \\ &= \sum_j \left|\lambda_j^{(n)}\right|^p \\ &> \sum_{\{j:\, \left|\lambda_j^{(n)}\right| > \epsilon\}} \epsilon^p \\ &= k_{\epsilon,n} \epsilon^p. \end{aligned}$$

In conclusion $k_{\epsilon,n} < c_p^p \epsilon^{-p}$ which is independent of n. This proves the first item; for the second item it is enough to observe that

$$\begin{aligned} P_n^{-1} A_n &= P_n^{-1}(P_n + (A_n - P_n)) \\ &= I_n + P_n^{-1}(A_n - P_n) \end{aligned}$$

and since $\{P_n^{-1}\}$ is uniformly bounded and $\{A_n - P_n\}$ is properly clustered at zero by the first item, it follows that $\{P_n^{-1}(A_n - P_n)\}$ is also properly clustered at zero and finally $\{P_n^{-1} A_n = I_n + P_n^{-1}(A_n - P_n)\}$ is properly clustered at 1.
The rest of the proof concerning the weak clustering is totally similar. •

By putting the singular values in place of the eigenvalues, along the very same lines it is possible to prove a generalization to generic matrices of the former result.

Theorem 8.5 *Let $\{A_n\}$ and $\{P_n\}$ be two sequences of matrices satisfying the condition $\|A_n - P_n\|_p \leq c_p$ for a certain $p \in [1, \infty)$ with c_p independent of n. Then the following is true:*

- *The sequence $\{A_n - P_n\}$ is properly clustered at zero in the sense of the singular values;*
- *if P_n are invertible for every n with uniformly bounded inverse (in spectral norm), then the sequence $\{P_n^{-1} A_n\}$ is properly clustered at one in the sense of the singular values.*

Moreover, if $\|A_n - P_n\|_p^p = o(d_n)$ with d_n size of A_n, then the following is true:

- *The sequence $\{A_n - P_n\}$ is weakly clustered at zero in the sense of the singular values;*
- *if P_n are invertible for every n and are sparsely vanishing (i.e. their inverses are sparsely unbounded in the sense of (71)), then the sequence $\{P_n^{-1} A_n\}$ is weakly clustered at one in the sense of the singular values.*

9 Appendix B: global distribution results for matrix sequences

In this appendix we present algebraic tools for establishing distribution results concerning matrix sequences: as an example of applications we identify the distribution of sequences obtained a linear combination of product of Toeplitz sequences [74]. A special case of the mentioned result coincides with the the Szegö Theorem concerning Toeplitz sequences in the full generality provided by Tyrtyshnikov and Zamarashkin [96].

First, let us introduce some notations and definitions. For any real valued function F defined on $\mathbf{R}$ and for any matrix $A \in M_N(\mathbf{C})$, by the symbol $\Sigma(F, A)$ we denote the mean

$$\frac{1}{N} \sum_{j=1}^{N} F[\sigma_j(A)]$$

and by the symbol $\|\cdot\|$ the spectral norm (Schatten p norms with $p = \infty$ [7]) where

$$\|A\| = \sigma_N(A), \quad \|A\|_p = \left[\sum_{j=1}^{N} \sigma_j^p \right]^{1/p} = [N \cdot \Sigma(|\cdot|^p, A)]^{1/p}$$

and $\sigma_1(A) \leq \sigma_2(A) \leq \cdots \leq \sigma_N(A)$ singular values of A: of course if the matrix is normal (a class containing all the Hermitian matrices) then the singular values $\sigma_j(A)$, $j = 1, \ldots, N$, coincide with

the modulus of the eigenvalues $\lambda_j(A)$, $j = 1, \ldots, N$. Moreover, in the case of Hermitian matrices, we use the symbol $\Sigma_\lambda(F, A)$ for denoting the eigenvalue mean

$$\frac{1}{N}\sum_{j=1}^{N} F[\lambda_j(A)].$$

Definition 9.1 *Given a sequence $\{A_n\}$ of matrices of size d_n with $d_n < d_{n+1}$ and given a function f defined over a set K equipped with a σ finite measure μ we say that $\{A_n\}$ is distributed as (f, K, μ) in the sense of the singular values (in the sense of the eigenvalues) if for any continuous F with bounded support the following limit relation holds*

$$\lim_{n\to\infty} \Sigma(F, A_n) = \frac{1}{\mu(K)}\int_K F(|f|)d\mu, \quad \left(\lim_{n\to\infty} \Sigma_\lambda(F, A_n) = \frac{1}{\mu(K)}\int_K F(f)d\mu\right).$$

In this case we write in short $\{A_n\} \sim_\sigma (f, K, \mu)$ ($\{A_n\} \sim_\lambda (f, K, \mu)$).

In the following the symbol μ is suppressed for many cases of interest (Toeplitz sequences, Generalized Locally Toeplitz sequences [78] etc.) since the measure always coincides with the standard Lebesgue measure $\mu\{\cdot\}$ on $\mathbf{R}^d$ for some positive integer d.

9.1 B.1. General tools for matrix sequences

We start with a perturbation result which is of paramount interest for estimating the spectral distribution from a quantitative point of view. This result is a generalization of the Wielandt-Hoffman inequality (see e.g. [7]) and represents a particular case of the Lidskii-Mirsky-Wielandt Theorem whose proof can be found in [7, Bhatia: Th. IV.3.4 and Ex. IV.3.5]:

Lemma 9.1 *Let $p \in [1, \infty)$. For any pair of matrices $A, B \in M_N(\mathbf{C})$, we have*

$$\left[\sum_{j=1}^{N} |\sigma_j(A) - \sigma_j(B)|^p\right]^{1/p} \leq \|A - B\|_p. \tag{56}$$

For $p = \infty$ we have $\max_j |\sigma_j(A) - \sigma_j(B)| \leq \|A - B\|$. If A and B are Hermitian, then we also have

$$\left[\sum_{j=1}^{N} |\lambda_j(A) - \lambda_j(B)|^p\right]^{1/p} \leq \|A - B\|_p \tag{57}$$

and, for $p = \infty$, $\max_j |\lambda_j(A) - \lambda_j(B)| \leq \|A - B\|$.

We now introduce some tools for dealing with matrix sequences: in the following we consider matrices A_n of size d_n which belong to a sequence of matrices $\{A_n\}$ with $d_n < d_{n+1}$. The idea is to give an elementary approximation theory for matrix sequences: we express complicate sequences in terms of simple splittings and/or simple elementary operations (linear combinations, products etc.) of simpler matrix sequences. We recover the spectral distribution of a complicate matrix sequence from the distributions of simpler ones.

