ELSEVIER Linear Algebra and its Applications 337 (2001) 37–78

LINEAR ALGEBRA AND ITS APPLICATIONS

www.elsevier.com/locate/laa

Spectral behavior of matrix sequences and discretized boundary value problems

Stefano Serra Capizzano

Dipartimento di Chimica, Fisica e Matematica, Università dell'Insubria, Sede di Como, Via Valleggio 11, 22100 Como, Italy

Received 24 November 1999; accepted 25 March 2001

Submitted by E. Tyrtyshnikov

Abstract

In this paper we provide theoretical tools for dealing with the spectral properties of general sequences of matrices of increasing dimension. More specifically, we give a unified treatment of notions such as distribution, equal distribution, localization, equal localization, clustering and sub-clustering. As a case study we consider the matrix sequences arising from the finite difference (FD) discretization of elliptic and semielliptic boundary value problems (BVPs). The spectral analysis is then extended to Toeplitz-based preconditioned matrix sequences with special attention to the case where the coefficients of the differential operator are not regular (belong to L^1) and to the case of multidimensional problems. The related clustering properties allow the establishment of some ergodic formulas for the eigenvalues of the preconditioned matrices. © 2001 Published by Elsevier Science Inc.

AMS classification: 15A12; 65N22; 65F10

Keywords: Spectral distribution; Cluster; Preconditioning; Ergodic formula

1. Introduction

In the numerical solution of infinite dimensional problems modeled by partial differential equations, integral equations, infinite systems of linear equations, etc. (fluidodynamics, elasticity, image processing, Markov chains and so on), we deal with sequences of finite dimensional systems $\{A_n\mathbf{u}_n = \mathbf{f}_n\}_n$ where the size of A_n is d_n with $d_n < d_{n+1}$ and is related to the finesse parameter of the discretization.

E-mail address: serra@mail.dm.unipi.it (S. Serra Capizzano).

0024-3795/01/$ - see front matter © 2001 Published by Elsevier Science Inc.
PII: S 0 0 2 4 - 3 7 9 5 (0 1) 0 0 3 3 5 - 4

Frequently the sequence $\{A_n\}_n$ inherits a kind of structure from the infinite dimensional model. This property and the large dimensions involved advise against direct methods and encourage the use of iterative techniques both for memory requirements/CPU time and precision of the solution. The theoretical counterpart of the research into fast iterative solvers is the in-depth study of the spectral properties and asymptotics of sequences of matrices especially in connection with the design and analysis of good preconditioners. Hence many concepts and tools such as distribution, equal distribution, localization, equal localization, clustering and sub-clustering have been more or less explicitly introduced and studied. Here, we first give a systematic and unified treatment of these notions by emphasizing relationships, similarities and differences. Then, as a case study, we consider a class of simple centered finite difference (FD) discretizations of a class of multidimensional boundary value problems of the form

$$\begin{cases} (-)^k \frac{\mathrm{d}^k}{\mathrm{d}x^k}\left(a(x)\frac{\mathrm{d}^k}{\mathrm{d}x^k}u(x)\right) = f(x), \quad x \in \Omega = (0,1), \ \ k \geqslant 1, \\ \text{Dirichlet B.C. on } \partial\Omega, \end{cases} \tag{1}$$

and of the form

$$\begin{cases} \sum_{i=1}^{d}(-)^k \frac{\partial^k}{\partial x_i^k}\left(a(\mathbf{x})\frac{\partial^k}{\partial x_i^k}u(\mathbf{x})\right) = f(\mathbf{x}), \\ \quad \mathbf{x} = (x_1,\ldots,x_d) \in \Omega = (0,1)^d, \ \ k \geqslant 1, \\ \text{Dirichlet B.C. on } \partial\Omega, \end{cases} \tag{2}$$

over hyperrectangular regions and with a nonnegative and *sparsely vanishing* [34] coefficient a.

We suppose that the function a is nonnegative and that the set of the essential zeros of a has zero Lebesgue measure (i.e., a is *sparsely vanishing*: see [34] and Definition 4.7), while generally [8,21], the function a is assumed to be positive on the whole domain. Moreover, again concerning the function a, we do not suppose any regularity except for the boundedness or the Lebesgue integrability.

In a preceding paper [25], we have analyzed the main structural and spectral properties of the FD matrices $A_n(a)$ coming from the discretization of (1) and (2) by centered FD formulae of order 2 with uniform mesh size $h = (n+1)^{-1}$ under the regularity assumption that a is (piecewise) twice continuously differentiable and by emphasizing the case where k is equal to 1. The case where k is greater than 1 and where high precision formulae are used is considered in [27,30]. However, we still assumed the (piecewise) continuity of the functional coefficient a.

In particular, we proposed two preconditioners in [16,25]. The first of these $\Delta_{2k}^{(n)}$ is obtained by the same discretization formula when applied to problem (1) with $a \equiv 1$. The second P_n is constructed as $\theta_k D_{n,a}^{1/2} \Delta_{2k}^{(n)} D_{n,a}^{1/2}$, where $D_{n,a}$ is the $n \times n$ diagonal matrix obtained by the diagonal part of $A_n(a)$ and θ_k is a suitable positive constant.

In [25], by supposing that the functional coefficient a has a continuous second derivative, we obtained an expansion formula for $A_n(a)$ in terms of the two preconditioners. In particular, by using this representation together with the assumption

that $a(x)$ is positive, we proved that the first preconditioner is optimal while the second is even superlinear (according to the notion of optimality of Axelsson and Neytcheva [2] reported in Definition 2.1). Using the same assumption of smoothness but supposing that $a(x)$ has a finite number of isolated zeros, we proved that the first preconditioner is not good because of a linear quantity (that is proportional to the dimension n of $A_n(a)$) of eigenvalues of the preconditioned matrix which accumulate in any ϵ-neighborhood of zero. On the other hand, the second preconditioner is still good at least in the sense that the related preconditioned matrix sequence has a *weak or general cluster* [34] of the eigenvalues around the unity. This "analytical" property does not theoretically guarantee a good behavior of the associated PCG method, but the *general clustering* of the eigenvalues does have a practical counterpart as shown in the numerical experiments discussed in [25] and in Section 7.

This paper can be divided into two parts: in the first part (Sections 3 and 4) we introduce new tools in order to analyze the spectral behavior of matrix-sequences. In the second part (Sections 5 and 6) we focus our attention on special matrix-sequences as $\{P_n\}_n$, $\{\Delta_{2k;d}^{(n)}\}_n$, $\{P_n^{-1}A_n(a)\}_n$, etc. In particular, we relax the regularity hypotheses on the function a by taking into account different cases and especially the case where $a \in L^\infty$ (or $a \in L^1$). In all these cases we prove that the preconditioned matrix sequence $\{P_n^{-1}A_n(a)\}_n$ has a general cluster at 1. We notice that the latter case where $a \in L^\infty$ or $a \in L^1$ poses some technical problems which are overcome by using some standard tools in functional analysis. In fact, the related spectral analysis of the clustering properties is carried out by using the Lusin Theorem [23] regarding the approximation in the measure of measurable functions by continuous functions when the definition space is locally compact. We notice that this approximation allows one to use the previously stated results [25,27,30], which hold in the continuous case, as an intermediate step. Moreover, in the case where a is a *sparsely vanishing function* [13,34], the presence of essential zeros is analyzed as well: the related results indicate that the behavior does not differ substantially with respect to the regular case considered and discussed in [25]. We recall that the analysis performed and reported in [25,27,30] was mainly concerned with the unilevel case. Several results regarding the asymptotical spectral distribution properties and the clustering of the preconditioned matrices are established here for the discrete approximation of multidimensional problems of the form displayed in (2).

The paper is organized as follows: in Section 2 we describe the preconditioning problem as a kind of "constrained approximation" and then we define the preconditioner P_n. In Section 3 we introduce some linear algebra tools for distribution and clustering that have been recently introduced by Tyrtyshnikov [35] in a Toeplitz preconditioning context. Section 4 is devoted to a systematic and unified analysis of the concepts of distribution, equal distribution, localization, equal localization, clustering and sub-clustering for sequences of vectors and matrices. In Section 5, we return to the case study in (1) and (2) by recalling some results regarding second-order BVPs in the presence of smooth coefficients. In Section 6, we present a thorough generalization of the results regarding the distribution and clustering of the spectra

in the case of Dirichlet problems of order $2k$ and "dimension" $d \geqslant 1$, by taking into account, not only the presence of zeros, but also the possibility that a is not regular ($a \in L^\infty$ or $a \in L^1$): while the latter case appears to be only academic, it becomes meaningful and useful when we consider the distributional "weak formulation" of the proposed problems. We report some numerical experiments in Section 7 and we summarize the results obtained in the paper in the concluding section (Section 8).

2. The preconditioning problem

2.1. A constrained approximation problem

When discretizing continuous problems as (1) and (2), we obtain a sequence $\{A_n\mathbf{x}_n = \mathbf{b}_n\}_n$ of linear systems of size d_n with $d_n < d_{n+1}$ (so that $\lim_{n\to\infty} d_n = \infty$). The higher n is, the more accurate the approximate solution $\mathbf{x}_n$ is: consequently, if good precision is required, we have to compute the solution $\mathbf{x}_n$ of a linear system of dimension d_n for a large value of n. The use of iterative methods [38] is recommended due to memory and accuracy requirements. However, in many cases the problem is the number of iterations that can grow as n or the cost per each iteration. Here we give a definition of optimality [2] for iterative methods applied to sequences of linear systems.

Definition 2.1. Given a sequence $\{A_n\mathbf{x}_n = \mathbf{b}_n\}_n$ of linear systems of size d_n, an iterative method is said to be *optimal* if its cost for computing $\mathbf{x}_n$ within a preassigned accuracy ε is $O(M(n))$, where $M(n)$ is the cost of the matrix–vector multiplication with matrix A_n and where the constant hidden in the $O(\cdot)$ term can depend on ε.

In a general iterative method the cost of a single iteration is basically reduced to matrix–vector multiplications where the involved matrix generally has the same pattern and the same structure of the original coefficient matrix A_n. As a consequence, Definition 2.1 implies that:

- the asymptotic number of iterations to reach the solution within the desired accuracy must be upperbounded by a constant and
- the cost of each single step is upperbounded by $O(M(n))$.

The second requirement is satisfied by the most popular iterative solvers so that the first requirement is the critical point. If the sequence of the (spectral) condition numbers of $\{A_n\}_n$ is unbounded, then it is generally difficult to solve the nth system within a number of steps independent of n.

With regard to this feature, one of the most successful iterative solvers is the preconditioned conjugate gradient (PCG) method. When applied without preconditioning, this method requires $O(d_n)$ iterations. On the other hand, the use of a preconditioner M_n can accelerate the convergence by reducing the number of steps.

In order to understand the convergence speed of this class of methods, the main objective is the localization and the distribution of the eigenvalues (singular values in the non-Hermitian case) of the sequence $\{M_n^{-1}A_n\}_n$. Let us start with two definitions.

Definition 2.2 [35]. Consider a sequence of $d_n \times d_n$ complex matrices $\{A_n\}_n$ and a set M in the nonnegative real line. Denote by M_ϵ the ϵ-extension of M, which is the union of all balls of radius ϵ centered at points of M. For any n, let $\gamma_n(\epsilon) \equiv \gamma_n(A_n, M, \epsilon)$ count those singular values of A_n that do not belong to M_ϵ.

- Assume that, for any $\epsilon > 0$,

$$\gamma_n(\epsilon) = o(d_n), \quad n \to \infty.$$

Then M is called a *general or weak cluster*.

- If, for any $\epsilon > 0$, there exists a constant $c(\epsilon)$ so that

$$\gamma_n(\epsilon) \leqslant c(\epsilon),$$

then M is called a *proper or strong cluster*.

- If $M = \{p\}$ is a *cluster*, then we say that $\{A_n\}_n$ is *clustered* at p.
- When the matrices A_n are Hermitian, the set M is allowed to belong to the whole real line and the given definitions apply to the eigenvalues in place of the singular values.

Definition 2.3. Let $\{A_n\}_n$, M, M_ϵ and $\gamma_n(\epsilon)$ be as in the preceding definition.

- The set M is a *sub-cluster* if

$$\lim_{\epsilon\to 0} \frac{1}{d_n} \liminf_{n\to\infty} \gamma_n(\epsilon) = c < 1.$$

- If $M = \{p\}$ is a *sub-cluster*, then we say that p is a *sub-cluster point* for $\{A_n\}_n$.

Remark 2.1. Let c be as in the first item of Definition 2.3 and let C be defined as

$$C = \lim_{\epsilon\to 0} \frac{1}{d_n} \limsup_{n\to\infty} \gamma_n(\epsilon).$$

Correspondingly we can give a more restrictive notion of *sub-cluster* identified by the relation $C < 1$. In essence, the latter definition is the one of Tyrtyshnikov and Zamarashkin in [37] which we will refer to as the concept of *TZ sub-cluster*. Table 1 helps to understand the relations among the notions of *weak cluster*, *sub-cluster* and *TZ sub-cluster*. We notice that the *weak clustering* is a special instance of the *sub-clustering* while the latter notion is a special case of the notion of *TZ sub-cluster*. The philosophical difference between *sub-clustering* and *TZ sub-clustering* can be summarized as follows: in the TZ definition we find a real sub-cluster for all the subsequences n_k extracted from n while in our definition, there exists at least a subsequence n_k such that M is a real sub-cluster for $\{A_{n_k}\}_k$. Indeed in the case 'C' or in the case 'F' there exist subsequences n_k for which M is a real sub-cluster of type 'A' or 'D' for $\{A_{n_k}\}_k$, respectively, but there exist subsequences $\hat{n}_k$ and $\epsilon > 0$ for

Table 1
Clustering and sub-clustering

		Weak cluster	TZ sub-cluster	Sub-cluster
A	$c = 0,\ C = 0$	Yes	Yes	Yes
B	$c = 0,\ C \in (0,1)$	No	Yes	Yes
C	$c = 0,\ C = 1$	No	No	Yes
D	$c = C \in (0,1)$	No	Yes	Yes
E	$0 < c < C < 1$	No	Yes	Yes
F	$c < 1,\ C = 1$	No	No	Yes
G	$c = 1,\ C = 1$	No	No	No

which M_ϵ contains at most $o(d_{\hat{n}_k})$ singular values of $\{A_{\hat{n}_k}\}_k$. Consequently M is not a sub-cluster for the extracted subsequence $\{A_{\hat{n}_k}\}_k$ and with regard to this subsequence we are in the case 'G'. Finally, we refer to the situation in 'D' as the canonical case.

Regarding the terminology of the preceding definitions, when the eigenvalues/singular values of $\{M_n^{-1}A_n - I\}_n$ are *properly clustered* at zero or when the sequence of the spectral condition numbers $\kappa(M_n^{-1}A_n)$ of $\{M_n^{-1}A_n\}_n$ is upperbounded by a constant independent of n, we know [1] that a constant number of iterations are required by the PCG method in order to solve a linear system with coefficient matrix A_n within a fixed accuracy. In particular, if $\{M_n^{-1}A_n - I\}_n$ is *properly clustered* and $\{A_n^{-1}M_n\}_n$ is spectrally bounded, then the related PCG method is optimal and, after a suitable constant number of iterations, the convergence is of a superlinear type (see, [1,9] for more details).

Therefore, we need to find a suitable preconditioner M_n such that:

1.a. $\kappa(M_n^{-1}A_n)$ is upperbounded by a constant independent of n (that is $\{M_n\}_n$ is "close" to $\{A_n\}_n$ in spectral norm) or

1.b. $\{M_n^{-1}A_n - I\}_n$ is properly clustered at 0 (that is $\{M_n\}_n$ is "close" to $\{A_n\}_n$ in the clustering sense);

2. a linear system involving M_n has a cost of $O(M(n))$.

Clearly, these two issues are often conflicting considering that when a matrix is too close to A_n it also requires the same computational effort to invert.

In general, when dealing with the PCG, given the class α of matrices arising from problems like (1) or (2), we proceed as follows:

A. Choose a suitable class β of matrices "close" enough to α whose elements are easy to invert.

B. Devise a suitable projection operator $\mathscr{P}_n : \alpha \to \beta$ to obtain a certain approximation $M_n \in \beta$ for any given $A_n \in \alpha$.

One of the possible ways to do this is to look for preconditioners within matrix algebras such as the circulant class [12] and the τ class [5] (point A) and to use the optimal approximation in Frobenius norm [8,10,29] (point B). However, as proved in [13] we do not generally meet points 1.a and 1.b due to asymptotical ill conditioning of the matrices discretizing problems of the form (1) or (2). Therefore we go on to use

a different strategy which is based on the information contained in the "continuous problem".

