Electronic Transactions on Numerical Analysis.
Volume 11, pp. 55-84, 2000.

ISSN 1068-9613.

ETNA
Kent State University
etna@mcs.kent.edu

HIGH-ORDER FINITE DIFFERENCE SCHEMES AND TOEPLITZ BASED PRECONDITIONERS FOR ELLIPTIC PROBLEMS*

STEFANO SERRA CAPIZZANO† AND CRISTINA TABLINO POSSIO‡

Abstract. In this paper we are concerned with the spectral analysis of the sequence of preconditioned matrices

$$\{P_n^{-1}(a,m,k)A_n(a,m,k)\}_n,$$

where $A_n(a,m,k)$ is the $n\times n$ symmetric matrix coming from a high–order Finite Difference discretization of the problem

$$\begin{cases} (-)^k\left(\dfrac{d^k}{dx^k}\left(a(x)\dfrac{d^k}{dx^k}u(x)\right)\right) = f(x) \quad \text{on } \Omega=(0,1), \\ \left(\dfrac{d^s}{dx^s}u(x)\right)_{|\partial\Omega} = 0 \quad s=0,\dots,k-1. \end{cases}$$

The coefficient function $a(x)$ is assumed to be positive or with a finite number of zeros. The matrix $P_n(a,m,k)$ is a Toeplitz based preconditioner constructed as $D_n^{1/2}(a,m,k)A_n(1,m,k)D_n^{1/2}(a,m,k)$, where $D_n(a,m,k)$ is the suitably scaled diagonal part of $A_n(a,m,k)$. The main result is the proof of the asymptotic clustering around unity of the eigenvalues of the preconditioned matrices. In addition, the "strength" of the cluster shows some interesting dependencies on the order k, on the regularity features of $a(x)$ and on the presence of the zeros of $a(x)$. The multidimensional case is analyzed in depth in a twin paper [38].

Key words. finite differences, Toeplitz and Vandermonde matrices, clustering and preconditioning, ergodic theorems, spectral distribution.

AMS subject classifications. 65N22, 65F10, 15A12.

1. Introduction.

The numerical solution of elliptic boundary value problems is a classical topic arising from a wide range of applications such as elasticity problems and nuclear and petroleum engineering [44]. In these contexts, the coefficient function $a(x)$ can be continuous or discontinuous, but being strictly positive the ellipticity of the continuous problem is therefore guaranteed. On the other hand, for the calculation of special functions or for applications to mathematical biology and mathematical finance, strict ellipticity is lost and indeed the function $a(x)$ may have isolated zeros generally located at the boundary of the definition domain. Therefore in the continuous problem, we simply assume that $a(x)\geq 0$ (with, at most, a finite number of zeros).

In preceding works [15, 16, 31, 34], we have considered these types of problems by focusing our attention on the Finite Differences (FD) of minimal order of accuracy. The resulting symmetric positive definite linear systems are solved by using preconditioned conjugate gradient (PCG) algorithms where the chosen preconditioners ensure the "optimality" of the method [2] and even a "clustering" [43] of the preconditioned spectra around unity [31].

In this paper we deal with high–order Finite Difference formulae for the approximation of the given elliptic (or semielliptic) differential problems. The motivation is given by the increased accuracy when the related solution is sufficiently regular. Nevertheless, the diminished sparsity of the resulting linear system has been considered an insurmountable obstacle

*Received June 10, 1999. Accepted for publication June 24, 2000. Recommended by L. Reichel.
†Dipartimento di Energetica "S. Stecco", Università di Firenze. Via Lombroso 6/17, 50134 Firenze, Italy. E-mail: serra@mail.dm.unipi.it
‡Dipartimento di Scienza dei materiali, Università di Milano Bicocca. Via Cozzi 53, 20126 Milano, Italy. E-mail: cristina.tablino.possio@mater.unimib.it

for the practical application of these methods. Here by looking at these structured linear systems from the point of view of their "locally Toeplitzness" [41], we arrive at different ways to overcome the difficulty represented by the diminished sparsity.