Definition 9.2 *Suppose a sequence of matrices $\{A_n\}$ of size d_n is given. We say that $\{\{B_{n,m}\}\}_m$, $m \in \mathbf{N}$ is an approximating class of sequences (a.c.s.) for $\{A_n\}$ if, for all sufficiently large $m \in \mathbf{N}$, the following splittings hold:*

$$A_n = B_{n,m} + R_{n,m} + N_{n,m}, \quad \forall n > n_m, \tag{58}$$

with

$$\mathrm{rank}(R_{n,m}) \leq d_n\, c(m), \quad \|N_{n,m}\| \leq \omega(m), \tag{59}$$

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

$$\lim_{m\to\infty} \omega(m) = 0, \ \lim_{m\to\infty} c(m) = 0. \tag{60}$$

Alternative (but equivalent) characterizations of the notion of a.c.s. can be provided in terms of other Schatten norms (see e.g. [78]). In the subsequent Lemma 9.2 and Lemma 9.3 we relate the notion in Definition 9.2 with the quantity $\Sigma(\cdot,\cdot)$.

Lemma 9.2 [74] *Suppose a sequence of matrices $\{A_n\}$ of size d_n is given and suppose that $\{\{B_{n,m}\}\}_m$, $m \in \mathbf{N}$ is an a.c.s. for $\{A_n\}$ with $R_{n,m} \equiv 0$. Then for any $F \in \mathcal{C}_0$ there exists $\theta(\cdot)$ with $\theta(m) \to 0$ as $m \to \infty$ such that*

$$|\Sigma(F, A_n) - \Sigma(F, B_{n,m})| \leq \theta(m).$$

Proof. From the assumption and from relation (60) in Definition 9.2, we know that

$$\|A_n - B_{n,m}\| \leq \omega(m), \ \ \theta(m) \to 0 \ \text{ as } m \to \infty.$$

We set $\omega_F : \mathbf{R}_0^+ \to \mathbf{R}_0^+$ the modulus of continuity (see e.g. [29]) of F defined as $\omega_F(\delta) = \sup_{x,y:\ |x-y|\leq\delta} |F(x) - F(y)|$: it is quite obvious to prove that

1. ω_F is a a nondecreasing function;
2. $\lim_{\delta\to 0^+} \omega_F(\delta) = \omega_F(0) = 0$ if and only if F is uniformly continuous in its domain.

Therefore

$$\begin{aligned}
|\Sigma(F, A_n) - \Sigma(F, B_{n,m})| &\leq \frac{1}{d_n}\sum_{j=1}^{d_n} |F[\sigma_j(A_n)] - F[\sigma_j(B_{n,m})]| \\
&\leq \frac{1}{d_n}\sum_{j=1}^{d_n} \omega_F(\max_j |\sigma_j(A_n) - \sigma_j(B_{n,m})|) \\
&= \omega_F(\max_j |\sigma_j(A_n) - \sigma_j(B_{n,m})|) \\
&\leq \omega_F(\|A_n - B_{n,m})\|) \leq \omega_F(\omega(m)).
\end{aligned}$$

where in the latter 3 steps we have used the definition of modulus of continuity, Lemma 9.1 with $p = \infty$, and the monotonicity of ω_F. It is clear that the desired $\theta(m)$ is $\omega_F(\omega(m))$ which tends to zero as m tends to infinity since F is continuous with bounded support and therefore uniformly continuous over $\mathbf{R}$. •

Lemma 9.3 [74] *Suppose a sequence of matrices $\{A_n\}$ of size d_n is given and suppose that $\{\{B_{n,m}\}\}_m$, $m \in \mathbf{N}$ is an a.c.s. for $\{A_n\}$ with $N_{n,m} \equiv 0$. Then, denoting by BV the set of functions with bounded variation on $\mathbf{R}$, for any $F \in \mathcal{C}_0 \cup BV$ there exists $\theta(\cdot)$ with $\theta(m) \to 0$ as $m \to \infty$ such that*

$$|\Sigma(F, A_n) - \Sigma(F, B_{n,m})| \leq \theta(m).$$

Proof. The statement for $F \in \mathcal{C}_0$ can be found in [28]. Here we give an alternative proof which is based on the BV case which is in turn based on the case where the test function is monotone and bounded.

For any $F \in \mathcal{C}_0$ and any $\epsilon > 0$ consider $F_\epsilon \in \mathcal{C}_0^1$ such that $\|F - F_\epsilon\|_\infty \leq \epsilon$. Since F_ϵ is Lipschitz continuous and its support K is compact we can write that F_ϵ is in BV (i.e. is of bounded variation) and therefore the thesis is reduced to the BV case.

Hence let $F \in BV$ and let us observe that $F = F^+ - F^-$ where F^+ and F^- are nondecreasing with

$$\mathrm{Var}(F_\epsilon) = \mathrm{Var}(F_\epsilon^+) + \mathrm{Var}(F_\epsilon^-).$$

Therefore both F_ϵ^+ and F_ϵ^- are in L^∞ and indeed have finite limits at $-\infty$ and ∞. Then we have

$$\begin{aligned} |\Sigma(F, A_n) - \Sigma(F, B_{n,m})| \quad \leq \quad & \left|\Sigma(F^+, A_n) - \Sigma(F^+, B_{n,m})\right| + \\ & \left|\Sigma(F^-, A_n) - \Sigma(F^-, B_{n,m})\right| \end{aligned} \tag{61}$$

so that it enough to manipulate separately the quantities involving F^+ and F^- and the BV case is reduced to the bounded monotone case.