2.2. The choice of the preconditioner

Let us consider the discretization of problem (1) on a uniform grid belonging to $\bar{\Omega} = [0, 1]$ with stepsize $h = (n + 1)^{-1}$, using centered finite differences of minimal precision order 2. After a scaling by h^{2k}, this discrete approximation leads to a $2k + 1$ band $n \times n$ linear system

$$A_n(a)\mathbf{y} = \mathbf{b}, \tag{3}$$

that belongs to the Toeplitz class if $a(x)$ is a constant function.

The first component of our preconditioner $P_n \equiv P_n(a)$ is the Toeplitz matrix $\Delta_{2k}^{(n)}$ obtained from the discretization of Eq. (1), where $a(x) \equiv 1$ and after the same scaling by h^{2k}:

$$\Delta_{2k}^{(n)} = \mathrm{Toep}_n[0, \ldots, 0, \tau_k, \ldots, \tau_0, \ldots, \tau_k, 0, \ldots, 0],$$

with

$$\tau_j = (-1)^j \binom{2k}{k-j}.$$

The second component is the diagonal matrix $D_{n,a}$ obtained by the main diagonal of $A_n(a)$ (see also [16]). Therefore we set

$$P_n = \binom{2k}{k}^{-1} D_{n,a}^{1/2} \Delta_{2k}^{(n)} D_{n,a}^{1/2}, \tag{4}$$

where

$$(A_n(a))_{j,j} = \sum_{i=0}^{k} b_i^{(k,0)} a(x_{j+i-k/2}),$$

$$(A_n(a))_{j,j+s} = \sum_{i=\max\{0,s\}}^{\min\{k,k+s\}} (-1)^s b_i^{(k,s)} a(x_{j+i-k/2}),$$

$$b_i^{(k,s)} > 0, \quad s \in \{-k, \ldots, k\}.$$

Moreover, by imposing $a \equiv 1$ we infer that

$$\sum_i b_i^{(k,s)} = \binom{2k}{k-|s|} = \left(\Delta_{2k}^{(n)}\right)_{j,j+s}. \tag{5}$$

With this choice, we note that the preconditioner is symmetric and is positive definite if the diagonal elements of $A_n(a)$ are positive. In fact the generic value $(A_n(a))_{j,j}$ is a sum with positive coefficients of $k + 1$ equispaced evaluations of the function $a(x)$. Therefore, if $a(x)$ is continuous and has, at most, only isolated zeros, it is evident

that $(A_n(a))_{j,j}$ is positive for a sufficiently fine mesh spacing. Of course, when a is not continuous but only L^1 it does not make any sense to take the evaluations of the function a in a discrete grid of points. In this case the symbol $a(x_j)$ denotes

$$(n+1)\int_{I_j} a(t)\mathrm{d}t, \quad I_j = [x_j, x_{j+1}], \quad x_{j+1} - x_j = h, \quad x_j = hj.$$

Hence, if the function a is sparsely vanishing [34] or equivalently [13] if the Lebesgue measure $m\{x \in [0,1] : a(x) = 0\}$ of the set of zeros of a is zero, then $A_n(a)$ and the preconditioner P_n are symmetric and positive definite since each value $a(x_j)$ is strictly positive.

In the case of problem (2), we discretize them on a uniform grid belonging to $\bar{\Omega} = [0,1]^d$ using centered finite differences of minimal precision order 2 with regard to each direction x_j. The stepsize with regard to the direction x_j, $j = 1, \ldots, d$, is $h_j = (n_j + 1)^{-1}$. Since the discretization of each term

$$\frac{\partial^k}{\partial x_j^k}\left(a(\mathbf{x})\frac{\partial^k}{\partial x_j^k}u(\mathbf{x})\right)$$

is represented by a matrix bounded in spectral norm divided by h_j^{2k}, it follows that the relation

$$h_{j_1}^{2k} = \mathrm{o}\left(h_{j_2}^{2k}\right)$$

for some $j_1 \neq j_2$ would imply that the "discrete" contribution of the operator

$$\frac{\partial^k}{\partial x_{j_2}^k}\left(a(\mathbf{x})\frac{\partial^k}{\partial x_{j_2}^k}u(\mathbf{x})\right)$$

is negligible in spectral norm with respect to the one of

$$\frac{\partial^k}{\partial x_{j_1}^k}\left(a(\mathbf{x})\frac{\partial^k}{\partial x_{j_1}^k}u(\mathbf{x})\right).$$

Therefore, except for some exceptional cases where it is required to discretize with different precisions in different directions, it is natural to think that h_{j_1} and h_{j_2} have the same asymptotical behavior. Consequently, we make the assumption that the stepsizes are equal up to suitable multiplicative constants, i.e., $\exists a_1, \ldots, a_d \in \mathcal{N}_+$ such that $n_j + 1 = va_j$, $v \in \mathcal{N}_+$.

Setting the multiindex $n = (n_1, n_2, \ldots, n_d)$ with $N(n) = n_1 n_2 \cdots n_d$ and after a scaling by v^{-2k}, the former discrete approximation leads to a multilevel $N(n) \times N(n)$ linear system

$$A_n(a)\mathbf{y} = \mathbf{b}, \tag{6}$$

having bandwidth $2k+1$ at each level j. The matrix $A_n(a)$ belongs to the multilevel Toeplitz class if $a(\mathbf{x})$ is a constant function.

The first component of our preconditioner $P_n \equiv P_n(a)$ is the Toeplitz matrix $\Delta_{2k;d}^{(n)}$ obtained from the discretization of Eq. (2), where $a(\mathbf{x}) \equiv 1$ and following the same scaling by v^{-2k}. Then the preconditioner $\Delta_{2k;d}^{(n)}$ is defined as

$$\Delta_{2k;d}^{(n)} = \sum_{j=1}^{d} X_{j,k}^{(n)},$$

where

$$X_{j,k}^{(n)} = a_j^{2k} \cdot I_{n_1} \otimes I_{n_2} \otimes \cdots I_{n_{j-1}} \otimes \Delta_{2k}^{(n_j)} \otimes I_{n_{j+1}} \otimes \cdots I_{n_d}$$

and $\otimes$ is the tensor product [3,18] ($A \otimes B = (a_{i,j}B)$).

The second component is the diagonal matrix $D_{n,a}$ extracted from the main diagonal of $A_n(a)$ so that

$$P_n = \theta_k{}^{-1} D_{n,a}^{1/2} \Delta_{2k;d}^{(n)} D_{n,a}^{1/2},$$

where

$$\theta_k = \left(\Delta_{2k;d}^{(n)}\right)_{i,i} = \left(\sum_{j=1}^{d} a_j^{2k}\right) \binom{2k}{k}$$

is the diagonal entry of the multilevel Toeplitz matrix $\Delta_{2k;d}^{(n)}$.

We note that the preconditioner P_n is symmetric and is positive definite if the diagonal elements of $A_n(a)$ are positive. In actuality, putting $j = (j_1, j_2, \ldots, j_d)$, the diagonal entry $(A_n(a))_{j,j}$ is a sum with positive coefficients of a constant number of evaluations of the function $a(\mathbf{x})$. Therefore, if $a(\mathbf{x})$ is continuous with, at most, only isolated zeros, then $(A_n(a))_{j,j}$ is positive when the mesh is fine enough. If $a \in L^1$ is not continuous, then it is senseless to take the evaluations of the function a in a discrete grid of points. Therefore, setting $\mathbf{x}_j = (x_{j_1}, x_{j_2}, \ldots, x_{j_d})$ and $\mathbf{e} = (1, 1, \ldots, 1)$, the symbol $a(\mathbf{x}_j)$ will denote

$$N(n+e) \int_{I_j} a(\mathbf{t}) \mathrm{d}\mathbf{t}, \quad I_j = \prod_{i=1}^{d} [x_{j_i}, x_{j_i+1}], \quad x_{j_i+1} - x_{j_i} = h_i, \quad x_{j_i} = h_i j_i.$$

So, if the function a is sparsely vanishing [34] or equivalently [13] if the multidimensional Lebesgue measure $m\{\mathbf{x} \in [0,1]^d : a(\mathbf{x}) = 0\}$ of the set of zeros of a is zero, then $A_n(a)$ and the preconditioner P_n are symmetric and positive definite since each $a(\mathbf{x}_j)$ is strictly positive.

3. Some linear algebra premises

We now introduce some results to be used as a tool in order to analyze the distribution of the eigenvalues of the preconditioned and nonpreconditioned matrices.

3.1. The Szegö–Tyrtyshnikov theory

Suppose that $f \in L^1(I^d, \mathscr{C})$ with $I = (-\pi, \pi)$. We define the (d-level) Toeplitz matrices $\{T_n(f)\}_n$, generated [19] by a Lebesgue-integrable function f as the matrices $T_n(f)$ whose entries along the diagonals are constant and are given by the Fourier coefficients $\{\tau_k\}_k$ ordered in a suitable way. In particular, setting

$$\begin{aligned} &n = (n_1, n_2, \ldots, n_d) \in \mathscr{N}_+{}^d, && N(n) = n_1 n_2 \cdots n_d, \\ &k = (k_1, k_2, \ldots, k_d), && s = (s_1, s_2, \ldots, s_d), \\ &t = (t_1, t_2, \ldots, t_d), && k_j, s_j, t_j \in \{-n_{j+1}, \ldots, n_{j-1}\} \end{aligned}$$

and

$$x = (x_1, x_2, \ldots, x_d)$$

we have

$$\left[T_n(f)\right]_{s,t} = \tau_{s-t}, \quad \tau_k = \frac{1}{[2\pi]^d} \int_{I^d} f(x) e^{-\mathbf{i}(k\cdot x)} \mathrm{d}x, \ \ \mathbf{i}^2 = -1. \tag{7}$$

To have an idea of the d-level structure, we must choose an ordering among the indices $\{k_j\}$. The matrix $T_n(f)$ has external dimension $n_1 \times n_1$, with $(d-1)$-level Toeplitz blocks of dimension $(n_2 \cdots n_d) \times (n_2 \cdots n_d)$. The description is naturally recursive so that when we arrive at expressing the first level, we then find the elements $\{\tau_k\}_k$ given in Eq. (7).

For instance, for $d = 2$, the expression $[T_n(f)]_{(s_1,s_2),(t_1,t_2)}$ indicates the entry of position (s_2, t_2) in the block (s_1, t_1) which is equal to $\tau_{s_1-t_1, s_2-t_2}$.

The following theorem gives a strong characterization of the spectra of multilevel Toeplitz matrices.

Theorem 3.1 [19,35,36]. *Let $f \in L^1(I^d, \mathscr{C})$ and let $\{\sigma_i^{(n)}\}$ be the singular values of $T_n(f)$. Then, for any continuous function F with bounded support the following asymptotic formula (the Szegö relation) holds true*:

$$\lim_{n\to\infty} \frac{1}{N(n)} \sum_{i=1}^{N(n)} F(\sigma_i^{(n)}) = \frac{1}{[2\pi]^d} \int_{I^d} F(|f(x)|) \mathrm{d}x. \tag{8}$$

Notice that when f is real-valued and nonnegative the singular values of $T_n(f)$ are its eigenvalues since $T_n(f)$ is Hermitian and nonnegative definite.

Theorem 3.2. *Let $f \in L^1(I^d, \mathscr{C})$ and $\{T_n(f)\}_n$ be the related Toeplitz sequence. Then p is a (singular value)* sub-cluster point *for $\{T_n(f)\}_n$ iff $m\{\mathbf{x} \in I^d : |f(\mathbf{x})| = p\} > 0$. If f is real-valued, then p is a (eigenvalue)* sub-cluster point *for $\{T_n(f)\}_n$ iff $m\{\mathbf{x} \in I^d : f(\mathbf{x}) = p\} > 0$. Here $m\{\cdot\}$ is the Lebesgue measure on $\mathbb{R}^d$.*

Proof. The proof can be handled by using limit relation (8) and straightforward measure theory arguments. □

Remark 3.1. Let $f \in L^1(I^d, \mathscr{C})$ and $\{T_n(f)\}_n$ be the related Toeplitz sequence. Then the set of the *sub-cluster points* is at most countable. Moreover, the *sub-cluster points* of Toeplitz sequences are all of canonical type in the sense of Remark 2.1 since the quantities c and C coincide.

The following further result holds.

Theorem 3.3. *Let $T_n(f)$ be as in Theorem* 3.1 *and f be a real-valued function. Let $\lambda(X)$ be the generic eigenvalue of the matrix X and let m_f and M_f be the essential infimum and supremum of f, respectively, i.e., the* inf f *and* sup f *up to within zero measure sets* [23]. *Then for any $n \in \mathscr{N}_+{}^d$ the following cases occur* [19]:
- $m_f < \lambda(T_n(f)) < M_f$ *if* $m_f < M_f$ *or*
- $\lambda(T_n(f)) = M$ *if* $m_f = M_f = M$.

In addition, if $\{\lambda_i^{(n)}\}$ is the complete set of the eigenvalues of $T_n(f)$ in nondecreasing order, then $\forall\{k(n)\}_n$ so that $k(n) = o(N(n))$, it follows that [35]

$$\lim_{n\to\infty} \lambda_{k(n)}^{(n)} = m_f \quad \textit{and} \quad \lim_{n\to\infty} \lambda_{N(n)-k(n)}^{(n)} = M_f.$$

Finally, if $f - m_f \sim \|\mathbf{x} - \mathbf{x}_0\|_2^{\gamma}$, then [26, 7]

$$\lambda_1^{(n)} - m_f \sim \sum_{j=1}^{d} n_j^{-\gamma}.$$

As an example, let us consider the band Toeplitz matrices $\{\Delta_{2k}^{(n)}\}_n$ related to the discretization of Eq. (1) with $a(x) \equiv 1$. It is evident that $\Delta_{2k}^{(n)}$ is the $n \times n$ one-level Toeplitz matrix ($d = 1$) generated by the polynomial $f(x_1) = (2 - 2\cos(x_1))^k$. Since $m_f = 0$ and $f(x_1) \sim x_1^{2k}$, according to Theorem 3.3 the matrix $\Delta_{2k}^{(n)}$ is positive definite, its minimal eigenvalue is asymptotic to n^{-2k} and its spectral condition number $\kappa(\Delta_{2k}^{(n)})$ grows as n^{2k}.

Let us consider the two-level Toeplitz matrices $\{\Delta_{4;2}^{(n)}\}_n$, $n = (n_1, n_2)$, related to the discretization of the bi-Laplacian (see Eq. (2) with $k = 2$) with $n_j + 1 = va_j$, $j = 1, 2$. Then $\Delta_{4;2}^{(n)} = T_n(f)$, where $f(x_1, x_2) = a_1^2(2 - 2\cos(x_1))^2 + a_2^2(2 - 2\cos(x_2))^2$.

In view of Theorem 3.3, we deduce that $\Delta_{4;2}^{(n)}$ is positive definite, its minimal eigenvalue is asymptotic to $n_1^{-4} + n_2^{-4}$ and its spectral condition number grows as $n_1^4 + n_2^4$.

4. The Weyl–Tyrtyshnikov equal distribution

Let us start with the basic definition of distribution.

Definition 4.1 [19, 35]. Two real sequences $\{a_i^{(n)}\}_{i\leqslant d_n}$, $\{b_i^{(n)}\}_{i\leqslant d_n}$ $(d_n < d_{n+1})$ are *equally distributed* (ED) if and only if, for any real-valued continuous function F with bounded support, the following relation holds:

$$\lim_{n\to\infty}\frac{1}{d_n}\sum_{i=1}^{d_n}F\left(a_i^{(n)}\right)-F\left(b_i^{(n)}\right)=0. \tag{9}$$

When the previous limit goes to zero as $O(d_n^{-1})$ and F is Lipschitz continuous, we say that there is *strong equal distribution* (SED). The same definition applies to the case of sequences of matrices $\{A_n\}_n$ and $\{B_n\}_n$ of dimension $d_n\times d_n$: in this case $\{a_i^{(n)}\}_{i\leqslant d_n}$ and $\{b_i^{(n)}\}_{i\leqslant d_n}$ are the sets of their singular values (or the eigenvalues if the involved matrices are Hermitian).

We now introduce two notions of *equal localization* that will be useful in the following.