In the following, we study some Toeplitz based preconditioners for matrices $A_n(a, m, k)$ coming from a large class of high–order FD discretizations of continuous problems of the form

$$\left\{ \begin{array}{l} (-)^k \left(\dfrac{d^k}{dx^k} \left(a(x) \dfrac{d^k}{dx^k} u(x) \right) \right) = f(x) \quad \text{on } \Omega = (0,1), \\ \left(\dfrac{d^s}{dx^s} u(x) \right)_{|\partial\Omega} = 0 \quad s = 0, \ldots, k-1. \end{array} \right. \tag{1.1}$$

The $2D$ continuous problem

$$\left\{ \begin{array}{l} (-)^k \left(\dfrac{\partial^k}{\partial x^k} \left(a(x,y) \dfrac{\partial^k}{\partial x^k} u(x,y) \right) + \dfrac{\partial^k}{\partial y^k} \left(a(x,y) \dfrac{\partial^k}{\partial y^k} u(x,y) \right) \right) = f(x,y) \text{ on } \Omega, \\ \left(\dfrac{\partial^s}{\partial \nu^s} u(x,y) \right)_{|\partial\Omega} = 0 \quad s = 0, \ldots, k-1, \end{array} \right. \tag{1.2}$$

where $\Omega = (0,1)^2$ and ν denotes the unit outward normal direction, is also considered, but, for reasons of notational complexity, the related analysis is reported in a twin paper [38], where the same efficiency of the proposed preconditioning technique is proved.

More specifically, by setting $P_n(a, m, k) = D_n^{1/2}(a, m, k) A_n(1, m, k) D_n^{1/2}(a, m, k)$, where $D_n(a, m, k)$ is the suitably scaled diagonal part of $A_n(a, m, k)$ and $A_n(1, m, k)$ is the symmetric positive definite Toeplitz matrix obtained when $a(x) \equiv 1$, asymptotic expansions concerning the preconditioned matrices $P_n^{-1}(a, m, k) A_n(a, m, k)$ are derived.

The spectral (distributional) analysis of sequences of matrices

$$\{A_n^{-1}(b, m, k) A_n(a, m, k)\}_n, \quad b > 0,$$

has been performed in [36] by focusing the attention on Szegő–like and Widom–like ergodic results [19, 45]. Here we analyze in detail the sequence of matrices $\{P_n^{-1}(a, m, k) A_n(a, m, k)\}_n$ by showing that it provides a clustering at unity of the related spectra. Calling h the mesh size of the discretization, the main results can be summarized as follows. Let $a(x)$ be a strictly positive function.
If $a(x) \in C^2(\overline{\Omega})$ then

$$P_n^{-1}(a, m, k) A_n(a, m, k) \sim_{\mathrm{S}} I_n + A_n^{-1}(1, m, k)[h^2 \Theta_n(a, m, k) + o(h^2)].$$

If $a(x) \in C^1(\overline{\Omega})$ then

$$P_n^{-1}(a, m, k) A_n(a, m, k) \sim_{\mathrm{S}} I_n + A_n^{-1}(1, m, k)[O(h\omega_{a_x}(h)) R_n(a, m, k) + o(h\omega_{a_x}(h))].$$

If $a(x) \in C(\overline{\Omega})$ then

$$P_n^{-1}(a, m, k) A_n(a, m, k) \sim_{\mathrm{S}} I_n + A_n^{-1}(1, m, k)[O(\omega_a(h)) R_n(a, m, k) + o(\omega_a(h))].$$

Here $X \sim_{\mathrm{S}} Y$ means that X and Y are similar, $\omega_f(\cdot)$ denotes the modulus of continuity of the function f, $\Theta_n(a, m, k)$ and $R_n(a, m, k)$ are bounded matrices with the same pattern as $A_n(a, m, k)$ and I_n denotes the identity matrix.

When we have less regularity and/or $a(x)$ is "sparsely vanishing" (i.e. the Lebesgue measure of the set of the zeros of $a(x)$ is zero), then the expansion can be restated in the

ETNA
Kent State University
etna@mcs.kent.edu

following way: for any positive ε there exists a sequence $\{D_n(\varepsilon)\}_n$ such that we have $\mathrm{rank}(D_n(\varepsilon)) \leq \varepsilon n$ for n large enough and

$$P_n^{-1}(a,m,k)A_n(a,m,k) \sim_{\mathrm{S}} I_n + A_n^{-1}(1,m,k)[\Theta_n(a,m,k,\varepsilon) + D_n(\varepsilon)],$$

with $\lim_{h\to 0} \Theta_n(a,m,k,\varepsilon) = 0$.