Consider a nondecreasing function $G \in L^\infty$ and consider the quantity

$$\left|\Sigma(G, A_n) - \Sigma(G, B_{n,m})\right|.$$

For any matrix A, let $S(A)$ be the vector of its singular values ordered non increasingly. Let $S(B_{n,m}, q)$, q integer number, be so that $(S(B_{n,m}, q))_i = (S(B_{n,m}))_{i+q}$, $i = 1, \ldots, d_n$ where $(S(B_{n,m}))_j = 0$ if $j \geq d_n + 1$ and $(S(B_{n,m}))_j = \max\{(S(A))_1, (S(B_{n,m}))_1\}$ if $j \leq 0$. Now, by the Cauchy interlace Theorem (see e.g. [7]), it is clear that

$$S(B_{n,m}, -2c(m)d_n) \geq S(B_{n,m}), S(A_n) \geq S(B_{n,m}, 2c(m)d_n)$$

where "$\geq$" is intended componentwise and $c(m)$ is the function considered in (59) and (60). Finally by monotonicity we deduce that

$$|\Sigma(G, A_n) - \Sigma(G, B_{n,m})| \leq$$

$$\left|\frac{1}{d_n} \sum_{i=1-2k,\ldots,2k, j=d_n-2k+1,\ldots,d_n+2k} G(\sigma_i(B_{n,m})) - G(\sigma_j(B_{n,m}))\right| \leq \frac{16k}{d_n}\|G\|_\infty$$

where $k = 2c(m)d_n$. Therefore we conclude that

$$|\Sigma(G, A_n) - \Sigma(G, B_{n,m})| \quad \leq \quad 32c(m)\|G\|_\infty. \tag{62}$$

Now we choose $\epsilon_m \to 0$ such that

$$\max\{\|F_{\epsilon_m}^-\|_\infty, \|F_{\epsilon_m}^+\|_\infty\} \leq (c(m))^{-1/2}$$

and finally by considering (61) and (62) we get

$$|\Sigma(F, A_n) - \Sigma(F, B_{n,m})| \leq \theta(m)$$

with $\theta(m) = 2\epsilon_m + 64(c(m))^{1/2}$ and the proof is over. •

Now we are ready for proving a result which is very useful for dealing with asymptotic distribution problems. Its proof can be found in [74] and it is essentially based on arguments introduced by Tilli in Proposition 2.7 of [90].

Proposition 9.1 *Let d_n be an increasing sequence of natural numbers. Suppose a sequence of matrices $\{A_n\}$ of size d_n is given such that $\{\{B_{n,m}\}\}_m$, $m \in \hat{\mathbf{N}} \subset \mathbf{N}$, $\#\hat{\mathbf{N}} = \infty$, is an a.c.s. for $\{A_n\}$ in the sense of Definition 9.2. Suppose that, for all sufficiently large $m \in \mathbf{N}$ and for all $F \in \mathcal{C}_0$, there exist the limits*

$$\lim_{n\to\infty} \Sigma(F, B_{n,m}) = \Phi_m(F), \quad \textit{and} \quad \lim_{m\to\infty} \Phi_m(F) = \Phi(F). \tag{63}$$

Then it necessarily holds

$$\lim_{n\to\infty} \Sigma(F, A_n) = \Phi(F), \quad \forall F \in \mathcal{C}_0. \tag{64}$$

Proof. Observe that, from (63), it easily follows that Φ and Φ_m are bounded linear functionals over $\mathcal{C}_0$, and $|\Phi(F)|, |\Phi_m(F)| \le \|F\|_\infty$. For any large m and for all $n > n_m$ it holds

$$|\Sigma(F, A_n) - \Phi(F)| \le \alpha_{n,m} + \beta_{n,m} + \gamma_{n,m} + |\Phi_m(F) - \Phi(F)|, \tag{65}$$

where

$$\alpha_{n,m} = |\Sigma(F, A_n) - \Sigma(F, B_{n,m} + N_{n,m})|,$$
$$\beta_{n,m} = |\Sigma(F, B_{n,m} + N_{n,m}) - \Sigma(F, B_{n,m})|,$$
$$\gamma_{n,m} = |\Sigma(F, B_{n,m}) - \Phi_m(F)|.$$

From Lemma 9.2 it follows $\limsup_{n\to\infty} \alpha_{n,m} \le \theta_1(m)$, for all m. From Lemma 9.3 we have $\limsup_{n\to\infty} \beta_{n,m} \le \theta_2(m), \quad \forall m$. From assumption (63) it follows $\limsup_{n\to\infty} \gamma_{n,m} = 0$, for all m. Since $\limsup$ is subadditive, from (65) we obtain

$$\begin{aligned} \limsup_{n\to\infty} |\Sigma(F, A_n) - \Phi(F)| \quad &\le \quad \theta_1(m) + \theta_2(m) + \\ &\qquad |\Phi_m(F) - \Phi(F)|, \quad \forall m. \end{aligned} \tag{66}$$

Taking the limit for $m \to \infty$, we obtain that the above $\limsup$ is zero, and (64) follows. •

It is evident that all the previous results can be adapted to the eigenvalues in the case of Hermitian matrices by using $\Sigma_\lambda(F, \cdot)$ in place of $\Sigma(F, \cdot)$. Moreover we stress that Proposition 9.1 is crucial for handling spectral distributions in the sense of Definition 9.1: in fact, with the choice of the functionals Φ and Φ_m of the form

$$\Phi[g](F) = \frac{1}{(2\pi)^d} \int_{I^d} F(|g(x)|)\, dx, \quad F \in \mathcal{C}_0,$$

g measurable function over a set K of finite measure we obtain that the result claimed in Proposition 9.1 reduces to the claim $\{A_n\} \sim_\sigma (f, K)$ where $\Phi = \Phi[f]$, $\Phi_m = \Phi[f_m]$ and f_m converging in measure to f. The latter reasoning is resumed in the following corollary.

Corollary 9.1 *Let d_n be an increasing sequence of natural numbers. Suppose a sequence of matrices $\{A_n\}$ of size d_n is given such that $\{\{B_{n,m}\}\}_m$, $m \in \hat{\mathbf{N}} \subset \mathbf{N}$, $\#\hat{\mathbf{N}} = \infty$, is an a.c.s. for $\{A_n\}$ in the sense of Definition 9.2. Suppose that, for all sufficiently large $m \in \mathbf{N}$*

$$\{B_{n,m}\} \sim_\sigma (f_m, K),\ m\{K\} < \infty \quad \textit{and} \quad \lim_{m\to\infty} f_m = f \ \text{ in measure.} \tag{67}$$

Then it necessarily holds

$$\{A_n\} \sim_\sigma (f, K). \tag{68}$$

Analogously, if A_n and $B_{n,m}$ are definitely Hermitian, suppose that, for all sufficiently large $m \in \mathbf{N}$,

$$\{B_{n,m}\} \sim_\lambda (f_m, K),\ m\{K\} < \infty \quad \textit{and} \quad \lim_{m\to\infty} f_m = f \ \text{ in measure.} \tag{69}$$

Then it necessarily holds

$$\{A_n\} \sim_\lambda (f, K). \tag{70}$$

9.1.1 Algebraization of matrix sequences

The aim of this section is to show that the notion of approximating class of sequences (a.c.s.) is stable under algebraization i.e. under linear combinations and products: the desired result is employed for reducing the spectral analysis of complicate matrix sequences to the spectral analysis of simple matrix sequences via Proposition 9.1.