Definition 4.2. Two real sequences $\{a_i^{(n)}\}_{i\leqslant d_n}$, $\{b_i^{(n)}\}_{i\leqslant d_n}$ $(d_n < d_{n+1})$ are *equally localized* (EL) if and only if, for any nontrivial interval $[\alpha,\beta]$ $(\alpha<\beta)$, the following relation holds:

$$\lim_{n\to\infty}\frac{1}{d_n}\left(\#\left\{i : a_i^{(n)}\in[\alpha,\beta]\right\}-\#\left\{i : b_i^{(n)}\in[\alpha,\beta]\right\}\right)=0. \tag{10}$$

When the previous limit goes to zero as $O(d_n^{-1})$, we say that there is strong equal localization (SEL). The same definition applies to the case of sequences of matrices $\{A_n\}_n$ and $\{B_n\}_n$ of dimension $d_n\times d_n$: in this case $\{a_i^{(n)}\}_{i\leqslant d_n}$ and $\{b_i^{(n)}\}_{i\leqslant d_n}$ are the sets of their singular values (or the eigenvalues if the involved matrices are Hermitian).

Definition 4.3. Two ordered real sequences $\{a_i^{(n)}\}_{i\leqslant d_n}$, $\{b_i^{(n)}\}_{i\leqslant d_n}$ $(d_n < d_{n+1})$ are ϵ *equally localized* (ϵ-EL) if and only if, for any $\epsilon>0$, the following relation holds.

$$\lim_{n\to\infty}\frac{1}{d_n}\#\left\{i : \left|a_i^{(n)}-b_i^{(n)}\right|>\epsilon\right\}=0. \tag{11}$$

When the previous limit goes to zero as $O(d_n^{-1})$, we say that there is ϵ *strong equal localization* (ϵ-SEL). The same definition applies to the case of sequences of matrices $\{A_n\}_n$ and $\{B_n\}_n$ of dimension $d_n\times d_n$: in this case $\{a_i^{(n)}\}_{i\leqslant d_n}$ and $\{b_i^{(n)}\}_{i\leqslant d_n}$ are the ordered sets of their singular values (or the eigenvalues if the involved matrices are Hermitian).

Definition 4.4. We say that a sequence $\{c_i^{(n)}\}_{i\leqslant d_n}$ is essentially bounded if there exists an interval $M=[\alpha,\beta]$ so that M is a general cluster for it. If M is a proper cluster, then we say that $\{c_i^{(n)}\}_{i\leqslant d_n}$ is properly bounded.

Definition 4.5. Given a sequence $\{c_i^{(n)}\}_{i\leqslant d_n}$, we say that $p\in\mathbb{R}$ is a *sub-cluster point* for $\{c_i^{(n)}\}_{i\leqslant d_n}$ iff

$$\lim_{\epsilon\to 0}\frac{1}{d_n}\limsup_{n\to\infty}\#\left\{i : c_i^{(n)}\in(p-\epsilon,\, p+\epsilon)\right\}=c>0.$$

A sequence $\{c_i^{(n)}\}_{i\leqslant d_n}$ without *sub-cluster points* is called *regular*.

Remark 4.1. We notice that if $\{c_i^{(n)}\}_{i\leqslant d_n}$ is, for any n, the complete set of the singular values (eigenvalues) of a (Hermitian) $d_n\times d_n$ matrix, then definition of *sub-cluster point* in Definition 4.5 reduces to the one in Definition 2.3 (see also [37]).

Remark 4.2. If $\{c_i^{(n)}\}_{i\leqslant d_n}$ is, for any n, the complete set of the singular values (eigenvalues) of a (Hermitian) $d_n\times d_n$ matrix A_n and if the singular values (eigenvalues) of $\{A_n\}_n$ enjoy a formula as (8) for some Lebesgue measurable function f, then f is *sparsely vanishing* according to Definition 4.7 iff 0 is not a *sub-cluster point*.

However all these concepts have deep relationships without being equivalent. Therefore in the following theorem we analyze the connections and the differences among them in detail.

Theorem 4.1. *Let* $\{a_i^{(n)}\}_{i\leqslant d_n}$ *and* $\{b_i^{(n)}\}_{i\leqslant d_n}$ $(d_n<d_{n+1})$ *be two ordered real sequences. The following facts hold true*:

1. *SED implies ED, SEL implies EL and* ϵ*-SEL implies* ϵ*-EL. These implications cannot be reversed.*
2. *EL implies ED.*
3. *SEL does not imply SED.*
4. *SED does not imply EL.*
5. ϵ*-EL implies ED.*
6. ϵ*-SEL does not imply SED.*
7. *SED does not imply* ϵ*-EL.*
8. ϵ*-SEL does not imply EL.*
9. *SEL does not imply* ϵ*-EL.*

Proof.

1. The implications SED $\Rightarrow$ ED and ϵ-SEL $\Rightarrow$ ϵ -EL are straightforward consequences of Definitions 4.1 and 4.3. SEL $\Rightarrow$ EL is a consequence of Definition 4.2 and of density of the Lipschitz continuous functions with bounded support into the class of continuous functions with bounded support. The sequences $\{a_i^{(n)}=1+1/\sqrt{n}\}_{i\leqslant n}$ and $\{b_i^{(n)}=1\}_{i\leqslant n}$ are ED but not SED. The sequences $\{a_i^{(n)}=i/n\}_{i\leqslant n}$ and $\{b_i^{(n)}=i/n-1/\sqrt{n}\}_{i\leqslant n}$ are EL but not SEL. The sequences $\{a_i^{(n)}=1\}_{i\leqslant n}$ and $\{b_i^{(n)}\}_{i\leqslant n}$, with $b_i^{(n)}=2$ if $i=2^k$ for some integer k and $b_i^{(n)}=1$ otherwise, are ϵ-EL but not ϵ-SEL.

2. In [35] Tyrtyshnikov observed that EL is a rewriting of the ED property where the continuous test functions are replaced by the staircase functions (i.e. linear combinations of the characteristic functions of intervals). The density of the second class into the first proves the desired result.
3. Let $\{a_i^{(n)} = 1 + 1/\sqrt{n}\}_{i\leqslant n}$ and $\{b_i^{(n)} = 1 + 2/\sqrt{n}\}_{i\leqslant n}$. Now for any nontrivial interval $[\alpha, \beta]$ $(\alpha < \beta)$, there exist integers $\bar{n}(\alpha, \beta)$ and $\tilde{n}(\alpha, \beta)$ so that

$$\#\left\{i : a_i^{(n)} \in [\alpha, \beta]\right\} = s(n) \quad \text{for } n \geqslant \bar{n}(\alpha, \beta)$$

and

$$\#\left\{i : b_i^{(n)} \in [\alpha, \beta]\right\} = s(n) \quad \text{for } n \geqslant \tilde{n}(\alpha, \beta),$$

where $s(n) = n$ if $1 \in [\alpha, \beta)$ and $s(n) = 0$ otherwise. This shows that the two sequences are SEL. Finally taking a continuous function F with bounded support which is identically equal to 1 over $[0, 3]$: we obtain

$$\frac{1}{n}\sum_{i=1}^{n} F\left(a_i^{(n)}\right) - F\left(b_i^{(n)}\right) = -1/\sqrt{n} \neq \mathrm{O}(n^{-1})$$

so that the two sequences are not SED.
4. For $\{a_i^{(n)} = 1 + 1/n^4\}_{i\leqslant n}$ and $\{b_i^{(n)} = 1 - 1/n^4\}_{i\leqslant n}$ it is quickly verifiable that they are SED but not EL since

$$\#\left\{i : a_i^{(n)} \in [0, 1]\right\} = 0 \quad \text{and} \quad \#\left\{i : b_i^{(n)} \in [0, 1]\right\} = n.$$

5. Given two ϵ-EL sequences $\{a_i^{(n)}\}_{i\leqslant d_n}$ and $\{b_i^{(n)}\}_{i\leqslant d_n}$, for any positive ϵ, consider $\gamma_n(\epsilon) = \#\{i : |a_i^{(n)} - b_i^{(n)}| > \epsilon\} = \mathrm{o}(d_n)$. Therefore, for any continuous F with bounded support and modulus of continuity ω_F we have

$$\begin{aligned}\left|\frac{1}{d_n}\sum_{i=1}^{d_n} F\left(a_i^{(n)}\right) - F\left(b_i^{(n)}\right)\right| &\leqslant \frac{1}{d_n}\sum_{i=1}^{d_n}\left|F\left(a_i^{(n)}\right) - F\left(b_i^{(n)}\right)\right| \\ &\leqslant \frac{1}{d_n}(2\|F\|_\infty \gamma_n(\epsilon) + \omega_F(\epsilon) d_n).\end{aligned}$$

Due to the arbitrariness of ϵ, the ED property follows.
6. It is enough to take the same sequences used for showing that SEL does not imply SED (part 3) to prove that ϵ-SEL does not imply SED.
7. For $\{a_i^{(n)} = n + i\}_{i\leqslant n}$ and $\{b_i^{(n)} = n + i + 1/2\}_{i\leqslant n}$ it is quickly verifiable that they are SED but not ϵ-EL.
8. For $\{a_i^{(n)} = 1 + 1/n^4\}_{i\leqslant n}$ and $\{b_i^{(n)} = 1 - 1/n^4\}_{i\leqslant n}$, it is trivial to verify that they are ϵ-SEL but not EL (take $[\alpha, \beta] = [1, 2]$).
9. For $\{a_i^{(n)} = n + i\}_{i\leqslant n}$ and $\{b_i^{(n)} = n + i + 1/2\}_{i\leqslant n}$ it is quickly verifiable that they are SEL but not ϵ-EL. □

We describe some more sophisticated relationships in the subsequent result.

Theorem 4.2. *Let* $\{a_i^{(n)}\}_{i\leqslant d_n}$ *and* $\{b_i^{(n)}\}_{i\leqslant d_n}$ $(d_n < d_{n+1})$ *be two ordered real sequences. The following facts hold true*:

1. *EL and essential boundedness imply* ϵ*-EL.*
2. *SEL and proper boundedness imply* ϵ*-SEL.*
3. *ED, essential boundedness and regularity imply* ϵ*-EL.*
4. *ED, essential boundedness and regularity imply EL.*
5. *Assumed essential boundedness and regularity imply EL iff* ϵ*-EL.*

Proof.

1. For the claim that "EL and essential boundedness imply ϵ-EL", it is enough to consider an equispaced partition of the bounded interval $M = [\alpha, \beta]$ of $\mathbb{R}$ to which $\{a_i^{(n)}\}_{i\leqslant d_n}$ and $\{b_i^{(n)}\}_{i\leqslant d_n}$ essentially belong. For any $\epsilon^* > 0$, we take the intervals $M_j = [x_j, x_{j+1}]$, where $x_j = \alpha + j(\beta-\alpha)\epsilon^*$, $j = 0, \ldots, \lceil(\epsilon^*)^{-1}\rceil$. Next we apply Eq. (10) to all of the intervals M_j. We find that, except for $\lceil(\epsilon^*)^{-1}\rceil o(d_n)$ indices, it holds that $|a_i^{(n)} - b_i^{(n)}| \leqslant \epsilon^*(\beta-\alpha)$ because $\epsilon^*(\beta-\alpha)$ is the diameter of each set M_j. Finally, the claim follows calling $\epsilon = \epsilon^*(\beta-\alpha)$.
2. It is just an adaptation of the proof of part 1.
3. Let $M = [\alpha, \beta] \subset \mathbb{R}$ be a general cluster for $\{a_i^{(n)}\}_{i\leqslant d_n}$ and $\{b_i^{(n)}\}_{i\leqslant d_n}$. For any $\epsilon^* > 0$, we take the intervals $M_j = [x_j, x_{j+1}]$, where $x_j = \alpha + j(\beta-\alpha)\epsilon^*$, $j = 0, \ldots, \lceil(\epsilon^*)^{-1}\rceil$. We now take $F = F_j$ being globally continuous, 1 over M_j, 0 over the complementary set of $M_{j-1} \cup M_j \cup M_{j+1}$, and linear over M_{j-1} and M_{j+1}. We apply Eq. (9) with $F = F_j$ to obtain

$$\lim_{n\to\infty} \frac{1}{d_n}\Big[T_1(j,\epsilon,n) + T_2(j,\epsilon,n) + \left(\#\left\{i : a_i^{(n)} \in M_j\right\} - \#\left\{i : b_i^{(n)} \in M_j\right\}\right)\Big] = 0,$$

where

$$T_1(j,\epsilon,n) = \sum_{i:\, a_i^{(n)} \in M_{j-1}\cup M_{j+1}} F_j\left(a_i^{(n)}\right),$$

$$T_2(j,\epsilon,n) = \sum_{i:\, b_i^{(n)} \in M_{j-1}\cup M_{j+1}} F_j\left(b_i^{(n)}\right).$$

Since $0 \leqslant F_j \leqslant 1$, it follows that $0 \leqslant T_1(j,\epsilon,n) \leqslant \#\{i : a_i^{(n)} \in M_{j-1} \cup M_{j+1}\}$ and $0 \leqslant T_2(j,\epsilon,n) \leqslant \#\{i : b_i^{(n)} \in M_{j-1} \cup M_{j+1}\}$. Both the sequences $\{a_i^{(n)}\}_{i\leqslant d_n}$ and $\{b_i^{(n)}\}_{i\leqslant d_n}$ are *regular* (no *sub-cluster points*) and consequently

$$\lim_{n\to\infty} \frac{1}{d_n} T_1(j,\epsilon,n) = \lim_{n\to\infty} \frac{1}{d_n} T_2(j,\epsilon,n) = 0$$

so that

$$\#\left\{i : a_i^{(n)} \in M_j\right\} = \#\left\{i : b_i^{(n)} \in M_j\right\} + o(d_n).$$

Recalling that the diameter of M_j is $(\beta-\alpha)\epsilon^*$ and that the union of finitely many M_j covers M which is a cluster for both the sequences, by setting $\epsilon = (\beta-\alpha)\epsilon^*$ we deduce that except for $o(d_n)$ indices, the values $a_i^{(n)}$ and $b_i^{(n)}$ belong to the same M_j for some j and therefore their distance is bounded by ϵ. The definition of the ϵ-EL property is easily recognized.

4. To prove the EL property, we take an interval $M = [\alpha, \beta]$ and we take $F = F_{\alpha,\beta}$ being globally continuous, 1 over M, 0 over the complementary set of $M_\epsilon = [\alpha - \epsilon, \beta + \epsilon]$ and linear over $[\alpha-\epsilon, \alpha]$ and $[\beta, \beta+\epsilon]$. We apply Eq. (9) with $F = F_{\alpha,\beta}$ and we obtain

$$\lim_{n\to\infty} \frac{1}{d_n}\left[T_1(j,\epsilon,n) + T_2(j,\epsilon,n) + \left(\#\left\{i : a_i^{(n)} \in M\right\} - \#\left\{i : b_i^{(n)} \in M\right\}\right)\right] = 0,$$

where

$$T_1(j,\epsilon,n) = \sum_{i:\, a_i^{(n)} \in M_\epsilon \setminus M} F_{\alpha,\beta}\left(a_i^{(n)}\right), \quad T_2(j,\epsilon,n) = \sum_{i:\, b_i^{(n)} \in M_\epsilon \setminus M} F_{\alpha,\beta}\left(b_i^{(n)}\right).$$

Just as in the proof of the preceding part, we find that

$$\lim_{n\to\infty} \frac{1}{d_n} T_1(j,\epsilon,n) = \lim_{n\to\infty} \frac{1}{d_n} T_2(j,\epsilon,n) = 0$$

so that

$$\#\left\{i : a_i^{(n)} \in M\right\} = \#\left\{i : b_i^{(n)} \in M\right\} + o(d_n)$$

and the proof is concluded.

5. Assuming the EL property, we deduce the ED property from part 2 of Theorem 4.1. Therefore ED, essential boundedness and regularity hold simultaneously so that, by part 3, the ϵ-EL property stands. The other case is symmetric. □

Remark 4.3. In the proof of the first three parts of Theorem 4.2 we have taken advantage of the fact that a bounded interval can be divided into a finite number of subintervals of radius as small as we desire. In essence, this is the notion of compactness in a metric space. Therefore the same proof applies unchanged if the sequences $\{a_i^{(n)}\}_{i\leqslant d_n}$ and $\{b_i^{(n)}\}_{i\leqslant d_n}$ are valued in a metric space $\mathscr{T}$ and if they are *essentially compact*, i.e., they are contained in a compact set $K \subset \mathscr{T}$ except at most $o(d_n)$ elements (compare this concept with Definition 4.4). A nontrivial case occurs when $\mathscr{T} = \mathbb{R}^s$ for some positive s. In this case we encounter the notion of grid-sequences over domains of $\mathbb{R}^s$ that are useful in constructive approximation and in the numerical treatment of differential equations.

In the following subsections we furnish some tools to evaluate the strength of the *equal distribution* and *equal localization* that are based upon estimates of the singular values and involve the Shatten p-norms [3].