All of these results are useful in order to understand the asymptotic behaviour of the PCG techniques when these Toeplitz based preconditioners are applied. Actually, by using the preceding expansions, we obtain an asymptotic estimate of the number of the eigenvalues of the preconditioned matrices which are not clustered at unity. Consequently, by virtue of the Axelsson and Lindskog Theorems [2], we deduce a sufficiently accurate upper bound on the number of PCG iterations that we need in order to reach the solution within a preassigned accuracy η. In this way the solution of a system with a coefficient matrix given by $A_n(a,m,k)$ is reduced to the solution of a few linear systems of diagonal and of band-Toeplitz types. The existence of very sophisticated numerical procedures for the computation of the solution of band-Toeplitz linear systems (see [13] and especially [6]) makes the proposed preconditioning techniques very attractive in the considered context of differential boundary value problems.

Indeed, both theoretical and practical comparisons prove that the new ideas are superior, especially in a multidimensional case or in a parallel model of computation, with respect to the classical techniques [1, 8, 22, 26, 27]. In particular the Strang circulant preconditioner can be singular [8, 17], while the T. Chan circulant preconditioner does not produce a strong cluster and is not optimal in the sense that the spectral condition number of the related preconditioned sequence goes to infinity as the size n goes to infinity [8]. Therefore, the given technique is optimal as the Gaussian elimination, but does not suffer from potential instabilities and accumulation of round-off errors which characterize direct methods applied to ill-conditioned linear systems; we recall that the spectral condition number of $A_n(a,m,k)$ grows at least as n^{2k} [36].

However, the motivation of the present analysis becomes much stronger in the multidimensional case [38] for at least three reasons: **1.** the Gaussian elimination is no longer optimal due to the "sparse" bandedness of the involved structures, **2.** any preconditioner chosen in a multilevel matrix algebra (circulants, trigonometric algebras etc.) cannot give a strong cluster at unity according to recent results of the first author and Tyrtyshnikov [39, 40], **3.** the incomplete LU factorization preconditioner has a linear cost per iteration, but the number of iterations cannot be bounded from above uniformly with respect to the dimension. Conversely, the multilevel version [38] of the ideas presented in this paper leads to preconditioning strategy having the following features:

1. a linear cost per iteration of the related PCG method,
2. a strong cluster at unity and a localization of the spectra in a positive interval independent of the dimension and bounded away from zero (at least when the coefficent $a(x)$ is positive and twice continuously differentiable).

Finally we mention that the theoretical results in the multidimensional context [38] are heavily based on the analysis reported in this paper.

The paper is organized as follows. In §2 we give preliminary results concerning the high–order FD matrices $A_n(a,m,k)$ and the Toeplitz matrices $A_n(1,m,k)$. Section 3 is devoted to the definition of the proposed Toeplitz based preconditioner and in §4 some numerical experiments are reported to show its practical effectiveness both in the $1D$ and $2D$ case. Section 5 is addressed to the theoretical clustering analysis of the preconditioned matrix sequences. In §6 we study the spectral distribution of the preconditioned matrices and we deal with the irregular case in which $a(x)$ is assumed $L^\infty(\Omega)$. In §7 we analyze the computational cost of the proposed PCG method and we make a direct comparison with the existing literature. Finally, some concluding remarks in §8 end the paper.

2. High–order FD matrices. In this section we summarize the main structural and spectral properties of the band matrices associated with high–order FD discretization of the continuous problem (1.1). Since these matrices reduce to Toeplitz structures in the case where $a(x)$ is a constant function, some crucial properties of Toeplitz structures are also considered.

It is worthwhile stressing that in our formulae we work with some extra points not belonging to $\Omega = (0, 1)$. So, for mathematical consistency, we need to define the coefficient function $a(x)$ over the set $\Omega^* = [-\tilde{\varepsilon}, 1 + \tilde{\varepsilon}]$, where $\tilde{\varepsilon}$ is some positive quantity. Therefore, when we write $a(x) \in C^s(\overline{\Omega})$ it is understood that $a(x)$ is simply defined in Ω^*, while the regularity is required in Ω. The only assumption we have to make is that for each $x \in \Omega^*$

$$\min_{y\in\overline{\Omega}} a(y) \le a(x) \le \max_{y\in\overline{\Omega}} a(y). \tag{2.1}$$

2.1. High–order FD formulae. In a previous paper [36] we highlighted some general features of high–order FD formulae for the discretization of the differential operator d^k/dx^k by using q equispaced mesh points. Here, we briefly report the essential notations and the key properties necessary to define and to analyze the arising FD matrices and the proposed Toeplitz based preconditioner.