First we recall that a sequence of matrices $\{A_n\}$ is sparsely unbounded (s.u.) if and only if, by definition, $\forall M > 0$, $\exists \bar{n}_M$ such that for $n \geq \bar{n}_M$ we have

$$\frac{\#\{j : \sigma_j(A_n) > M\}}{d_n} \leq r(M), \quad \lim_{M\to\infty} r(M) = 0. \tag{71}$$

Therefore by invoking the singular value decomposition, we have

$$A_n = A^{(1)}_{n,M} + A^{(2)}_{n,M}, \ \ \|A^{(1)}_{n,M}\| \leq M, \ \ \mathrm{rank}(A^{(2)}_{n,M}) \leq r(M)d_n. \tag{72}$$

It is almost trivial to see that if $\{A_n\} \sim_\sigma \theta$ with measurable θ taking values on $\mathbf{C} \cup \{\infty\}$, then $\{A_n\}$ s.u if and only if θ is sparsely unbounded that is $\lim_{M\to\infty} m\{x : |\theta(x)| > M\} = 0$. Furthermore, we observe that any function a belonging to $L^1(I^d)$ is sparsely unbounded and that the product $\nu(x)$ of a finite number of s.u. functions is s.u. since the Lebesgue measure of the set where $|\nu(x)| = \infty$ is zero.

Proposition 9.2 *Let $\{A_n\}$ and $\{B_n\}$, $A_n, B_n \in M_{d_n}(\mathbf{C})$, be two given sparsely unbounded (s.u.) matrix sequences. Suppose that*

$$\{\{Y_{A,n,m}\}\}_m \ \text{ and } \ \{\{Y_{B,n,m}\}\}_m,$$

$m \in \hat{\mathbf{N}} \subset \mathbf{N}$, $\#\hat{\mathbf{N}} = \infty$, are two a.c.s., in the sense of Definition 9.2, for $\{A_n\}$ and $\{B_n\}$, respectively. Then $\{\{Y_{A,n,m}Y_{B,n,m}\}\}_m$ is an a.c.s. for the sequence $\{A_nB_n\}$.

Proof. Consider the product A_nB_n. Then by exploiting the splittings of A_n and B_n given in (58) and (59), we have

$$\begin{aligned} A_nB_n \ &= \ Y_{A,n,m}Y_{B,n,m} + R_{A,n,m}B_n + \\ &\quad N_{A,n,m}B_n + Y_{A,n,m}N_{B,n,m} + Y_{A,n,m}R_{B,n,m} \end{aligned}$$

with

$$\begin{gathered} \max\{\|N_{A,n,m}\|, \|N_{B,n,m}\|\} \leq \omega(m), \\ \max\{\mathrm{rank}(R_{A,n,m}), \mathrm{rank}(R_{B,n,m})\} \leq c(m)d_n \end{gathered}$$

and

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

Therefore $\mathrm{rank}(R_{A,n,m}B_n + Y_{A,n,m}R_{B,n,m}) \leq 2c(m)d_n$. In order to prove that $\{\{Y_{A,n,m}Y_{B,n,m}\}\}$ is an a.c.s. for $\{A_nB_n\}$, we still need to prove that the matrix $N_{A,n,m}B_n + Y_{A,n,m}N_{B,n,m}$ can be decomposed as a sum of a term bounded in spectral norm by a quantity depending only on m and going to zero as m tends to infinity and of a term whose rank divided by d_n is bounded from above by another quantity depending only on m and going to zero as m tends to infinity. This is proved by using the assumption that $\{A_n\}$ and $\{B_n\}$ are s.u. Indeed, by this hypothesis, by (72), and choosing $M \equiv M_m = [\omega(m)]^{-1/2}$, it follows that

$$A_n = A^{(1)}_{n,M_m} + A^{(2)}_{n,M_m}, \ \ B_n = B^{(1)}_{n,M_m} + B^{(2)}_{n,M_m},$$

with

$$\|A^{(1)}_{n,M_m}\|, \|B^{(1)}_{n,M_m}\| \le M_m, \;\; \mathrm{rank}(A^{(2)}_{n,M_m}), \mathrm{rank}(B^{(2)}_{n,M_m}) \le \theta(M_m)d_n.$$

Consequently we have

$$\begin{aligned} N_{A,n,m}B_n + Y_{A,n,m}N_{B,n,m} &= N_{A,n,m}B^{(1)}_{n,M_m} + N_{A,n,m}B^{(2)}_{n,M_m} + \\ &\quad A_nN_{B,n,m} - N_{A,n,m}N_{B,n,m} - R_{A,n,m}N_{B,n,m} \\ &= N_{A,n,m}B^{(1)}_{n,M_m} + N_{A,n,m}B^{(2)}_{n,M_m} + \\ &\quad A^{(1)}_{n,M_m}N_{B,n,m} + A^{(2)}_{n,M_m}N_{B,n,m} - \\ &\quad -N_{A,n,m}N_{B,n,m} - R_{A,n,m}N_{B,n,m} \end{aligned}$$

where

$$\|N_{A,n,m}B^{(1)}_{n,M_m}\|, \|A^{(1)}_{n,M_m}N_{B,n,m}\| \le [\omega(m)]^{1/2},$$

$$\mathrm{rank}(N_{A,n,m}B^{(2)}_{n,M_m}), \mathrm{rank}(A^{(2)}_{n,M_m}N_{B,n,m}) \le \theta(M_m)d_n,$$

$$\|N_{A,n,m}N_{B,n,m}\| \le [\omega(m)]^2 \;\text{ and }\; \mathrm{rank}(R_{A,n,m}N_{B,n,m}) \le c(m)d_n.$$

By using the subadditivity of the rank and of the norm the claimed thesis follows. •

9.2 B.2. Applications to structured matrix sequences

Let f be a d variate (2π)-periodic complex valued (Lebesgue) integrable function, defined over the hypercube I^d, with $I = [-\pi, \pi)$ and $d \ge 1$. From the Fourier coefficients of f

$$a_j = \frac{1}{(2\pi)^d}\int_{I^d} f(x)e^{-\mathbf{i}(j,x)}\,dx, \qquad \mathbf{i}^2 = -1, \quad j = (j_1, \ldots, j_d) \in \mathbf{Z}^d$$

with $x = (x_1, \ldots, x_d)$, $(j, x) = \sum_{k=1}^d j_k x_k$, $n = (n_1, \ldots, n_d)$ and $N(n) = n_1 \cdots n_d$, we consider the sequence of Toeplitz matrices $\{T_n(f)\}$, where $T_n(f) = \{a_{j-i}\}^n_{i,j=e^T} \in M_{N(n)}(\mathbf{C})$, $e^T = (1, \ldots, 1) \in \mathbf{N}^d$ is said to be the Toeplitz matrix of order n generated by f. We recall that $n \to \infty$ with $n = (n_1, \ldots, n_d)$ being a multi-index, is equivalent to write $\min_{1\le j\le d} n_j \to \infty$.