Definition 4.6. We denote by M_s the linear space $\mathscr{C}^{s\times s}$ of all square matrices of order s with complex entries. If $A \in M_s$, then the symbol $\sigma_j(A)$ denotes the jth singular value of A and if A is Hermitian, then the symbol $\lambda_j(A)$ denotes the jth eigenvalue of A where both the sets are arranged in a nonincreasing order. The space M_s is equipped with the Shatten p-norm defined as [3]

$$\|A\|_{\mathrm{S},p} = \left[\sum_{j=1}^{s} \sigma_j(A)^p\right]^{1/p}, \quad p \in [1,\infty),$$

and

$$\|A\|_{\mathrm{S},\infty} = \sigma_1(A), \quad p = \infty.$$

For $p = 2$ we find the classical Frobenius norm and for $p = \infty$ we obtain the so-called "spectral" norm. When choosing $p = 1$, we find the so-called "trace" norm that for Hermitian nonnegative definite matrices equals the trace of the matrix. In the approximation of sequence of matrices of increasing dimension in simpler spaces of matrices, the preferred norm is generally the Frobenius norm. The first motivation is "practical" in the sense that this is the only Shatten p-norm whose calculation is computationally not expensive:

$$|A\|_{\mathrm{S},2} \equiv \left[\sum_{i,j=1}^{s} |(A)_{i,j}|^2\right]^{1/2}.$$

The second motivation is theoretical: actually the Frobenius norm is the only Shatten p-norm induced by an inner product which makes the space M_s into a Hilbert space. More specifically, setting $\langle A, B\rangle = \mathrm{trace}(A^{\mathrm{H}}B)$, we simply deduce that $\|A\|_{\mathrm{F}} \equiv \|A\|_{\mathrm{S},2} = \langle A, A\rangle^{1/2}$.

Therefore, if we want to solve a linear system $A_n\mathbf{y} = \mathbf{b}$ with A_n of "large" dimension d_n, we look for a convex closed set of matrices in which the computation (matrix inversion, matrix product, etc.) is inexpensive. Consequently, it is natural to consider the "least square approximation" problem of the given matrix A_n in order to devise a suitable preconditioner. This approach leads to the Frobenius-optimal approximation in algebras considered in [10,29] in the Toeplitz context and in [8] in the context of finite difference matrices discretizing elliptic differential operators. However, owing to the general analysis given in [13], it is easy to recognize that the Frobenius-optimal approximation in algebra is not completely satisfactory when the matrix sequence A_n is asymptotically ill conditioned with regard to n [13] or when the algebraic system is multilevel and the number d of levels is large [31,32]. It is worth noticing that multidimensional differential problems as well as multiresolution

or image processing naturally lead to multilevel structured matrices [9]. Therefore in this paper we look for approximations constructed using a different philosophy: in particular, as shown in the previous section, we invoke a "structural" approach [24,25,27,28,30] where the preconditioner approximating the matrix A_n is defined by using the continuous information contained in the model problem.

4.1. Analysis of the equal distribution

The following perturbation result is of paramount interest for estimating the *equal distribution* from a quantitative point of view. This result is a generalization of the Wielandt–Hoffman inequality (see e.g. [3]) and represents a particular case of the Lidskii–Mirsky–Wielandt Theorem whose proof can be found in [3, Theorem IV.3.4 and Example IV.3.5].

Lemma 4.1. *Let $p \in [1, \infty)$. For any pair of matrices $A, B \in M_s$, we have*

$$\left[\sum_{j=1}^{s} |\sigma_j(A) - \sigma_j(B)|^p\right]^{1/p} \leqslant \|A - B\|_{\mathrm{S},p}. \tag{12}$$

If A and B are Hermitian, then we also have

$$\left[\sum_{j=1}^{s} |\lambda_j(A) - \lambda_j(B)|^p\right]^{1/p} \leqslant \|A - B\|_{\mathrm{S},p}. \tag{13}$$

Given a function F continuous and with bounded support and the sequence $\{A_n\}_n$ and $\{B_n\}_n$ of $d_n \times d_n$ square matrices, we define

$$\Sigma(F, A_n, B_n) = \frac{1}{d_n} \sum_{i=1}^{d_n} F(\sigma_i(A_n)) - F(\sigma_i(B_n)) \quad (d_n < d_{n+1}).$$

Theorem 4.3. *Let $\{A_n\}_n$ and $\{B_n\}_n$ be two sequences of $d_n \times d_n$ matrices and let us suppose that $\|A_n - B_n\|_{\mathrm{S},p} \leqslant C(n)$. Assume that F is a Hölder continuous function with Hölder parameter $\alpha \in (0, 1]$. Then there exists a constant M so that*

$$|\Sigma(F, A_n, B_n)| \leqslant M[C(n)[d_n]^{-1/p}]^{\alpha}.$$

In particular, if $C(n) \leqslant c^*$, then

$$|\Sigma(F, A_n, B_n)| \leqslant M(c^*)^{\alpha}[d_n]^{-\alpha/p}.$$

Finally, if $C(n) \leqslant c^*$ and $p = \alpha = 1$, then we have strong equal distribution (SED).

Proof. By the assumptions it follows that $F \in \mathrm{Lip}(\alpha, M)$ with $\alpha \in (0, 1]$ and M positive constant. Therefore, we directly infer that

$$|\Sigma(F, A_n, B_n)| \leqslant \frac{1}{d_n} M \sum_{i=1}^{d_n} |\sigma_i(A_n) - \sigma_i(B_n)|^{\alpha}. \tag{14}$$

We call $\mathbf{s}(\alpha)$ the d_n-dimensional vector whose jth entry is $|\sigma_j(A_n) - \sigma_j(B_n)|^{\alpha}$ and we call $\mathbf{1}$ the d_n-dimensional vector of all ones. In this way inequality (14) can be rewritten as

$$|\Sigma(F, A_n, B_n)| \leqslant \frac{1}{d_n} M(\mathbf{s}(\alpha), \mathbf{1}), \tag{15}$$

where $(\mathbf{x}, \mathbf{y}) = \sum \bar{x}_i y_i$ is the usual inner product of $\mathscr{C}^{d_n}$. We now recall the Hölder inequalities, that is

$$|(\mathbf{x}, \mathbf{y})| \leqslant \|\mathbf{x}\|_t \|\mathbf{y}\|_{q(t)} \tag{16}$$

for any $t \geqslant 1$ and with $q(t)$ being the conjugated exponent so that $t^{-1} + [q(t)]^{-1} = 1$. Here $\|\mathbf{x}\|_s$ indicates $(\sum |x_i|^s)^{1/s}$ if $s \in [1, \infty)$ and $\max |x_i|$ if $s = \infty$.

The idea is to apply the Hölder inequality with $t = p/\alpha$ to Eq. (15) by obtaining the following chain of inequalities:

$$\begin{aligned} |\Sigma(F, A_n, B_n)| &\leqslant \frac{1}{d_n} M \|\mathbf{s}(\alpha)\|_{p/\alpha} \|\mathbf{1}\|_{q(p/\alpha)} \\ &= \frac{1}{d_n} M \|\mathbf{s}(1)\|_p^{\alpha} [d_n]^{1-\alpha/p}. \end{aligned}$$

At this point, by using the formal expression of the vector $\mathbf{s}(1)$ and Lemma 4.1 (in particular Eq. (12)), we obtain the following relation:

$$|\Sigma(F, A_n, B_n)| \leqslant M[d_n]^{-\alpha/p} \|A_n - B_n\|_{\mathrm{S},p}^{\alpha}. \tag{17}$$

By recalling that $\|A_n - B_n\|_{\mathrm{S},p} \leqslant C(n)$, the proof of the first part of the theorem is concluded. The other cases follow directly since they are special instances of the general formula displayed in (17). □

4.2. Analysis of the equal localization and of the clustering

Lemma 4.2. *Let $\{A_n\}_n$ and $\{B_n\}_n$ be two sequences of $d_n \times d_n$ matrices.*

1. *Assume* $\operatorname{rank}(A_n - B_n) = o(d_n)$. *Then the sequences $\{A_n\}_n$ and $\{B_n\}_n$ are equally localized* (*EL*) *and equally distributed* (*ED*).
2. *If* $\operatorname{rank}(A_n - B_n) = O(1)$, *then $\{A_n\}_n$ and $\{B_n\}_n$ are strongly equal localized* (*SEL*) *and strongly equally distributed* (*SED*).

Proof.

1. Let $r_n = \operatorname{rank}(A_n - B_n)$. As a consequence of the Cauchy interlace theorem [18] we have $\sigma_{i-2r_n}(B_n) \geqslant \sigma_i(A_n) \geqslant \sigma_{i+2r_n}(B_n)$ for $i = 2r_n + 1, \ldots, d_n - 2r_n$. Therefore, for any interval $[\alpha, \beta]$ we have

$$\#\{i : \sigma_i(A_n) \in [\alpha, \beta]\} = \#\{i : \sigma_i(B_n) \in [\alpha, \beta]\} + e_n \quad |e_n| \leqslant 4r_n. \qquad (18)$$

Consequently $r_n = o(d_n)$ and then the sequences $\{A_n\}_n$ and $\{B_n\}_n$ are equally localized (EL). The use of part 2 of Theorem 4.1 then leads to the equal distribution (ED).

2. If $r_n = O(1)$, then there is SEL by (18). For the proof of the last part, recall that F is Lipschitz continuous with bounded support contained in $M = [\alpha, \beta]$. Owing to its Lipschitzness, F is of bounded variation ($F \in BV$) too. Therefore it can be expressed as the sum of two monotone functions. By linearity it is enough to focus our attention on the monotone functions restricted to M. Let $S(A_n)$ and $S(B_n)$ be the sets of the singular values ordered nonincreasingly. Let q be an integer number and let $S(B_n, q)$ be such that $(S(B_n, q))_i = (S(B_n))_{i+q}$, $i = 1, \ldots, d_n$, where $(S(B_n))_j = \min\{\alpha, (S(B_n))_{d_n}\}$ if $j \geqslant d_n + 1$ and $(S(B_n))_j = \max\{\beta, (S(B_n))_1\}$ if $j \leqslant 0$. Now supposing that $r_n = O(1)$ i.e., $r_n \leqslant k$ for some positive k, we find that $S(B_n, -2k) \geqslant S(B_n)$, $S(A_n) \geqslant S(B_n, 2k)$, where "$\geqslant$" is intended componentwise. Finally, by monotonicity we deduce that

$$\begin{aligned}
&|\Sigma(F, A_n, B_n)| \\
&\quad \leqslant |\Sigma(F, S(B_n, -2k), S(B_n, 2k))| \\
&\quad = \left| \frac{1}{d_n} \sum_{i=1-2k,\ldots,2k,\, j=d_n-2k+1,\ldots,d_n+2k} F(\sigma_i(B_n)) - F(\sigma_j(B_n)) \right| \\
&\quad = O(d_n^{-1})
\end{aligned}$$

and the proof is complete. □

Lemma 4.3. *Let $\{A_n\}_n$ and $\{B_n\}_n$ be two sequences of $d_n \times d_n$ matrices.*

1. *If $\|A_n - B_n\|_{S,p}^p = o(d_n)$, $p \in [1, \infty)$ or $\|A_n - B_n\|_{S,\infty} = o(1)$, then $\{A_n\}_n$ and $\{B_n\}_n$ are ϵ equally localized (ϵ-EL) and equally distributed (ED).*
2. *When $\|A_n - B_n\|_{S,p} = O(1)$, $p \in [1, \infty)$, or $\|A_n - B_n\|_{S,\infty} = O(d_n^{-1})$, then $\{A_n\}_n$ and $\{B_n\}_n$ are ϵ strongly equally localized (ϵ-SEL).*
3. *If $p = 1$ and $\|A_n - B_n\|_{S,p} = O(1)$, then $\{A_n\}_n$ and $\{B_n\}_n$ are strongly equal distributed (SED).*

Proof.

1. We follow an idea indicated by Tyrtyshnikov in [35] for the case where $p = 2$. Let ϵ be a positive arbitrary number and $\gamma_n(\epsilon) = \#\{i : |\sigma_i(A_n) - \sigma_i(B_n)| > \epsilon\}$. By Theorem 4.1 (inequality (12)) for $p \in [1, \infty)$ we have

$$\sum_{i=1}^{d_n} |\sigma_i(A_n) - \sigma_i(B_n)|^p \leqslant \|A_n - B_n\|_{S,p}^p = o(d_n).$$

Now by definition of $\gamma_n(\epsilon)$ we deduce that

$$o(d_n) = \|A_n - B_n\|_{S,p}^p \geqslant \sum_{i=1}^{d_n} |\sigma_i(A_n) - \sigma_i(B_n)|^p \geqslant \gamma_n(\epsilon)\epsilon^p$$

that is $\gamma_n(\epsilon) = o(d_n)$. The latter relationship is by definition equivalent to ϵ-EL. In the case of $p = \infty$ the proof is trivial. Now by part 5 of Theorem 4.1 we deduce the ED property.

2. When $\|A_n - B_n\|_{S,p} = O(1)$, $p \in [1, \infty)$, or $\|A_n - B_n\|_{S,\infty} = O(d_n^{-1})$, then ϵ-SEL property is easily deduced by using the same argument as in the preceding part.
3. Finally if $p = 1$ and $\|A_n - B_n\|_{S,p} = O(1)$, then $\{A_n\}_n$ and $\{B_n\}_n$ are strongly equal distributed by the last part of Theorem 4.3. □

Theorem 4.4. *Let $\{A_n\}_n$ and $\{B_n\}_n$ be two sequences of $d_n \times d_n$ matrices.*

1. *If $\|A_n - B_n - D_n\|_{S,p}^p = o(d_n)$ with $p \in [1, \infty)$ and $\operatorname{rank}(D_n) = o(d_n)$, then $\{A_n\}_n$ and $\{B_n\}_n$ are equally distributed (ED).*
2. *If $\|A_n - B_n - D_n\|_{S,1} = O(1)$ with $\operatorname{rank}(D_n) = O(1)$, then $\{A_n\}_n$ and $\{B_n\}_n$ are strongly equal distributed (SED).*

Proof.

1. Let $X_n = B_n + D_n$. Then $\{A_n\}_n$ and $\{X_n\}_n$ fulfill the assumptions of part 1 of Lemma 4.3. Therefore $\{A_n\}_n$ and $\{X_n\}_n$ are ED. Moreover, $\{B_n\}_n$ and $\{X_n\}_n$ fulfill the assumptions of part 1 of Lemma 4.2 and consequently are ED. Since the ED relation is an equivalence relation, the transitivity yields the claimed result.
2. Let $X_n = B_n + D_n$. Therefore $\{A_n\}_n$ and $\{X_n\}_n$ are SED by part 3 of Lemma 4.3. Moreover $\{B_n\}_n$ and $\{X_n\}_n$ fulfill the assumptions of part 2 of Lemma 4.2 and consequently are SED. Since the SED relation is an equivalence relation, the proof is concluded by applying the transitivity. □

We prove the following corollaries with similar tools. In particular, the essentials of the proof of Corollary 4.1 can be found in [37].

Corollary 4.1. *Let $\{A_n\}_n$ and $\{B_n\}_n$ be two sequences of $d_n \times d_n$ matrices.*

1. *Suppose that $\|A_n - B_n\|_{S,p}^p = o(d_n)$ and $p \in [1, \infty)$. Then M is a cluster for $\{A_n\}_n$ iff it is a cluster for $\{B_n\}_n$.*
2. *When $\|A_n - B_n\|_{S,p} = O(1)$ with $p \in [1, \infty)$, then M is a proper cluster for $\{A_n\}_n$ iff it is a proper cluster for $\{B_n\}_n$.*

Proof.

1. Let M be a cluster for $\{A_n\}_n$. Then for any $\epsilon > 0$ we have

$$\gamma_n(A_n, M, \epsilon) = o(d_n), \quad \gamma_n(A_n, M, 2\epsilon) = o(d_n),$$

where the function γ_n is the one considered in Definition 2.2. More precisely $\gamma_n(A_n, M, \epsilon)$ measures the cardinality of $I_n(A_n, M, \epsilon)$ being the set of indices j

so that $\sigma_j(A_n) \notin M_\epsilon$. Now for any positive ϵ^*, let $J_n(A_n, B_n, \epsilon^*)$ be the set of indices j such that $|\sigma_j(A_n) - \sigma_j(B_n)| > \epsilon^*$. By Lemma 4.3, it holds that $\{A_n\}_n$ and $\{B_n\}_n$ are ϵ-EL and consequently $\#J_n(A_n, B_n, \epsilon^*) = o(d_n)$ for an arbitrary $\epsilon^* > 0$. For every $i \in U_n(\epsilon, \epsilon^*) \equiv J_n^c(A_n, B_n, \epsilon^*) \cap I_n^c(A_n, M, \epsilon)$ it simultaneously holds that

$$\sigma_i(A_n) \in M_\epsilon \quad \text{and} \quad |\sigma_i(A_n) - \sigma_i(B_n)| \leqslant \epsilon^*.$$

If $\epsilon^* < \epsilon$ and $i \in U_n(\epsilon, \epsilon^*)$, by triangle inequality it follows that $\sigma_i(B_n) \in M_{2\epsilon}$. Finally, recalling that $\#J_n^c(A_n, B_n, \epsilon^*) = d_n - o(d_n)$, $\#I_n^c(A_n, M, \epsilon) = d_n - o(d_n)$, it is transparent that

$$\#U_n(\epsilon, \epsilon^*) = n - o(d_n).$$

Since $U_n(\epsilon, \epsilon^*) \subset \{j : \sigma_j(B_n) \in M_{2\epsilon}\}$ and since ϵ is arbitrary it follows that M is a cluster for $\{B_n\}$ and the proof of the first part is concluded.

2. When $\|A_n - B_n\|_{S,p} = O(1)$ with $p \in [1, \infty)$, by following the same argument and by replacing each $o(d_n)$ by $O(1)$ we obtain the desired result. □

Corollary 4.2. *Let $\{A_n\}_n$ and $\{B_n\}_n$ be two sequences of $d_n \times d_n$ matrices and let M be a set of the real line so that for any positive ϵ, the set M_ϵ is made up of a finite number of intervals.*

1. *Suppose $\|A_n - B_n - D_n\|_{S,p}^p = o(d_n)$, $p \in [1, \infty)$ and* $\operatorname{rank}(D_n) = o(d_n)$. *Then M is a cluster for $\{A_n\}_n$ iff it is a cluster for $\{B_n\}_n$.*
2. *If $\|A_n - B_n - D_n\|_{S,p} = O(1)$ with* $\operatorname{rank}(D_n) = O(1)$, *$p \in [1, \infty)$, then M is a proper cluster for $\{A_n\}_n$ iff it is a proper cluster for $\{B_n\}_n$.*

Proof.

1. Let $X_n = B_n + D_n$. Then $\{A_n\}_n$ and $\{X_n\}_n$ have the same clusters by Corollary 4.1. But $\{X_n\}_n$ and $\{B_n\}_n$ fulfill the hypotheses of Lemma 4.2 so that $\{A_n\}_n$ and $\{B_n\}_n$ are EL. Therefore, by definition of EL matrix sequences, it follows that for any nontrivial interval $[\alpha, \beta]$ ($\alpha < \beta$), we have

$$\#\{i : \sigma_i(A_n) \in [\alpha, \beta]\} = \#\{i : \sigma_i(B_n) \in [\alpha, \beta]\} + o(d_n).$$

Since M_ϵ is (for any ϵ) a finite union of nontrivial intervals, the proof is concluded.

2. When $\|A_n - B_n - D_n\|_{S,p} = O(1)$ with $p \in [1, \infty)$ and $\operatorname{rank}(D_n) = O(1)$, by following the same argument and by replacing each $o(d_n)$ by $O(1)$ we obtain the desired result. □

Remark 4.4. In a certain sense, the limitations on M are academical. Indeed if M does not fulfill the requests of Corollary 4.2, then M must be unbounded and made up of an infinite number of unconnected parts. Notice that the fact that M_ϵ was finitely unconnected is essential in the proof of the preceding corollary. Indeed, if there exists a positive ϵ such that the set M_ϵ is made up of an infinite number of nonintersecting intervals, then it is possible to construct sequences $\{A_n\}_n$ and $\{B_n\}_n$

with $\operatorname{rank}(A_n - B_n) = 1$ for which M is a cluster for the first sequence but not for the second.

Some further consequences of Theorem 4.4 are listed below. For $p \in [1, \infty)$ it holds:

[Cl1] $\|A_n - B_n - D_n\|_{\mathrm{S},p} = \mathrm{O}(1)$, with $\operatorname{rank}(D_n) = \mathrm{O}(1)$ implies that $\forall \epsilon > 0$, all the singular values of $A_n - B_n$ belong to $[0, \epsilon)$ except $N_\mathrm{o} \equiv N_\mathrm{o}(n, \epsilon) = \mathrm{O}(1)$ outliers.

[Cl2] $\|A_n - B_n - D_n\|_{\mathrm{S},p}^p = \mathrm{o}(d_n)$, with $\operatorname{rank}(D_n) = \mathrm{o}(d_n)$ implies that $\forall \epsilon > 0$, all the singular values of $A_n - B_n$ belong to $[0, \epsilon)$ except $N_\mathrm{o} \equiv N_\mathrm{o}(n, \epsilon) = \mathrm{o}(d_n)$ outliers. The same is true when $\forall \epsilon > 0$, whereby there is a sequence of matrices $\{D_n(\epsilon)\}_n$ so that $\operatorname{rank}(D_n(\epsilon)) \leqslant \epsilon d_n$ and $\|A_n - B_n - D_n(\epsilon)\|_{\mathrm{S},p}^p \leqslant \epsilon d_n$ (see [36, Theorem 2]).

[Cl3] $\|A_n - B_n - D_n\|_{S,p} = \mathrm{O}(1)$, with $\operatorname{rank}(D_n) = \mathrm{O}(1)$ and such that the minimal singular value of B_n is greater than a fixed constant $\delta > 0$, imply that $\forall \epsilon > 0$, $\{B_n^{-1}(A_n - B_n)\}_n$ is properly clustered. (*Strong or proper clustering*.)

[Cl4] $\|A_n - B_n - D_n\|_{\mathrm{S},p}^p = \mathrm{o}(d_n)$, with $\operatorname{rank}(D_n) = \mathrm{o}(d_n)$ and such that the minimal singular value of B_n is greater than a fixed constant $\delta > 0$, imply that $\forall \epsilon > 0$, $\{B_n^{-1}(A_n - B_n)\}_n$ is generally clustered. (*Weak or general clustering*.)

Of course, if $\{A_n\}_n$ and $\{B_n\}_n$ are Hermitian and B_n is positive definite, then the properties and definitions [Cli], $i = 1, 2, 3, 4$ are in the sense of the eigenvalues. By referring to the case [Cl3], when the quantity $N_\mathrm{o} \equiv N_\mathrm{o}(n, \epsilon)$ goes to infinity as n goes to infinity and ϵ goes to zero (see [25]), then we have the "weakest" case of strong clustering. As we will see in the sequel this is one of the "peculiar" cases that we encounter and we will call it *Weakest Strong Clustering*.

Finally, with regard to [Cl3] and [Cl4], we remark that the assumption that "the minimal singular value of B_n greater than a fixed $\delta > 0$" is necessary and cannot be removed (take $B_n = n^{-10} I$ and $A_n = n^{-2} I$ for a counterexample both for [Cl3] and [Cl4]). But, if B_n has, at most, $\mathrm{o}(d_n)$ singular values going to zero as n goes to infinity, then we obtain again a weak clustering. The question is connected with the concept of *sparsely vanishing* functions discussed in [13,34] and is analyzed in Section 6.

Definition 4.7. A real-valued nonnegative measurable function $a(\mathbf{x})$ defined in $K \subset \Re^d$, K compact, is sparsely vanishing if

$$\lim_{\epsilon \to 0} m\{\mathbf{x} \in K : a(\mathbf{x}) \leqslant \epsilon\} = 0$$

with $m\{\cdot\}$ being the Lebesgue measure on $\mathbb{R}^d$.

The following technical lemmata are useful in order to overcome the restriction about the minimal singular value of B_n emphasized in [Cl4].

Lemma 4.4. *Suppose that $\{A_n\}_n$ and $\{B_n\}_n$ are two sequences of matrices of dimension $d_n \times d_n$, and that B_n is nonsingular and let us call σ_i the singular values of B_n. If there exists a nonnegative function $x(\epsilon)$ independent of n so that, for any ϵ and n large enough*,

$$\frac{\#\{i : \sigma_i \leqslant \epsilon\}}{d_n} \leqslant x(\epsilon), \quad \lim_{\epsilon \to 0} x(\epsilon) = 0 \tag{19}$$

and if there exist matrices $\{D_n(\epsilon)\}_n$, $D_n = D_n(\epsilon)$ such that

$$\lim_{n \to \infty} \|A_n - B_n - D_n\|_{\mathrm{S},\infty} = 0 \tag{20}$$

with $\mathrm{rank}(D_n) \leqslant \epsilon d_n$, *then the weak clustering property holds.*

Proof. By the assumptions, for any $\epsilon > 0$, we find matrices $\{D_n(\epsilon)\}_n$, $D_n = D_n(\epsilon)$ such that for *n* large enough

$$\|A_n - B_n - D_n\|_{S,\infty} \leqslant \epsilon^2,$$

with $\mathrm{rank}(D_n) \leqslant \epsilon d_n$. Let us consider $B_n^{-1}A_n$ and let us analyze its singular values distribution. We set $X_n = A_n - B_n - D_n$; therefore

$$B_n^{-1}A_n = I + B_n^{-1}D_n + B_n^{-1}X_n$$

with $\mathrm{rank}(B_n^{-1}D_n) = \mathrm{rank}(D_n) \leqslant \epsilon d_n$. So we have to evaluate the structure of $B_n^{-1}X_n$. Let $B_n(\delta)$ be the matrix B_n written in its singular value decomposition where each singular value less than δ has been replaced by δ. Clearly $Y_n = B_n - B_n(\delta)$ has rank at most equal to $b(\delta) = \#\{i : \sigma_i \leqslant \delta\}$ and so

$$\begin{aligned} B_n^{-1}X_n &= (B_n(\delta) + Y_n)^{-1}X_n \\ &= (I + B_n(\delta)^{-1}Y_n)^{-1}B_n(\delta)^{-1}X_n \\ &= (I + W_n)B_n(\delta)^{-1}X_n, \end{aligned}$$

where W_n is obtained by applying the Shermann–Morrison–Woodbury formula for inverting $(I + B_n(\delta)^{-1}Y_n)$. Therefore W_n has, at most, the same rank as Y_n. Consequently, the matrix $B_n^{-1}X_n$ has been written as the sum of $W_nB_n(\delta)^{-1}X_n$ and $B_n(\delta)^{-1}X_n$, the first having rank bounded by $b(\delta) = x(\delta)d_n$, with $\lim_{\delta \to 0} x(\delta) = 0$, the second having spectral norm bounded by

$$\frac{\|X_n\|_{\mathrm{S},\infty}}{\delta} \leqslant \frac{\epsilon^2}{\delta}.$$

Now, by choosing $\delta = \epsilon$ and by applying the minmax theorem (modified for dealing with the singular values), the result is plainly proved. □

Remark 4.5. Observe that the assumption of nonsingularity of the matrices B_n can be easily removed if we consider the pseudo-inverse of Moore–Penrose B_n^+ instead of the usual inverse B_n^{-1}.

Remark 4.6. It is worth pointing out that relation (20) can be replaced by the following:

$$\lim_{n\to\infty} \frac{\|A_n - B_n - D_n\|_{\mathrm{S},p}^p}{d_n} = 0$$

with any choice of $p \in [1, \infty)$.

Remark 4.7. Assume that $\{A_n\}_n$ and $\{B_n\}_n$ are as in Lemma 4.4. If in addition the matrices $\{A_n\}_n$ and $\{B_n\}_n$ are Hermitian and $\{B_n\}_n$ are positive definite, then the *weak clustering* property holds both for the singular values and for the eigenvalues. To see how to translate from singular values to eigenvalues use relations (19). Finally notice that the matrices of $\{B_n^{-1}A_n\}_n$ are not necessarily Hermitian.

In order to meet the hypotheses of the latter lemma in the context of our differential problems, this result will be useful.

Lemma 4.5. *If $a : K \to \mathbb{R}$, $K = \bar{\Omega} = [0, 1]^d$ is nonnegative and sparsely vanishing, then*

$$\lim_{\epsilon\to 0} \limsup_{n\to\infty} \frac{\#\{i : a(\mathbf{x}_i) \leqslant \epsilon\}}{N(n)} = 0.$$

Here, if a is continuous, then $a(\mathbf{x}_i)$ has the usual meaning. Otherwise the symbol $a(\mathbf{x}_i)$ indicates the quantity $N(n)\int_{I_i} a(\mathbf{t})\mathrm{d}\mathbf{t}$, $I_i = \prod_{j=1}^d [\mathbf{x}_{i_j}, \mathbf{x}_{i_j+1}]$.

Proof. The proof is given in the unidimensional case $d = 1$. The multidimensional case can be treated in the same way.

We essentially use the definition of sparsely vanishing function and the relation $x_{i+1} - x_i = (n + 1)^{-1}$. More precisely, let us call J_ϵ the set of indices $\{i : a(x_i) \leqslant \epsilon\}$. The cardinality of J_ϵ is a monotone nondecreasing function with ϵ. Therefore, there exists a nonnegative $c \leqslant 1$ such that

$$\lim_{\epsilon\to 0} \limsup_{n\to\infty} \frac{\#\{i : a(x_i) \leqslant \epsilon\}}{n} = c.$$

Let us call $X_\epsilon = \bigcup_{j\in J_\epsilon} I_j$. Therefore, from the preceding limit relation, we deduce that $m(X_\epsilon) = \sum_{j\in J_\epsilon} m(I_j) = c + \mathrm{o}(1)$, where the $\mathrm{o}(1)$ is with respect to ϵ and n separately. Now, let us take δ positive value and let us define $X_{\epsilon,\delta} = \{x \in X_\epsilon : a(x) \leqslant \delta\}$. The following facts hold true:

- $\int_{X_\epsilon} a(t)\,\mathrm{d}t = \sum_{j\in J_\epsilon} \int_{I_j} a(t)\mathrm{d}t \leqslant (c + \mathrm{o}(1))\epsilon$.
- $h(\delta) = m(X_{\epsilon,\delta}) \leqslant m(x : a(x) \leqslant \delta)$ and so, since a is sparsely vanishing, we have $\lim_{\delta\to 0} h(\delta) = 0$.

- $\int_{X_\epsilon} a(t)\,\mathrm{d}t \geqslant \int_{X_\epsilon / X_{\epsilon,\delta}} a(t)\,\mathrm{d}t$ which is greater than $\delta m(X_\epsilon / X_{\epsilon,\delta}) \geqslant \delta(c + \mathrm{o}(1) - h(\delta))$.

We now join the preceding results by obtaining

$$(c + \mathrm{o}(1))\epsilon \geqslant \delta(c + \mathrm{o}(1) - h(\delta)).$$

Choose $\delta = \sqrt{\epsilon}$ and divide by δ.

$$(c + \mathrm{o}(1))\sqrt{\epsilon} \geqslant (c + \mathrm{o}(1) - h(\sqrt{\epsilon})).$$

Finally, if we calculate the limit as ϵ goes to zero, we deduce $0 \geqslant c$ that is $c = 0$. □

5. Some results on the smooth case

We now come back to the structured sequences of matrices that arise in the FD discretization of problems (1) and (2). We analyze them in some detail by showing several connections with the results of Section 4.

Preliminarily we recall a chain of results on the relationships among $\Delta_{2k;d}^{(n)}$, $A_n(a)$ and P_n proved in [25,27,30] under the assumption that a is smooth (at least continuous).

From hereon we make the assumption that $\exists a_1, \ldots, a_d$ with $a_j \in \mathcal{N}_+$ such that the multiindex $n = (n_1, \ldots, n_d)$ is such that $n_j + 1 = v a_j$, $v \in \mathcal{N}_+$.