The asymptotic distribution of eigenvalues and singular values of a sequence of Toeplitz matrices has been thoroughly studied in the last century (for example see [96, 91, 16] and the references reported therein). The starting point of this theory, which contains many extensions and other results, is a famous theorem of Szegő [37], which we report in the Tyrtyshnikov and Zamarashkin generalized version:

Theorem 9.1 (Tyrtyshnikov-Zamarashkin, [96]) *If f is integrable over I^d, and if $\{T_n(f)\}$ is the sequence of Toeplitz matrices generated by f, then it holds*

$$\{T_n(f)\} \sim_\sigma (f, I^d). \tag{73}$$

Moreover, if f is also real valued, then each matrix $T_n(f)$ is Hermitian and

$$\{T_n(f)\} \sim_\lambda (f, I^d). \tag{74}$$

The following two preparatory lemmas are the basic building blocks for the proof of Theorem 9.1.

Lemma 9.4 *Let p be a complex d variate polynomial. Then $\{\{\mathcal{A}_n(p)\}\}_m$ is an a.c.s. for $\{T_n(p)\}$ where $\mathcal{A}_n$ denotes the space of the circulant matrices and the symbol $\mathcal{A}_n(p)$ has been defined in (33).*

Proof. By Theorem 5.1, for $d = 1$, we have

$$T_n(p) = \mathcal{A}_n(p) + R_{n,m}$$

with $\mathrm{rank}(R_{n,m}) \leq c$ with c constant depending only on the degree of p. Therefore by virtue of Definition 9.2 we deduce that $\{\{\mathcal{A}_n(p)\}\}_m$ is an a.c.s. for $\{T_n(p)\}$. For $d \geq 2$ we reduce the analysis to the univariate case. The complex polynomial p can be written as a finite linear combination of monomials of the form $m(x) = \prod_{j=1}^{d} e^{\mathbf{i}\alpha_j x_j}$: as a consequence $T_n(p)$ and $\mathcal{A}_n(p)$ can be written according to the same linear combination of

$$T_{n_1}(e^{\mathbf{i}\alpha_1 x_1}) \otimes - \cdots \otimes T_{n_d}(e^{\mathbf{i}\alpha_d x_d}), \quad \mathcal{A}_{n_1}(e^{\mathbf{i}\alpha_1 x_1}) \otimes - \cdots \otimes \mathcal{A}_{n_d}(e^{\mathbf{i}\alpha_d x_d})$$

respectively. But every $T_{n_j}(e^{\mathbf{i}\alpha_j x_j})$ differs from the corresponding circulant approximation $\mathcal{A}_{n_j}(e^{\mathbf{i}\alpha_j x_j})$, $j = 1, \ldots, d$, by a term of constant rank and therefore (by linearity)

$$T_n(p) = \mathcal{A}_n(p) + R_{n,m}$$

with $\mathrm{rank}(R_{n,m}) \leq cN(n)(\sum_{j=1}^{d} n_j^{-1})$ with c universal constant. Since $N(n)(\sum_{j=1}^{d} n_j^{-1}) = o(N(n))$ by Definition 9.2 the desired result directly follows. •

Lemma 9.5 *Let $f \in L^1(I^d)$ and $\{p_m\}$ be a sequence of polynomials converging to f in the L^1 norm. Then $\{\{T_n(p_m)\}\}_m$ is an a.c.s. for $\{T_n(f)\}$.*

Proof. We point out that, for any m, the sequence $\{T_n(f) - T_n(p_m)\}$ coincides with $\{T_n(f - p_m)\}$ and therefore it is enough to exploit the singular value decomposition of $T_n(f - p_m)$ for large m and n and the assumption that p_m converges in L^1 norm (and therefore in measure) to f. Indeed, by the assumption, there exists a function $k(m)$ going to zero such that $\|f - p_m\|_{L^1} \leq (2\pi)^d\, k(m)$; moreover for any $f \in L^p$ with $p \in [1, \infty)$, we have (see [83])

$$\|T_n(f)\|_p^p \leq (2\pi)^{-d}\, N(n) \|f\|_{L^p}^p. \tag{75}$$

Therefore, by considering the case of $p = 1$ the relation

$$\|T_n(f - p_m)\|_p = \sum_{j=1}^{N(n)} \sigma_j(T_n(f - p_m)) \leq N(n)k(m) \tag{76}$$

holds $\forall n$ and $\forall m$. Therefore we have $\#\{j :\ \sigma_j(T_n(f - p_m)) > \sqrt{k(m)}\} \leq N(n)\sqrt{k(m)}$ and, by the singular value decomposition, we deduce that

$$T_n(f - p_m)) = R_{n,m} + N_{n,m}$$

with $\mathrm{rank}(R_{n,m}) \leq N(n)\, c(m)$ and $\|N_{n,m}\| \leq \omega(m)$, where $c(m) = \omega(m) = \sqrt{k(m)}$. •

Proof of Theorem 9.1

Step 1. For every trigonometric polynomial p, the circulant matrix $\mathcal{A}_n(p)$ has eigenvalues given by

a equispaced sampling of p on the definition domain I^d. Moreover if p is real valued then $\mathcal{A}_n(p)$ is Hermitian. Therefore by a direct check we obtain that