Lemma 5.1 [25,27]. *If a is nonnegative with at most isolated zeros, then the preconditioned matrix $P_n^{-1} A_n(a)$ is similar to $[\Delta_{2k;d}^{(n)}]^{-1} A_n^*(a)$ where, for n large enough, $A_n^*(a)$ is the SPD matrix given by*

$$t_k D_{n,a}^{-1/2} A_n(a) D_{n,a}^{-1/2} \quad \text{for } t_k = \left(\sum_{j=1}^{d} a_j^{2k} \right) \binom{2k}{k}.$$

Theorem 5.1 [25,30]. *The eigenvalues of $[\Delta_{2k;d}^{(n)}]^{-1} A_n(a)$ belong to $[a, A]$, where $a = \inf a(x)$ and $A = \sup a(x)$. Moreover, if $a \in \mathbf{C}^2([0,1]^d)$ is strictly positive and $k = 1$, then there exist two positive constants c, C such that the spectrum of $P_n^{-1} A_n(a)$ is contained in $[c, C]$.*

Lemma 5.2 [25,30]. *Assume that a is strictly positive. If $a \in \mathbf{C}^2([0,1]^d)$, then the matrix $A_n^*(a)$ can be expanded in the following way*:

$$\begin{aligned} A_n^*(a) &= \Delta_{2k}^{(n)} + \mathrm{O}(\|h\|_\infty^2) E, \\ h &= (h_1, h_2, \ldots, h_d), \\ h_j &= (n_j + 1)^{-1}, \end{aligned} \tag{21}$$

where E has the same pattern as $A_n(a)$ and is bounded in spectral norm. If $a \in \mathbf{C}([0,1]^d)$, then $A_n^*(a) = \Delta_{2k}^{(n)} + \Theta_h$, where Θ_h has the same pattern as $A_n(a)$ and the magnitude of its entries is $\mathrm{O}(\omega_a(\|h\|_\infty))$.

Theorem 5.2 [25,30]. *If $a \in \mathbf{C}^2([0,1]^d)$ is strictly positive and $k = 1$, then, for any sequence ϵ_n decreasing to zero (as slowly as we want), $\forall \epsilon > 0$, $\exists \bar{n}$ such that, if $n \geqslant \bar{n}$, then $N(n) - 2\lceil \epsilon_n^{-1} \rceil$ eigenvalues of the preconditioned matrix $P_n^{-1} A_n(a)$ are in $(1-\epsilon, 1+\epsilon)$ (weakest strong clustering property). On the other hand, if $k > 1$, then the cluster is weak.*

The matrices $\Delta_{2k;d}^{(n)}$ deserve some attention because their eigenvalues behave like the sampling of sparsely vanishing functions. This fact is observed in the subsequent theorem and remarks.

Theorem 5.3. *Let $\Delta_{2k}^{(n)}$ be the $n \times n$ Toeplitz matrix obtained by the FD discretization of problem* (1) *with precision order* 2 *and $a \equiv 1$. The matrix $\Delta_{2k}^{(n)}$ is positive definite and the limit relation*

$$\lim_{\epsilon \to 0} \limsup_{n \to \infty} \frac{\#\{i : \lambda_i^{(n)} \leqslant \epsilon\}}{n} = 0 \tag{22}$$

holds true, where $\{\lambda_i^{(n)}\}$ are the eigenvalues of $\Delta_{2k}^{(n)}$ arranged in nondecreasing order.

Moreover, if $\Delta_{2k;d}^{(n)}$ is the $N(n) \times N(n)$ multilevel Toeplitz matrix generated by the FD discretization of problem (2) *with precision order* 2 *and $a \equiv 1$, then $\Delta_{2k;d}^{(n)}$ is positive definite and the limit relation*

$$\lim_{\epsilon \to 0} \limsup_{n \to \infty} \frac{\#\{i : \lambda_i^{(n)} \leqslant \epsilon\}}{N(n)} = 0 \tag{23}$$

hold true. Here $\{\lambda_i^{(n)}\}$ denotes the complete set of the eigenvalues of $\Delta_{2k;d}^{(n)}$ arranged in nondecreasing order.

Proof. From the analysis of the coefficients, it follows that the generating function of the Toeplitz matrix $\Delta_{2k}^{(n)}$ is $(2 - 2\cos(x))^k = 2^{2k} \sin^{2k}(x/2)$ which has a unique zero in $I = (-\pi, \pi)$ at $x = 0$ and is strictly positive elsewhere: as a consequence it is *sparsely vanishing*. From this and according to Theorem 3.3 and Remark 4.2, we deduce that

- $\Delta_{2k}^{(n)}$ is positive definite and its singular values coincide with the eigenvalues.
- 0 is not a *sub-cluster point* for the eigenvalues.

The latter two statements are just a rewriting of (22).

In the multilevel setting we observe that the generating function of $\Delta_{2k;d}^{(n)}$ is the multivariate function

$$\sum_{j=1}^{d} a_j^{2k}(2-2\cos(x_j))^k$$

which is zero at zero and is positive elsewhere. From now on the proof is identical to the unidimensional case. □

It is possible to consider a more "constructive" way for proving Theorem 5.3 that uses a standard procedure in structured linear algebra. Indeed we "correct" the matrix $\Delta_{2k}^{(n)}$ in the τ algebra [5]. More specifically, let $H_n(k)$ be the $n \times n$ symmetric Hankel matrix [5] whose first row is given by $[\tau_2, \ldots, \tau_k, \mathbf{0}]$ and whose last row is given by the reversed vector $[\mathbf{0}, \tau_k, \ldots, \tau_2]$, the coefficients $\{\tau_j\}$ being those of the Toeplitz matrix $\Delta_{2k}^{(n)}$ (see Sections 2.2 and 3.1). Then the matrix

$$\tau(\Delta_{2k}^{(n)}) = \Delta_{2k}^{(n)} - H_n(k)$$

belongs to the τ algebra [5]. Moreover, it is easy to verify that the ith eigenvalue of $\tau(\Delta_{2k}^{(n)})$ coincides with $\hat{\lambda}_i^{(n)} = 2^{2k}\sin^{2k}(x_i^{(n)}/2)$, where $x_i^{(n)} = \mathrm{i}\pi/(n+1)$ (see [14]) and the function $2^{2k}\sin^{2k}(x/2)$ is sparsely vanishing since it is continuous and has a unique zero at $x = 0$. Therefore, in the light of Lemma 4.5, it follows that

$$\lim_{\epsilon\to 0}\limsup_{n\to\infty}\frac{\#\{i : \hat{\lambda}_i^{(n)} \leqslant \epsilon\}}{n} = 0.$$

We now observe that $H_n(k)$ has rank $r_k = 2(k-1)$ which is constant with respect to the dimension n. Consequently, by part 2 of Lemma 4.2, we plainly infer that $\{\tau(\Delta_{2k;d}^{(n)})\}_n$ and $\{\Delta_{2k;d}^{(n)}\}_n$ are SEL so that

$$\lim_{\epsilon\to 0}\limsup_{n\to\infty}\frac{\#\{i : \lambda_i^{(n)} \leqslant \epsilon\}}{n} = 0,$$

where $\{\lambda_i^{(n)}\}$ denotes the eigenvalues of $\Delta_{2k}^{(n)}$ and so the proof in the unidimensional case is concluded.

Analogously, in the multidimensional case we correct the Toeplitz matrix $\Delta_{2k;d}^{(n)}$ in the d-level τ algebra [5] so that

$$\tau(\Delta_{2k;d}^{(n)}) = \Delta_{2k;d}^{(n)} - H_n(k;d),$$

where the matrix $H_n(k;d)$ is a d-level Hankel matrix whose rank $r_{k,n}$ is such that

$$r_{k,n} = \mathrm{o}\left(N(n)\sum_{j=1}^{d} n_j^{-1}\right).$$

Since $r_{k,n} = \mathrm{o}(N(n))$, the application of part 1 of Lemma 4.2 yields the desired result.

Remark 5.1. What the preceding linear algebra proof revealed is that the matrix sequences $\{\tau(\Delta_{2k;d}^{(n)})\}_n$ and $\{\Delta_{2k;d}^{(n)}\}_n$ are EL and properly bounded (in the sense of

Definition 4.4). Therefore by part 2 of Theorem 4.1 and part 1 of Theorem 4.2 they are ED and ϵ-EL. Moreover, for $d = 1$ they are SEL and consequently by part 2 of Theorem 4.2 they are also ϵ-SEL. Finally for $d = 1$ we observe that $\|\tau(\Delta_{2k;d}^{(n)}) - \Delta_{2k;d}^{(n)}\|_{S,1} = O(1)$ and consequently by Theorem 4.3 we deduce that $\{\tau(\Delta_{2k;d}^{(n)})\}_n$ and $\{\Delta_{2k;d}^{(n)}\}_n$ are SED.

Remark 5.2. Since $\{\Delta_{2k;d}^{(n)}\}_n$ is distributed as the polynomial $f(\mathbf{x}) = \sum_{i=1}^{d} a_j^{2k}(2 - 2\cos(x_i))^k$ by Theorem 3.1, and f is a polynomial not identically zero, it follows that $m\{\mathbf{x} \in I^d : |f((\mathbf{x})| = p\} = 0$ for any $p \in \mathbb{R}$. Therefore in the light of Theorem 3.2 we infer that $\{\Delta_{2k;d}^{(n)}\}_n$ has no sub-cluster points. Now $\{\tau(\Delta_{2k;d}^{(n)})\}_n$ and $\{\Delta_{2k;d}^{(n)}\}_n$ are ED by Remark 5.1 and therefore the sequence $\{\tau(\Delta_{2k;d}^{(n)})\}_n$ has no sub-cluster points.

The following results now take into account the presence of isolated zeros.

Theorem 5.4 [25,30]. *If a has a unique zero at $\mathbf{x} = 0$ of order α, then preconditioned matrix*

$$\left[\Delta_{2k;d}^{(n)}\right]^{-1} A_n(a)$$

has eigenvalues in $(0, A]$, *A being the maximum of the function a. In addition, the lower bound is tight in the sense that the smallest eigenvalue*

$$\lambda_1\left(\left[\Delta_{2k;d}^{(n)}\right]^{-1} A_n(a)\right)$$

of the first preconditioned matrix tends to zero as n tends to infinity. In particular we have

$$\lambda_1 = O(\|h\|_\infty^\alpha), \quad h = (h_1, h_2, \ldots, h_d), \quad h_j = (n_j + 1)^{-1}.$$

The same is true if the unique zero is located elsewhere.

Therefore, if a has zeros, then $\{\Delta_{2k;d}^{(n)}\}_n$ cannot be a good preconditioner for $\{A_n(a)\}_n$ according to points 1.a and 1.b in Section 2.1.

6. General results on distribution and clustering

The aim of this section is to give general results on distribution and clustering for the matrix sequences $\{A_n^*(a)\}$, $\{\Delta_{2k;d}^{(n)}\}$, and $\{P_n^{-1}A_n(a)\}$. First we consider the case where a is (at least) continuous and then the case where a is not.

Lemma 6.1. *Let $A_n^*(a)$ be the symmetrical scaling of the matrix related to the problem* (1) *or* (2) *as in Lemma* 5.1. *Let $\Delta_{2k;d}^{(n)}$ be the multilevel Toeplitz matrix defined as $A_n(1)$. If a is continuous, strictly positive and ω_a is its modulus of continuity and $j = (j_1, j_2, \ldots, j_d)$, $i = (i_1, i_2, \ldots, i_d)$, then we find that*

$$(A_n^*(a))_{i,i\pm j} = (\Delta_{2k;d}^{(n)})_{i,i\pm j} + \mathrm{O}(\omega_a(\|h\|_\infty))$$

for $j \leqslant k \cdot e^{\mathrm{T}}$ (and where the meaning of $\leqslant$ is in the sense of the partial ordering of $\mathbb{R}^d$ and $e^{\mathrm{T}} = (1, \ldots, 1)$). If a is nonnegative with m isolated zeros, then for any positive ϵ there exist matrices $D_n^{(1)}, \ldots, D_n^{(m)}$, $D_n^{(j)} = D_n^{(j)}(\epsilon)$ having rank bounded by $\epsilon N(n)$ such that $A_n^{**}(a) = A_n^*(a) - (D_n^{(1)} + \cdots + D_n^{(m)})$ and

$$(A_n^{**}(a))_{i,i\pm j} = (\Delta_{2k;d}^{(n)})_{i,i\pm j} + \mathrm{O}(\omega_a(\|h\|_\infty))$$

for $j \leqslant ke^{\mathrm{T}}$.

Proof. For $j = 0$ we find $(A_n^*(a))_{i,i} = (\Delta_{2k;d}^{(n)})_{i,i}$. For $j \neq 0$, each element $(A)_{i,i\pm j}$ is a finite sum (at most dk terms) of evaluations of a in close points (their distance is bounded by $\|k\|_\infty \|h\|_\infty$) multiplied by values $b_t^{(k,j)}$ whose sum over the indices t is exactly $(\Delta_{2k;d}^{(n)})_{i,i\pm j}$ (see Eq. (5)). Now the result follows from the definition of $A_n^*(a)$. When a is nonnegative having exactly m zeros, the proof can be performed as in Theorem 4.2 of [25]. □

With the help of the preceding lemma, we can prove the following theorem.

Theorem 6.1. *Let $A_n^*(a)$ be the matrix related to problem* (1) *or* (2) *symmetrically scaled as in Lemma* 5.1 *and $\Delta_{2k;d}^{(n)}$ be the multilevel Toeplitz matrix defined as $A_n(1)$. If a is continuous and strictly positive and ω_a is its modulus of continuity, then we have*:

$$\left\| A_n^*(a) - \Delta_{2k;d}^{(n)} \right\|_{\mathrm{S},p}^p = N(n)\mathrm{O}(\omega_a^p(\|h\|_\infty))$$

with $p \in [1, \infty)$ and

$$\left\| A_n^*(a) - \Delta_{2k:d}^{(n)} \right\|_{\mathrm{S},\infty} = \mathrm{O}(\omega_a(\|h\|_\infty)).$$

If a is nonnegative with m isolated zeros, then for any positive ϵ there exists a matrix $D_n = D_n^{(1)} + \cdots + D_n^{(m)}$ having rank bounded by $\epsilon N(n)$, ($D_n^{(j)} = D_n^{(j)}(\epsilon)$ as in the previous lemma) such that $A_n^{**}(a) = A_n^*(a) - (D_n^{(1)} + \cdots + D_n^{(m)})$ and

$$\left\| A_n^{**}(a) - \Delta_{2k;d}^{(n)} \right\|_{\mathrm{S},p}^p = N(n)\mathrm{O}(\omega_a^p(\|h\|_\infty))$$

with $p \in [1, \infty)$ and

$$\left\| A_n^{**}(a) - \Delta_{2k}^{(n)} \right\|_{\mathrm{S},\infty} = \mathrm{O}(\omega_a(\|h\|_\infty)).$$

Here the constants hidden in the "big O*" terms can depend on ϵ.*

Proof. It is a simple consequence of the preceding Lemma 6.1 of the bandedness of all the involved matrices. □

Remark 6.1. The latter result, in view of Theorem 4.4, tells us that $\{A_n^*(a)\}_n$ and $\{\Delta_{2k;d}^{(n)}\}_n$ are ED in the sense of the eigenvalues ($A_n^*(a)$ and $\Delta_{2k;d}^{(n)}$ are symmetric). But $\Delta_{2k;d}^{(n)}$ is the Toeplitz matrix generated by $\sum_{i=1}^d a_j^{2k}(2-2\cos(x_i))^k$ and therefore, by taking into account the ergodic theorem, Theorem 3.1, we have

$$\lim_{n\to\infty}\frac{1}{N(n)}\sum_{i=1}^{N(n)}F\left(\lambda_i^{(n)}\right)$$
$$=\frac{1}{[2\pi]^d}\int_{(-\pi,\pi)^d}F\left(\sum_{i=1}^d a_j^{2k}(2-2\cos(x_i))^k\right)\mathrm{d}\mathbf{x} \tag{24}$$

with $\lambda_i^{(n)}$, $i=1,2,\ldots,N(n)$, being the eigenvalues of $A_n^*(a)$.

Remark 6.2. For $d=1$, Widom has proven a second-order result [39] for the eigen/singular values of Toeplitz matrices generated by symbols in the Krein algebra $\mathscr{K}$ of all the essential bounded functions over $I=(-\pi,\pi)$ whose Fourier coefficients $\{\tau_k\}$ are such that $\sum_k |k||\tau_k|^2<\infty$. More specifically the quoted result is the following:

$$\frac{1}{n}\sum_{i=1}^n F(\lambda_i^{(n)})-\frac{1}{2\pi}\int_{(-\pi,\pi)}F(f(x))\mathrm{d}x=\mathrm{O}(n^{-1})$$

with $\lambda_i^{(n)}$, $i=1,2,\ldots,n$, being the eigenvalues of $T_n(f)$, $f\in\mathscr{K}$ and F regular enough. We now remark that all the trigonometric polynomials are in the Krein algebra and that (by Theorem 6.1)

$$\left\|A_n^*(a)-\Delta_{2k}^{(n)}\right\|_{S,1}\leqslant n\mathrm{O}(\omega_a(\|h\|_\infty))$$

when a is positive and continuous. Therefore, if a is also Lipschitz continuous, we deduce that $\|A_n^*(a)-\Delta_{2k}^{(n)}\|_{S,1}=\mathrm{O}(1)$ so that the application of Theorem 4.3 tells us that $\{A_n^*(a)\}_n$ and $\{\Delta_{2k}^{(n)}\}_n$ are SED. Finally, the combination of the Widom result with the SED property yields the following relation:

$$\frac{1}{n}\sum_{i=1}^n F(\lambda_i^{(n)})-\frac{1}{2\pi}\int_{(-\pi,\pi)}F((2-2\cos(x))^k)\mathrm{d}x=\mathrm{O}(n^{-1})$$

with $\lambda_i^{(n)}$, $i=1,2,\ldots,n$, being the eigenvalues of $A_n^*(a)$.

Remark 6.3. Through Remarks 5.1 and 5.2, we know that $\{\Delta_{2k;d}^{(n)}\}_n$ is distributed as the polynomial $f(\mathbf{x})=\sum_{i=1}^d a_j^{2k}(2-2\cos(x_i))^k$, has no sub-cluster points (regularity), and is properly bounded. Now $\{A_n^*(a)\}_n$ and $\{\Delta_{2k;d}^{(n)}\}_n$ are ED by Remark

6.1 and therefore the sequence $\{\tau(\Delta_{2k;d}^{(n)})\}_n$ has no sub-cluster points (regularity) and is essentially bounded (properly bounded if a is strictly positive). Therefore parts 3 and 4 of Theorem 4.2 imply that $\{A_n^*(a)\}_n$ and $\{\Delta_{2k;d}^{(n)}\}_n$ are EL and ϵ-EL.

Theorem 6.2. *If $a \in \mathbf{C}([0,1]^d)$ and has a unique isolated zero, then all the eigenvalues of the preconditioned matrix $P_n^{-1}A_n(a)$ lie in $(1-\epsilon, 1+\epsilon)$ except N_0 outliers with $N_0 \equiv N_0(n,\epsilon) = o(N(n))$ (weak clustering property).*

Proof. In the light of Lemma 5.1, we can analyze the spectrum of the matrix $[\Delta_{2k;d}^{(n)}]^{-1}A_n^*(a)$. Moreover, by the preceding theorem, we find that for any ϵ and n large enough, we deduce that

$$\operatorname{rank}(A_n^{**}(a) - A_n^*(a)) \leqslant \epsilon N(n)$$

and $\|A_n^{**}(a) - \Delta_{2k;d}^{(n)}\|_{S,\infty} = o(1)$. Finally, in the light of Theorem 5.3, the eigenvalues of $\Delta_{2k;d}^{(n)}$ behave as the sampling of a sparsely vanishing function. Therefore, by setting $A_n = A_n^*(a)$ and $B_n = \Delta_{2k;d}^{(n)}$, the hypotheses of Lemma 4.4 and Remark 4.7 are fulfilled and, consequently, we have that

$$\left\{ P_n^{-1}A_n(a) \sim \left[\Delta_{2k;d}^{(n)}\right]^{-1} A_n^*(a) \right\}_n$$

has a weak clustered spectrum both with regard to the singular values and to the eigenvalues. □

Moreover, it is worth pointing out that Eqs. (1) and (2) impose that $a \in C^q$, $q = \|k\|_\infty$ and so it seems that a refined analysis is just an academic exercise. However, when we consider the "weak formulation" [11], problem (1) is transformed into an integral problem. Therefore, in this sense, the given analysis becomes again meaningful.

We are able to prove something more concerning this fact. From the Lusin Theorem, we know that it is possible to approximate $a \in L^\infty(K)$ (with *compact* K), by a family $c_\epsilon \in C(K)$ with respect to the topology induced by the convergence in measure. This result is used to prove that the preconditioned matrix sequence $\{P_n^{-1}A_n(a)\}_n$ is clustered at one with a being just L^∞.

Theorem 6.3. *Let $A_n^*(a)$ be the matrix related to problem* (1) *symmetrically scaled as in Lemma* 5.1 *and $\Delta_{2k}^{(n)}$ be the Toeplitz matrix defined as $A_n(1)$ with $d = 1$. Here the coefficients $a(x_i)$ should be replaced by the mean value on the interval $I_i = [x_i, x_{i+1}]$ in the sense that $a(x_i)$ means $(n+1)\int_{I_i} a(t)$.*

- *If $a \in L^\infty$ is nonnegative and sparsely vanishing, then we find that there exist matrices D_n having $o(n)$ rank such that*

$$\left\| A_n^*(a) - \Delta_{2k}^{(n)} - D_n \right\|_{S,p}^p = o(n)$$

and

$$\left\| A_n^*(a) - \Delta_{2k}^{(n)} - D_n \right\|_{S,\infty} = \mathrm{o}(1).$$

In addition, the number of outliers of $\{P_n^{-1}A_n(a)\}_n$ *is generically* $\mathrm{o}(n)$.

- *If* a *is not sparsely vanishing, then* $A_n(a)$ *and* P_n *may fail to be invertible or* $\{P_n^{-1}A_n(a)\}_n$ *may not have clustered eigenvalues.*

Proof. In view of the Lusin Theorem [23], for any positive ϵ, we find a continuous function $c = c_\epsilon$ such that a differs from c only in an ϵ measure set and is bounded in infinity norm by $\|a\|_\infty$. In addition, since a is sparsely vanishing, it follows that $\{c_\epsilon\}_\epsilon$ is such that

$$m\{x \in [0,1] : \ |c_\epsilon| \leqslant \epsilon\} \to 0 \quad \text{for } \epsilon \to 0.$$

On the other hand, we may replace c with $\max\{c, \epsilon\}$ to avoid the possibility that c takes nonpositive values. Now $A_n(a) = A_n(c) + (A_n(a) - A_n(c))$ and so

$$A_n^*(a) = A_n^*(c) + X_1 + X_2 + X_3,$$

where

$$\begin{aligned} X_1 &= D_{n,a}^{-1/2} A_n(a-c) D_{n,a}^{-1/2}, \\ X_2 &= \left(D_{n,a}^{-1/2} - D_{n,c}^{-1/2} \right) A_n(c) D_{n,a}^{-1/2}, \\ X_3 &= D_{n,c}^{-1/2} A_n(c) \left(D_{n,a}^{-1/2} - D_{n,c}^{-1/2} \right). \end{aligned}$$

Clearly, in view of the preceding results, we have $A_n^*(c) = \Delta_{2k}^{(n)} + \mathrm{O}(\omega_c(n^{-1})) + DD_n$, where DD_n has $\mathrm{o}(n)$ rank and is due to the presence of zeros. On the other hand, since $a - c \neq 0$ only in a set of measure ϵ and c is bounded, it follows that c is an L^1 approximation of a in the sense that $\|a - c\|_{L^1} \leqslant \epsilon 2\|a\|_\infty$. From this, we can deduce that $A_n(a-c)$ is a matrix which can be written as the sum of two matrices, the first of rank $\mathrm{o}(n)$ and the second of small norm (this is evident and trivial when the set $\{x : \ a(x) \neq c(x)\}$ is made up of a finite number of intervals). Actually, the key point is the remark that

$$\begin{aligned} \sum_{i=1}^{n} |a_i - c_i| &= \sum_{i=1}^{n} (n+1) \left| \int_{I_i} a(t) - c(t) \, \mathrm{d}t + \mathrm{o}(1) \right| \\ &\leqslant (n+1) \left(\int |a(t) - c(t)| \mathrm{d}t + \mathrm{o}(1) \right) \\ &\leqslant 4\epsilon \|a\|_\infty (n+1). \end{aligned}$$

Since ϵ can be chosen arbitrarily small, it is evident that $|a_i - c_i| \leqslant \epsilon$ with the exception of, at most, $\alpha(\epsilon)n$ indices with $\alpha(\epsilon)$ infinitesimal as ϵ. Therefore there exist nonnegative functions $\alpha_i(\epsilon)$, $i = 1, 2, 3, 4$, such that $\alpha_i(\epsilon)$ goes to zero as ϵ goes to zero and such that

f1. $|(D_{n,a} - D_{n,c})_{i,i}| \leqslant \epsilon$ except, at most, $\alpha_1(\epsilon)n$ positions where the distance is bounded.
f2. $|(D_{n,a}^{-1/2} - D_{n,c}^{-1/2})_{i,i}| \leqslant \epsilon$ except, at most, $\alpha_2(\epsilon)n$ positions.
f3. $|(A_n(a-c))_{i,j}| \leqslant \epsilon$ except, at most, $\alpha_3(\epsilon)n$ positions where the considered quantity is bounded.
f4. $(D_{n,a}^{-1/2})_{i,i}$ and $(D_{n,c}^{-1/2})_{i,i}$ bounded by a fixed constant, except for, asymptotically, $\alpha_4(\epsilon)n$ positions (due to the fact that a is sparsely vanishing and c is chosen so that $m\{x \in [0,1] : a \neq c\} < \epsilon$).

In other words, by taking into account the crucial information that all the involved matrices are banded, we deduce that

$$A_n^*(a) = A_n^*(c) + LN(n) + LR(n), \tag{25}$$

where $\|LN(n)\|_{S,\infty} \leqslant \epsilon$ and $\operatorname{rank}(LR(n)) \leqslant h(\epsilon)n$ with $\lim_{\epsilon\to 0} h(\epsilon) = 0$.

From this, setting $D_n = DD_n + LR(n)$ and owing to the bandedness of all the involved matrices, we find that $\|A_n^*(a) - \Delta_{2k}^{(n)} - D_n\|_{S,p}^p = o(n)$ and $\|A_n^*(a) - \Delta_{2k}^{(n)} - D_n\|_{S,\infty} = o(1)$. To conclude, recall that $P_n^{-1}A_n(a)$ is similar to $[\Delta_{2k}^{(n)}]^{-1}A_n^*(a)$. Moreover, in the light of Theorem 5.3 the matrix $\Delta_{2k}^{(n)}$ has positive eigenvalues that can be seen, roughly speaking, as a sampling of the continuous sparsely vanishing function $2^{2k}\sin^{2k}(x/2)$ (compare (22)) over an equispaced mesh. The weak cluster of the singular values of $Z_n(a) = P_n^{-1}A_n(a)$ now follows from Lemma 4.4. Since $Z_n(a)$ is symmetrizable and all its eigenvalues are positive, we infer the weak cluster of the eigenvalues at 1 of the eigenvalues of $Z_n(a)$ (see Remark 4.7).

Finally, if a is not sparsely vanishing, then $A_n(a)$ and $P_n(a)$ may fail to be invertible (for instance if a is identically zero in an interval $[s,t]$, $s < t$) and the algebraic problem and the differential one may fail to have solution. □

The interesting fact in the proof of the preceding result is that it can be generalized in a very natural and simple way even in the multidimensional case.

Theorem 6.4. *Let $A_n^*(a)$ be the symmetrical scaling of the matrix related to problem* (2) *as in Lemma* 5.1 *and with Dirichlet boundary conditions. Let $\Delta_{2k;d}^{(n)}$ be the Toeplitz matrix defined as $A_n(1)$. Here each coefficient $a(\mathbf{x}_i)$ should be replaced by the mean value $N(n+e)\int_{I_i} a(\mathbf{t})\mathrm{d}\mathbf{t}$.*

- *If $a \in L^\infty$ is nonnegative and sparsely vanishing, then we find that there exist matrices D_n having $o(N(n))$ rank such that*

$$\left\|A_n^*(a) - \Delta_{2k;d}^{(n)} - D_n\right\|_{S,p}^p = o(N(n))$$

and

$$\left\|A_n^*(a) - \Delta_{2k;d}^{(n)} - D_n\right\|_{S,\infty} = o(1).$$

In addition, the number of outliers of $\{P_n^{-1}A_n(a)\}_n$ is generically $o(N(n))$.

- *If a is not sparsely vanishing, then $A_n(a)$ and $P_n(a)$ may fail to be invertible or $\{P_n^{-1}A_n(a)\}_n$ may not have clustered eigenvalues.*

Remark 6.4. If a is L^1, then the Lusin approximation c_ϵ of a is replaced by a non-negative polynomial being an L^1 approximation of a (this is possible since $[0,1]^d$ has finite measure). Now the proof of Theorem 6.3 is substantially unchanged except for the item [f1] which is replaced by "[f1*] $|(D_{n,a}-D_{n,c})_{i,i}|\leqslant\epsilon$ except, at most, $\alpha_1(\epsilon)N(n)$ positions (where the distance is not necessarily bounded)" and for the item [f3] which is replaced by "[f3*] $|(A_n(a-c))_{i,j}|\leqslant\epsilon$ except, at most, $\alpha_3(\epsilon)N(n)$ positions (where the considered quantity is not necessarily bounded)". We notice that this slight change does not spoil the proof because in the expression of $A_n^*(a)$ in Eq. (25) we have to add two other terms of "small" rank.

Finally, if $a\in L^1_{\mathrm{loc}}(\Omega)$ that is the restriction of a to any compact set $K\subset\Omega$ (Ω is open) belongs to L^1, then the proof and the statements of Theorems 6.3 and 6.4 still work with a bit different definition of the coefficient matrix $A_n(a)$. More precisely the symbol $a(\mathbf{x}_j)$ will denote

$$N(n+e)\int_{I_j}a(\mathbf{t})\mathrm{d}\mathbf{t},\quad I_j=\prod_{i=1}^{d}[x_{j_i},x_{j_i+1}],\ \ x_{j_i+1}-x_{j_i}=h_i,\ \ x_{j_i}=h_i j_i,$$

if $I_j\cap\partial\Omega=\emptyset$ and is 1 otherwise. In fact, we point out that the set I_j is a compact set and is contained in Ω if $I_j\cap\partial\Omega=\emptyset$ so that the integral appearing above makes sense.

Remark 6.5. By Theorems 6.3 and 6.4 and by using the same arguments as in Remarks 6.1 and 6.3, we deduce that $\{A_n^*(a)\}_n$ and $\{\Delta_{2k;d}^{(n)}\}_n$ are ED, EL and ϵ-EL.

6.1. Some computational remarks

In a sequential model of computation, system (3) can be solved directly and with an optimal cost by using very classic band solvers [18]. Here the optimality is with respect to the dimension n because these methods require $\mathrm{O}(n)$ arithmetic operations and the matrices $A_n(a)$ are defined by $\mathrm{O}(n)$ parameters. However, if we consider the dependence on the bandwidth in the asymptotic cost, that is, the dependence on k, then we remark that the Golub band solvers [18] based on the Gaussian elimination have a quadratic cost with respect to k. So, by taking into account the parameter k, there exist methods which are much more convenient. More specifically, we approximated $\{A_n(a)\}_n$ by the matrix sequence $\{P_n\}_n$ defined in (4). Therefore, up to the operations involving the diagonal matrices $D_{a,n}$ the computation is reduced to a band Toeplitz computation. The following very fast methods can be applied to banded Toeplitz structures:

- multigrid methods requiring $O(nk)$ ops and $O(\log n)$ parallel steps with $O(nk)$ processors in the parallel PRAM model [22] of computation [14,15] (linear dependence on k);
- a recursive displacement-rank [20]-based technique requiring $O(n \log k + k \log^2 (k) \log(n))$ ops (logarithmic dependence on k in the $O(n)$ term) and $O(\log n)$ parallel steps with $O(nk)$ processors [6].

Clearly, in order to obtain the total computational cost, the quoted costs have to be multiplied by the number of iterations which is constant with respect to n at least when a is positive (see Theorem 5.1), and added to the cost of few matrix–vector multiplications (recall the PCG algorithm [1]). The overall cost is of $O(n \log k)$ ops and $O(\log n)$ steps with $O(nk)$ processors in the PRAM model of computation.

Moreover, with special attention to the parallel model of computation, in [17,25] two matrix-algebra parallel strategies have been proposed. The first two strategies are based on the possibility of expressing these Toeplitz matrices as low-rank corrections of matrices belonging to some matrix algebras, such as the circulant class C_n [12] and the τ_n class [5]. These decompositions suggest the use of the Sherman–Morrison–Woodbury [18] (SMW) formula to obtain an efficient computation of the solution of the considered system ($O(\log n + \log^2 k)$ parallel steps with $O(n + k^3)$ processors).

In conclusion, in [25] and when a is regular enough, we have reduced the asymptotic cost of these band systems to the cost of the band-Toeplitz systems for which the recent literature provides very sophisticated algorithms [6,14,15]. Here we extended this result to the case where a is not smooth.

When we consider 2D differential operators such as

$$(-)^p \frac{\partial^p}{\partial x^p}\left(a(x,y)\frac{\partial^p}{\partial x^p}\right) + (-)^p \frac{\partial^p}{\partial y^p}\left(a(x,y)\frac{\partial^p}{\partial y^p}\right), \tag{26}$$

we construct the preconditioner the same way as in the scalar case (see Section 2.2).