$$\{\mathcal{A}_n(p)\} \sim_\sigma (p, I^d);$$

moreover if p is real valued then we also have

$$\{\mathcal{A}_n(p)\} \sim_\lambda (p, I^d).$$

Step 2. $\{\{\mathcal{A}_n(p)\}\}_m$ is an a.c.s. for $\{T_n(p)\}$ by Lemma 9.4 and $\{\mathcal{A}_n(p)\} \sim_\sigma (p, I^d)$ by **Step 1**. Therefore by Corollary 9.1 we have $\{T_n(p)\} \sim_\sigma (p, I^d)$. Moreover, if p is real valued then both $\mathcal{A}_n(p)$ and $T_n(p)$ are Hermitian with $\{\mathcal{A}_n(p)\} \sim_\lambda (p, I^d)$ by **Step 1** and hence again by Corollary 9.1 it follows $\{T_n(p)\} \sim_\lambda (p, I^d)$.
Step 3. By Lemma 9.5, given any Lebesgue integrable function f there exists a sequence of trigonometric polynomials p_m converging in L^1 norm (and a fortiori in measure) to f such that $\{\{T_n(p_m)\}\}_m$ is an a.c.s. for $\{T_n(f)\}$. Furthermore we have $\{T_n(p_m)\} \sim_\sigma (p_m, I^d)$ by **Step 2**. Consequently by Corollary 9.1, we infer $\{T_n(f)\} \sim_\sigma (f, I^d)$. Finally, if f is real valued, then we can choose p_m real valued and therefore the matrices $T_n(f)$ and $T_n(p_m)$ are all Hermitian. Since $\{T_n(p_m)\} \sim_\lambda (p_m, I^d)$ by **Step 2**, the use of Corollary 9.1 allows one to conclude $\{T_n(f)\} \sim_\lambda (f, I^d)$. •

9.2.1 The algebra of Toeplitz sequences

In this subsection Theorem 9.1 is used as basic block for proving distributional results for more involved sequences.

Lemma 9.6 *Let f and g be two functions belonging to $L^1(I^d)$. Then*

$$\{T_n(f)T_n(g)\}$$

is distributed as the measurable function fg.

Proof. We remark that $\{T_n(f)\}$ and $\{T_n(g)\}$ are sparsely unbounded since they are distributed as sparsely unbounded functions. Let $\{p_{f,m}\}$ and $\{p_{g,m}\}$ be two sequences of polynomials converging to f and g in the L^1 norm, respectively. By Lemma 9.5, $\{\{T_n(p_{f,m})\}\}_m$ is an a.c.s. for $\{T_n(f)\}$ and $\{\{T_n(p_{g,m})\}\}_m$ is an a.c.s. for $\{T_n(g)\}$. Then, by Proposition 9.2, we conclude that the collection $\{\{T_n(p_{f,m})T_n(p_{g,m})\}\}_m$ is an a.c.s. for $\{T_n(f)T_n(g)\}$. But by direct inspection we see that $T_n(p_{f,m})T_n(p_{g,m}) = T_n(p_{f,m}p_{g,m}) + R_{n,m}$ where $R_{n,m}$ is of rank bounded by $k(m)N(n)\sum_{j=1}^d n_j^{-1}$ (recall that $T_n(p)$ is a multilevel band matrix if p is polynomial of fixed degree) and therefore, by definition, $\{\{T_n(p_{f,m}p_{g,m})\}\}_m$ is an a.c.s. for $\{T_n(f)T_n(g)\}$ as well. Finally, by Theorem 9.1, we have

$$\{T_n(p_{f,m}p_{g,m})\} \sim_\sigma p_{f,m}p_{g,m}$$

and $p_{f,m}p_{g,m}$ converges in measure to fg so that the application of Corollary 9.1 concludes the proof. •

Lemma 9.7 *Let k be a positive natural number and $\{f_\alpha : \alpha = 1, \ldots, k\}$ be a finite set of functions belonging to $L^1(I^d)$. Then*

$$\{\prod_{\alpha=1}^k T_n(f_\alpha)\}$$

is distributed as the measurable function $\prod_{\alpha=1}^k f_\alpha$.

Proof. It is substantially the same argument as in Lemma 9.6. If $\{p_{f_\alpha,m}\}$ is a sequence of polynomials converging in L^1 norm to f_α then $\prod_{\alpha=1}^k p_{f_\alpha,m}$ converges in measure to $\prod_{\alpha=1}^k f_\alpha$. A repeated application of Proposition 9.2 tells us that $\{\{\prod_{\alpha=1}^k T_n(p_{f_\alpha,m})\}\}_m$ is an a.c.s. for $\{\prod_{\alpha=1}^k T_n(f_\alpha)\}$ and therefore, since any $p_{f_\alpha,m}$ is a polynomial, the collection $\{\{T_n(\prod_{\alpha=1}^k p_{f_\alpha,m})\}\}_m$ is a new a.c.s. for $\{\prod_{\alpha=1}^k T_n(f_\alpha)\}$. Finally, the use of Theorem 9.1 and a final application of Corollary 9.1 conclude the proof. •

Theorem 9.2 *Let k and q_α, $\alpha = 1, \ldots, k$ be positive natural numbers and $\{f_{\alpha,\beta} : \alpha = 1, \ldots, k,\ \beta = 1, \ldots, q_\alpha\}$ be a finite set of functions belonging to $L^1(I^d)$. Then*

$$\{\sum_{\alpha=1}^k \prod_{\beta=1}^{q_\alpha} T_n(f_{\alpha\beta})\} \sim_\sigma \theta = \sum_{\alpha=1}^k \prod_{\beta=1}^{q_\alpha} f_{\alpha\beta}.$$

Proof. It is enough to remark that the sum of a finite collection $\{\{B_{n,m}^{(\alpha)}\}\}_m$ of a.c.s. for the sequence $\{A_n^{(\alpha)}\}$ is an a.c.s. for $\{\sum_\alpha A_n^{(\alpha)}\}$. In our case, by Lemma 9.7, we deduce that

$$A_n^{(\alpha)} = \prod_{\beta=1}^{q_\alpha} T_n(f_{\alpha\beta})$$

and

$$B_{n,m}^{(\alpha)} = T_n(\prod_{\beta=1}^{q_\alpha} p_{f_{\alpha\beta},m})$$

where $p_{f_{\alpha\beta},m}$ converges in the L^1 norm to $a_{\alpha\beta}$. To conclude, use Theorem 9.1 and the key Corollary 9.1 as in the previous lemmas. •

We end this subsection with two remarks.
• If $\theta = \sum_{\alpha=1}^k \prod_{\beta=1}^{q_\alpha} f_{\alpha\beta}$ belongs to the L^1 class then it makes sense to consider the sequence $\{T_n(\theta)\}$. Indeed it is easy to see that

$$\{\sum_{\alpha=1}^k \prod_{\beta=1}^{q_\alpha} T_n(f_{\alpha\beta})\} \text{ and } \{T_n(\theta)\}$$

are equally distributed since $\{\sum_{\alpha=1}^k \prod_{\beta=1}^{q_\alpha} T_n(f_{\alpha\beta}) - T_n(\theta)\}$ is clustered at zero (to see this it is enough to think again to the proofs of Lemma 9.6, Lemma 9.7 and Theorem 9.2 by taking into account that $\theta \in L^1$).

• The most classical (and successful) approach to the asymptotics for finite Toeplitz structures consists in using the corresponding infinite dimensional Toeplitz operators $T(\cdot) = T_\infty(\cdot)$ (see e.g. [16]). This clearly works if the symbols are univariate and continuous since $T(f)T(g) = T(fg) + \mathcal{K}$ where $\mathcal{K}$ is a compact operator (see the beautiful formula due to Widom [102]). However if f and g belong to L^1 (or if the symbols f and g are multivariate), then the above formula is not well defined since, in general, fg may fail to belong to L^1 so that we cannot give sense to $T(fg)$ (or $\mathcal{K}$ is not compact if we consider the multivariate case). Hence, the "finite dimensional" approach described in this appendix seems to be more versatile and flexible at least in this context.

9.3 B.3. Further generalizations

Further results concerning spectral distribution formulas and other asymptotics of structured matrix sequences can be found in [16] and reference therein and in [67, 90, 78, 91, 73, 80]. We recall that

this kind of asymptotics are useful for an efficient numerical solution of several problems in applied mathematics arising in signal and image processing, time series, PDEs etc. (see e.g. [21]). Here we just give some examples of application and some theoretical extensions to differential problems.

Let $f(s) = (2 - 2\cos(s))^2 = 6 - 8\cos(s) + 2\cos(2s) = 6 - 4e^{\mathbf{i}s} - 4e^{-\mathbf{i}s} + e^{2\mathbf{i}s} + e^{-2\mathbf{i}s}$: then we have

$$T_n(f) = \begin{pmatrix} 6 & -4 & 1 & & \\ -4 & \ddots & \ddots & \ddots & \\ 1 & \ddots & & \ddots & 1 \\ & \ddots & \ddots & \ddots & -4 \\ & & 1 & -4 & 6 \end{pmatrix}. \tag{77}$$

This matrix is related to the fourth order derivative. Indeed if we consider the model problem

$$(b(x)u'')'' = f(x), \;\; x \in \Omega = (0,1),$$

with $b(x) \leq b_* > 0$ and homogeneous boundary conditions, its centered equispaced Finite Differences discretization of precision order 2 with minimal bandwidth leads to a system for $A_n(b)\mathbf{u} = h^4\mathbf{f}$, $h = (n+1)^{-1}$ where $A_n(1) = T_n(f)$. In the general case, by mirroring the function b outside the domain $\Omega = (0,1)$, we find

$$A_n(b) = \sum_{j \in \mathbf{Z}} b_j^{(n)} \begin{pmatrix} 0 \\ -1 \\ 2 \\ -1 \\ 0 \end{pmatrix}_j \begin{pmatrix} 0 \\ -1 \\ 2 \\ -1 \\ 0 \end{pmatrix}_j^T, \quad b_j^{(n)} = b(jh), \tag{78}$$

where

$$\begin{pmatrix} 0 \\ -1 \\ 2 \\ -1 \\ 0 \end{pmatrix}_j = 2\mathbf{e}_j - \mathbf{e}_{j-1} - \mathbf{e}_{j+1} \in \mathbf{R}^n, \;\; j \in \mathbf{Z}$$

with $\mathbf{e}_k$ denoting the k-th vector of the canonical basis and with the understanding that $\mathbf{e}_k = 0$ if $k \leq 0$ or $k \geq n+1$. The following facts hold:

- $A_n(1) = T_n(f)$, $f(s) = (2 - 2\cos(s))^2 = 6 - 8\cos(s) + 2\cos(2s)$,
- $A_n(\cdot)$ is a linear positive operator;
- $(\inf_\Omega b)\, A_n(1) \leq A_n(b) \leq (\sup_\Omega b)\, A_n(1)$,
- $T_n(p)$ is an optimal preconditioner for $A_n(b)$,
- $\lambda_{\min}(A_n(b)) \sim \lambda_{\min}(T_n(f)) \sim n^{-4}$,
- $\{T_n(p)\} \sim_\lambda (f, I)$.

The first three items are direct consequences of the diadic representation in (78) and of the mirroring boundary conditions; the fourth relation and $\lambda_{\min}(A_n(b)) \sim \lambda_{\min}(T_n(f))$ (first part of the fifth relation) are a consequence of the third one and the positivity of b. Finally $\lambda_{\min}(T_n(f)) \sim n^{-4}$ is a consequence of Corollary 2.1 since $f \geq 0$ has a unique zero of order 4 and the sixth is a special case of the Szegö Theorem.

Now we study the spectral distribution of the positive definite sequence distribution of $\{A_n(b)\}$. For every $m > 0$, we use centered equispaced Finite Differences of precision order 2 with minimal bandwidth for

$$(b_m(x)u'')'' = f(x), \;\; x \in \Omega_j = (x_j, x_{j+1}),$$

with $x_j \equiv x_j^{(m)} = j/m$, $j = 0, \ldots, m-1$, $b_m(x) = b(x_j)$ if $x \in \Omega_j$ and homogeneous boundary conditions on every Ω_j.
Therefore we find a global system $A_{n,m}\mathbf{u} = h^4\mathbf{f}_{n,m}$, $h = (n+1)^{-1}$ where $A_{n,m} = \oplus_{j=0}^{m-1} b(x_j)T_{n/m}(p)$ i.e.

$$A_{n,m} = \begin{pmatrix} b(x_0)T_{n/m}(p) & & & \\ & \ddots & & \\ & & \ddots & \\ & & & b(x_{m-1})T_{n/m}(p) \end{pmatrix}.$$

The following items are true:

- The eigenvalues of $A_{n,m}$ are $\{b(j/m)\lambda_i(T_{n/m}(p)) : j = 0, \ldots, m-1, i = 1, \ldots, n/m\}$;
- $\{A_{n,m}\} \sim_\lambda (b_m(x)(2-2\cos(s))^2, [0,1] \times I))$;
- $b_m(x)(2-2\cos(s))^2 \to_\mu b(x)(2-2\cos(s))^2$ (convergence in measure);
- $\{\{A_{n,m}\}\}_m$ is an a.c.s. for $\{A_n(b)\}$;

where the last is implied by

$$\begin{aligned} \|[A_n(b) - A_{n,m}]_{(j-1)m:jm}\| &= \|[A_n(b)]_{(j-1)m:jm} - b(j/m)T_{n/m}(p)\| \\ &\leq \|p\|_\infty \omega_b(1/m), \\ \inf_{\|Y\|\leq\epsilon} \{\operatorname{rank}[\Delta_{n,m} + Y]\} &\leq 2m \\ \Delta_{n,m} &= A_n(b) - \oplus_{j=0}^{m-1}[A_n(b)]_{(j-1)m:jm}. \end{aligned}$$

here with the MATLAB like notation $X_{s:t}$, $t \geq s$, we denote the submatrix of X of size $t-s+1$ defined by rows and columns of X in the range $\{s, s+1, \ldots, t\}$. From Corollary 9.1 we have

$$\{A_n(b)\} \sim_\lambda (b(x)(2-2\cos(s))^2, [0,1] \times I)).$$

9.3.1 Multivariate generalizations

We consider second order elliptic PDEs in d-dimensional domains: in the first case we have a discretization by centered formulae on a equispaced grid; then we consider tensor grids which are not equispaced and finally an example of discretization by linear Finite Elements (see [78, 81, 6]). In the following, $\circ$ denotes the component wise "Hadamard product" and H_u is the Hessian i.e. the dyad of operators:

$$H_u = \begin{pmatrix} \frac{\partial}{\partial x_1} \\ \vdots \\ \frac{\partial}{\partial x_d} \end{pmatrix} \begin{pmatrix} \frac{\partial}{\partial x_1} \\ \vdots \\ \frac{\partial}{\partial x_d} \end{pmatrix}^T.$$

With the previous notations the following PDE

$$\begin{gathered} -\textstyle\sum_{i,j=1}^d \frac{\partial}{\partial x_i}\left(b_{i,j}(x)\frac{\partial}{\partial x_j}u(x)\right) = b(x), \\ \text{if} \quad x \in \Omega_d^\circ \subset (0,1)^d, \\ + \qquad \text{boundary conditions,} \end{gathered}$$

is written as

$$\begin{gathered} -e^T[B(x) \circ H_u(x)]e + \text{first order terms} = b(x) \\ \text{if} \quad x \in \Omega_d^\circ \subset (0,1)^d, \\ + \qquad \text{boundary conditions.} \end{gathered}$$

Denoting $e^T = (1, \ldots, 1)$ and by $A_n(B, P, W_\alpha)$ the discretization of the former problem, we have [78]

$$\{A_n(B, P, W_\alpha)\} \sim_{\sigma,\lambda} (e^T[B(x) \circ P(s) \circ W_\alpha]e, \Omega_d \times I^d)$$

where $P(s) \circ W_\alpha$ is the Finite Differences representation of the operator matrix $-H_u$ on the grid $\mathcal{G}$ with $n_j + 1 = \alpha_j v$. Here the grid sequence $\mathcal{G} = \{\mathcal{G}_n\}$, $\mathcal{G}_n = \mathcal{G}_{n_1} \times \cdots \times \mathcal{G}_{n_d}$

$$\mathcal{G}_{n_i} = \left\{ \frac{j}{n_i + 1} : \ j = 1, \ldots, n_i \right\}.$$

Moreover, the dyad of functions

$$P(s) \circ W_\alpha = \begin{pmatrix} p_1(s_1)\alpha_1 \\ \vdots \\ p_d(s_d)\alpha_d \end{pmatrix} \begin{pmatrix} p_1(s_1)\alpha_1 \\ \vdots \\ p_d(s_d)\alpha_d \end{pmatrix}^H,$$

represent the FD formulae on the Fourier domain. In the case of non uniform grids we find (see e.g. [81])

$$\{A_n(B, P, W_\alpha)\} \sim_{\sigma,\lambda} (e^T J^{-1}(x)[B(\Phi(x)) \circ P(s) \circ W_\alpha]J^{-T}(x)e, \Omega_d \times I^d)$$

where $P(s) \circ W_\alpha$ is the Finite Differences representation of the operator matrix $-H_u$ on the grid $\mathcal{G}$ with $n_j + 1 = \alpha_j v$, $\Phi(\mathcal{G})$ is the non uniform grid and $J(x)$ is the Jacobian of Φ.

Finally, when considering linear Finite Elements, we have [16]

$$\{A_n(B, T)\} \sim_\lambda |\det(J(x))| \cdot e^T J^{-1}(x)[B(\Phi(x)) \circ P(s)]J^{-T}(x)e, \Omega_d \times I^d)$$

where $P(s)$ is the Finite Elements representation of the operator matrix $-H_u$ over the uniform triangulation $\mathcal{U}$, Φ is the transform such that $T = \Phi(\mathcal{U})$ and $J(x)$ is the Jacobian of Φ.

We recall that in all the three cases (i.e. uniform and non uniform FD and linear Finite Elements), the hypotheses on the problem data are very weak i.e.

• The domain Ω should be at least measurable according to Peano-Jordan.

• The coefficients $b_{i,j}$ should be at least integrable in the Riemann sense.

9.3.2 Spectral distributions and preconditioning

Finally, we just mention that the same tools (a.c.s. + Corollary 9.1 + algebraization results) can be used for finding preconditioning results: let $T_n = A_n(1)$, $D_n(b)$ "diagonal and scaled part of $A_n(b)$", then we set

$$P_n(b) = D_n^{1/2}(b) T_n D_n^{1/2}(b).$$

Corollary 9.1 + the algebraization of the a.c.s. imply

$$\{P_n^{-1}(b)A_n(b)\} \sim_\lambda \ (1, [0, 1]^d \times I^d)$$

which is another way of writing that the preconditioned sequence $\{P_n^{-1}(b)A_n(b)\}$ is weakly clustered at 1 in the sense of the eigenvalues.

Under the regularity assumptions $C^2([0, 1]^d)$ of b, it has been proven [70, 80] the proper clustering at the unity of $\{P_n^{-1}(b)A_n(b)\}$ in the eigenvalue sense and the uniform spectral boundedness of $\{P_n^{-1}(b)A_n(b)\}$ and $\{A_n^{-1}(b)P_n(b)\}$. Therefore, by Theorem 8.3, we observe an optimal convergence and superlinear behavior of the associated PCG method.

9.3.3 Final remarks

The distribution results can be useful in many directions: among them we recall a finer analysis of the convergence of (P)CG methods according to the results of Beckermann and Kuijlaars [5], heuristics for the preconditioning, spectral information to be utilized for the design of fast iterative methods (especially of multigrid type).

Some open questions remain and, in particular, future developments should include:

- a uniform approach for a larger class of Finite Elements in analogy with the Finite Differences case;
- a stochastic approach to the convergence theory of iterative solvers: in that case, it is interesting to point out that the global distribution of the spectra plays a role in place of the spectral radius.