In general, the matrix $\Delta_{2p;2}^{(n)}$ discretizing operator (26) is a double-banded matrix with external bandwidth $2p+1$ (p is the order of the operator with respect to y) and internal bandwidth $2p+1$ (p is the order of the operator with regard to x) and its generating function is nonnegative and has only one zero in $(x, y) = (0, 0)$ [24]. Therefore, as in the scalar case we can "correct" [17] $\Delta_{2p;2}^{(n)}$ in two different block matrix algebras: namely the block circulant class $C_{n,n}$ and the block τ algebra $\tau_{n,n}$ [5].

Consequently, by using the SMW formula and recalling that the computational cost of a bidimensional discrete Fourier or sine transform is $O(n^2 \log n)$ arithmetic operations and $O(\log n)$ parallel steps, we have to perform $O(\log n)+O(\log^2 n)$ parallel steps where the term $O(\log^2 n)$ is due to the inversion of the "smaller" matrix in the SMW formula.

As observed in [17], the best idea in the block case is the use of an algebraic multigrid method; in [15] it is shown that, in practice, the cost of the solution of $\Delta_{2p;2}^{(n)}\mathbf{x} = \mathbf{b}$ is $O(n^2)$ arithmetic operations and $O(\log n)$ parallel steps (a formal proof of convergence within a constant number of iterations can be found in [33]).

7. Numerical experiments

We restrict our attention to the irregular case for the presentation of the numerical experiments since the case where the coefficient a is continuous (or piecewise continuous) has been tested extensively in [25,27,30].

We consider problems in d dimensions with $d \leqslant 2$. For $d = 1$ the choice of the coefficient a varies among the following:

1. $a(x) \equiv a_1(x) = \lceil x^{-1/2} \rceil$ for $x \in (0, 1]$ and 1 for $x = 0$;
2. $a(x) \equiv a_2(x) = \lceil x^{-1} \rceil / \left(1 + \lceil x \rceil^{-1}\right)$ for $x \in (0, 1]$ and 1 for $x = 0$;
3. $a(x) \equiv a_3(x) = x a_2(x)$.

All these functions have an infinite but countable number of discontinuity points. The first belongs to $L^1 \backslash L^\infty$, while a_2 and a_3 belong to L^∞. The functions a_1 and a_2 are essentially positive while a_3 essentially vanishes at $x = 0$ and is sparsely vanishing. In Table 2, we report the number of PCG iterations, where $n \in \{500 + j100 : j = 0, \ldots, 5\}$. The test functions a are listed in the first column and the preconditioners are given in the heading. We show the number of PCG iterations in each row when the data vector is made up by all ones.

In Table 3, we give the number of outliers with respect to a cluster at 1 with radius 0.1, namely, we count the number N_o of eigenvalues of $P_n^{-1} A_n(a)$ for $n = 150, 300, 600$ not belonging to $(0.9, 1.1)$. The number N_o is written as $N_o(+) + N_o(-)$, where $N_o(-)$ counts those outliers less than 0.9 and $N_o(+)$ counts those outliers bigger than 1.1.

Some remarks are needed:

- Concerning Table 2, we observe that the number of PCG iterations is constant when the preconditioner is $\Delta_{2k;d}^{(n)}$ or P_n and the functional coefficient a is strictly positive and bounded. This independence with regard to n fully agrees with the spectral clustering theorems proved in this paper and with the spectral analysis of $\{(\Delta_{2k;d}^{(n)})^{-1} A_n(a)\}_n$ given in [27]. Notice that the simple preconditioner $D_{n,a}$ is never good since $\{(D_{n,a})^{-1} A_n(a)\}_n$ distributes as $\{\Delta_{2k;d}^{(n)}\}_n$: to see this notice $(D_{n,a})^{-1} A_n(a)$ is similar $A_n^*(a)$ and then refer to Remark 6.1. Now for any neighborhood $I_\epsilon = (0, \epsilon)$ with $\epsilon > 0$ we observe that $\{\Delta_{2k;d}^{(n)}\}_n$ shows $O(\sqrt{\epsilon} n)$ eigenvalues belonging to I_ϵ (see Theorem 5.3 and the subsequent linear algebra proof). In the light of the convergence analysis reported in [1], we know that the number of iterations is substantially equal to the dimension of the matrix and this is evident from Table 2.
- When zero belongs to the essential range of a or a is unbounded, it is immediate to observe that the only working preconditioner is P_n. Also this result agrees with the theoretical expectations of this paper. In this case, as shown in Table 3, the number of outlying eigenvalues grows very slowly (only logarithmically as n) and this behavior is much better when compared with the theoretical results. Regarding the preconditioner $\Delta_{2k;d}^{(n)}$, it is worthwhile observing that the case where a has zeros

Table 2
PCG iterations: case $n \in [500, 1000]$, $d = 1$

$a(x)$	D	Δ	P
$a_1(x)$, $n = 500$	500	17	11
$n = 600$	600	18	12
$n = 700$	700	19	13
$n = 800$	800	20	14
$n = 900$	900	20	14
$n = 1000$	1000	21	14
$a_2(x)$, $n = 500$	500	9	9
$n = 600$	600	9	9
$n = 700$	700	9	9
$n = 800$	800	9	10
$n = 900$	900	9	10
$n = 1000$	1000	10	10
$a_3(x)$, $n = 500$	500	115	10
$n = 600$	600	126	11
$n = 700$	700	136	11
$n = 800$	800	146	11
$n = 900$	900	155	12
$n = 1000$	1000	164	12

Table 3
Outliers: case $n = 150, 300, 600$, $d = 1$, P_n

$a(x)$	$n = 150$	$n = 300$	$n = 600$
$a_1(x)$	$4 + 2$	$5 + 3$	$6 + 4$
$a_2(x)$	$2 + 1$	$2 + 2$	$3 + 2$
$a_3(x)$	$1 + 3$	$2 + 4$	$2 + 5$

is much worse when compared to the case of a unbounded: this is in accordance with the analysis of Axelsson and Lindskog [1] that showed that "small" outliers slow down the convergence much more than "big" outliers.

- If a is essentially positive, then the presence of a countable (infinite) number of jumps of a does not spoil the performances of the associated PCG methods when $\Delta_{2k;d}^{(n)}$ or P_n are used as preconditioners according to the results of Theorems 5.1 and 6.3. Finally, notice the similarity of these results with respect to the case where a is smooth [25].

For $d = 2$ the choice of a is the following:

1. $a(x, y) \equiv a_1(x, y) = a_1(x) + a_1(y)$;
2. $a(x, y) \equiv a_2(x, y) = a_2(x) + a_2(y)$;

Table 4
PCG iterations: case $n = 10^2, 20^2, 30^2$, $d = 2$

$a(x, y)$		D	Δ	P
$a_1(x, y)$,	$n = 10^2$	27	9	6
	$n = 20^2$	53	12	7
	$n = 30^2$	79	13	8
$a_2(x, y)$,	$n = 10^2$	27	7	5
	$n = 20^2$	53	8	6
	$n = 30^2$	80	8	5
$a_3(x, y)$,	$n = 10^2$	29	18	5
	$n = 20^2$	57	26	6
	$n = 30^2$	88	32	6
$a_4(x, y)$,	$n = 10^2$	30	12	6
	$n = 20^2$	59	15	8
	$n = 30^2$	95	17	9

Table 5
Outliers: case $n = 10^2, 20^2, 30^2$, $d = 2$, P_n

$a(x, y)$	$n = 10^2$	$n = 20^2$	$n = 30^2$
$a_1(x, y)$	1	$8 + 2$	$14 + 1$
$a_2(x, y)$	0	3	0
$a_3(x, y)$	0	1	0
$a_4(x, y)$	4	$10 + 4$	$15 + 3$

3. $a(x, y) \equiv a_3(x, y) = (x + y)a_2(x, y)$;
4. $a(x, y) \equiv a_4(x, y) = \exp(a_2(x))a_1(y) + y$.

All these functions have an infinite but countable number of discontinuity points. The first and the fourth belong to $L^1 \backslash L^\infty$, while a_2 and a_3 belong to L^∞. The functions a_1, a_2 and a_4 are essentially positive while a_3 essentially vanishes at $(x, y) = (0, 0)$ and is sparsely vanishing. In Table 4, we report the number of PCG iterations, where $n \in \{(j10)^2 : j = 1, 2, 3\}$. The test functions a are listed in the first column and the preconditioners are given in the heading. In each row, we show the number of PCG iterations when the data vector is made up by all ones.

In Table 5, we give the number of outliers with respect to a cluster at 1 with radius 0.1, namely, we count the number of eigenvalues of $P_n^{-1}A_n(a)$ for $n = 100, 400, 900$ not belonging to $(0.9, 1.1)$. The number N_o is written as $N_o(+) + N_o(-)$, where $N_o(-)$ and $N_o(+)$ have the same meaning as before.

We remark that the behavior of the PCG method in the two-dimensional case is not substantially different from the unidimensional case.

- Concerning Table 4, we observe that the number of PCG iterations is constant when the preconditioner is $\Delta_{2k;d}^{(n)}$ or P_n and the functional coefficient a is strictly positive and bounded that is for $a = a_3(x, y)$.
- When zero belongs to the essential range of a or a is unbounded, the only working preconditioner is P_n. In this case, as shown in Table 5, the number of outlying eigenvalues grows very slowly (only logarithmically as n) and this behavior is much better when compared with the theoretical results.
- If a is essentially positive, then the presence of a countable (infinite) number of jumps of a does not spoil the performances of the associated PCG methods when $\Delta_{2k;d}^{(n)}$ or P_n are used as preconditioners according to the results of Theorems 5.1 and 6.4. Finally, notice the similarity of these results with respect to the case where a is smooth [25].

8. Conclusive remarks

To conclude, in this paper, we have introduced new tools in order to study the spectral behavior of matrix-sequences. As a case study we have discussed the asymptotical distributional properties of the spectra of Toeplitz-based preconditioned matrix-sequences under the assumptions that the functional coefficient a is not regular and the differential problems are of the form (1) or (2). We have proved that the general clustering of the spectra still holds in the irregular and multilevel case. Moreover, the results indicate that a possible deterioration of the convergence properties of the associated PCG methods occurs when the function a is not strictly positive.

Acknowledgements

Hearty thanks to Paolo Tilli and to Eugene Tyrtyshnikov for their nice discussions.

References

[1] O. Axelsson, G. Lindskog, On the rate of convergence of the preconditioned conjugate gradient method, Numer. Math. 52 (1986) 499–523.
[2] O. Axelsson, M. Neytcheva, The algebraic multilevel iteration methods—theory and applications, in: D. Bainov (Ed.), Proceedings of the 2nd International Colloquium on Numerical Analysis, Plovdiv, Bulgaria, August 1993, pp. 13–23.
[3] R. Bhatia, Matrix Analysis, Springer, New York, 1997.
[5] D. Bini, M. Capovani, Spectral and computational properties of band symmetric matrices, Linear Algebra Appl 52 (1983) 99–125.
[6] D. Bini, B. Meini, Effective methods for solving banded Toeplitz systems, SIAM J. Matrix Anal. Appl. 20 (3) (1999) 700–719.

[7] A. Böttcher, S. Grudsky, On the condition numbers of large semi-definite Toeplitz matrices, Linear Algebra Appl. 279 (1998) 285–301.
[8] R.H. Chan, T.F. Chan, Circulant preconditioners for elliptic problems, J. Numer. Linear Algebra Appl. 1 (1992) 77–101.
[9] R.H. Chan, M. Ng, Conjugate gradient methods for Toeplitz systems, SIAM Rev. 38 (1996) 427–482.
[10] T.F. Chan, An optimal circulant preconditioner for Toeplitz systems, SIAM J. Sci. Statist. Comput. 9 (1988) 766–771.
[11] P. Ciarlet, The Finite Element Method for Elliptic Problems, North-Holland, Amsterdam, 1978.
[12] H. Davis, Circulant Matrices, Wiley, New York, 1979.
[13] F. Di Benedetto, S. Serra Capizzano, A unifying approach to matrix algebra preconditioning, Numer. Math. 82 (1) (1999) 57–90.
[14] G. Fiorentino, S. Serra, Multigrid methods for Toeplitz matrices, Calcolo 28 (1991) 283–305.
[15] G. Fiorentino, S. Serra, Multigrid methods for symmetric positive definite block Toeplitz matrices with nonnegative generating functions, SIAM J. Sci. Comput. 17 (1996) 1068–1081.
[16] G. Fiorentino, S. Serra, Tau preconditioners for (high order) elliptic problems, in: P. Vassilevski (Ed.), Proceedings of the 2nd IMACS Conference on Iterative Methods in Linear Algebra, Blagoevgrad, Bulgaria, June 1995, pp. 342–353.
[17] G. Fiorentino, S. Serra, Fast parallel solvers for elliptic problems, Comput. Math. Appl. 32 (1996) 61–68.
[18] G. Golub, C. Van Loan, Matrix Computations, The Johns Hopkins University Press, Baltimore, 1983.
[19] U. Grenander, G. Szegö, Toeplitz Forms and Their Applications, second ed., Chelsea, New York, 1984.
[20] T. Kailath, S. Kung, M. Morf, Displacement ranks of matrices and linear equations, J. Math. Anal. Appl. 68 (1979) 395–407.
[21] I. Lirkov, S. Margenov, P. Vassilevsky, Circulant block factorization for elliptic problems, Computing 53 (1994) 59–74.
[22] M.J. Quinn, Parallel Computing: Theory and Practice, McGraw-Hill, New York, 1994.
[23] W. Rudin, Real and Complex Analysis, McGraw-Hill, Singapore, 1986.
[24] S. Serra, Preconditioning strategies for asymptotically ill-conditioned block Toeplitz systems, BIT 34 (1994) 579–594.
[25] S. Serra, The rate of convergence of Toeplitz based PCG methods for second order nonlinear boundary value problems, Numer. Math. 81 (3) (1999) 461–495.
[26] S. Serra, On the extreme eigenvalues of Hermitian (block) Toeplitz matrices, Linear Algebra Appl 270 (1998) 109–129.
[27] S. Serra Capizzano, C. Tablino Possio, Spectral and structural analysis of high precision finite difference matrices for elliptic operators, Linear Algebra Appl. 293 (1999) 85–131.
[28] S. Serra, Optimal quasi-optimal and superlinear band-Toeplitz preconditioners for asymptotically ill-conditioned positive definite Toeplitz systems, Math. Comp. 66 (1997) 651–665.
[29] S. Serra, A Korovkin-type Theory for finite Toeplitz operators via matrix algebras, Numer. Math. 82 (1) (1999) 117–142.
[30] S. Serra Capizzano, C. Tablino Possio, High-precision finite difference schemes and Toeplitz based preconditioners for elliptic problems, Electr. Trans. Numer. Anal. 11 (2000) 55–84.
[31] S. Serra Capizzano, E. Tyrtyshnikov, Multilevel Toeplitz matrices and approximation by matrix algebras, in: F. Luk (Ed.), Proceedings in Advanced Signal Processing Algorithms, Architectures, and Implementations VIII—SPIE Conference, San Diego, CA, July 1998, pp. 393–404.
[32] S. Serra Capizzano, E. Tyrtyshnikov, Any circulant-like preconditioner for multilevel matrices is not superlinear, SIAM J. Matrix Anal. Appl. 21 (2) (1999) 431–439.
[33] H. Sun, X. Jin, Q. Chang, Multigrid scheme for ill-conditioned block Toeplitz linear systems, TR no. 12, Department of Mathematics, Chinese University of Hong Kong, 1998.

[34] E. Tyrtyshnikov, Circulant preconditioners with unbounded inverses, Linear Algebra Appl. 216 (1995) 1–23.
[35] E. Tyrtyshnikov, A unifying approach to some old and new theorems on distribution and clustering, Linear Algebra Appl. 232 (1996) 1–43.
[36] E. Tyrtyshnikov, N. Zamarashkin, Spectra of multilevel Toeplitz matrices: advanced theory via simple matrix relationships, Linear Algebra Appl. 270 (1998) 15–27.
[37] E. Tyrtyshnikov, N. Zamarashkin, Thin structure of eigenvalue clusters for non-Hermitian matrices, Linear Algebra Appl. 292 (1999) 297–310.
[38] R. Varga, Matrix Iterative Analysis, Prentice-Hall, Englewood Cliffs, NJ, 1962.
[39] H. Widom, On the singular values of Toeplitz matrices, Zeit. Anal. Anw. 8 (1989) 221–229.