As usual, we assume that the discretization of $(d^k u(x)/dx^k)_{|x=x_r}$ with q equispaced mesh points ($q \ge k+1$) involves $m = \lfloor q/2 \rfloor$ mesh points less than x_r, $m = \lfloor q/2 \rfloor$ greater than x_r, plus the point x_r if q is odd. More precisely, if $q = 2m+1$ the mesh points are defined as $x_j = x_r + jh$, $j = -m, \ldots, m$, while if $q = 2m$ as $x_j = x_r + (j - 1/2)h$, $j = 1, \ldots, m$ and $x_j = x_r + (j + 1/2)h$, $j = -m, \ldots, -1$.

Let $\mathbf{c} \in \mathbb{R}^q$ be the coefficient vector defining a FD formula that shows an order of accuracy ν under the assumption of a sufficient regularity of the function $u(x)$, i.e.

$$(d^k u(x)/dx^k)_{|x=x_r} = h^{-k} \sum_j c_j u(x_j) + O(h^\nu).$$

Such a coefficient vector can be obtained as the solution of a Vandermonde-like linear system [36]. As a consequence, the following statement holds true.

LEMMA 2.1. [36] *Let $\mathbf{c} \in \mathbb{R}^q$ be the coefficient vector defining the maximal order FD formula discretizing d^k/dx^k by using q ($q \ge k+1$) equispaced mesh points. Then $\mathbf{c}$ is unique and its entries are rational, $\mathbf{c}$ is symmetric or antisymmetric with respect to its middle according to whether the quantity $k(\mathrm{mod}\ 2)$ equals 0 or 1. Finally, the FD formula order of accuracy ν equals $q - k + 1$ if $k + q$ is odd and equals $q - k$ if $k + q$ is even.*

To obtain symmetric FD matrices, welcome from a computational point of view, we leave the operator in “divergence form” and we discretize the inner and the outer derivatives separately. Although the FD discretization process could be performed in a more general way (refer to [36]), here we limit ourselves to the case where both the inner and the outer operator are discretized by means of the FD formula of maximal order of accuracy with respect to q mesh points. It is worthwhile noticing that, owing to the comparison between the computational cost and the order of accuracy (see Lemma 2.1), we will always consider q odd when k is even and vice versa.

DEFINITION 2.2. [36] *The symbol $A_n(a, m, k)$, with $m = \lfloor q/2 \rfloor$, denotes the $n \times n$ symmetric band matrix discretizing the problem (1.1) through the maximal order FD formula with respect to q ($q \ge k+1$) equispaced mesh points related to the coefficient vector $\mathbf{c} \in \mathbb{R}^q$ for both the inner and the outer derivatives.*

Hereafter, the relations defining the nonzero lower triangular entries of the generic r^{th} rows of the symmetric band matrix $A_n(a,m,k)$ are reported.

Case k odd ($q=2m$): As a consequence of Lemma 2.1, we are dealing with an antisymmetric coefficient vector $\mathbf{c}=(-c_m,\,\dots,\,-c_1,\,c_1,\dots,c_m)$. Therefore, by Definition 2.2, it follows that

$$(A_n)_{r,r}=\sum_{j=1}^{m}\left(a\left(x_{r-j+1/2}\right)+a\left(x_{r+j-1/2}\right)\right)c_j^2, \tag{2.2}$$

$$(A_n)_{r,r-p}=\sum_{j=1}^{m-p}\left(a\left(x_{r-p-j+1/2}\right)+a\left(x_{r+j-1/2}\right)\right)c_jc_{j+p} \tag{2.3}$$

$$-\sum_{j=1}^{p}a\left(x_{r-j+1/2}\right)c_jc_{p+1-j},\qquad \text{if } p=1,\dots,m-1,$$

$$(A_n)_{r,r-p}=-\sum_{j=p+1-m}^{m}a\left(x_{r-j+1/2}\right)c_jc_{p+1-j},\qquad \text{if } p=m,\dots,2m-1. \tag{2.4}$$

Case k even ($q=2m+1$): As a consequence of Lemma 2.1, we are dealing with a symmetric coefficient vector $\mathbf{c}=(c_m,\,\dots,\,c_1,\,c_0,\,c_1,\dots,c_m)$. Therefore, by Definition 2.2, it follows that

$$(A_n)_{r,r}=a(x_r)c_0^2+\sum_{j=1}^{m}(a(x_{r-j})+a(x_{r+j}))c_j^2, \tag{2.5}$$

$$(A_n)_{r,r-p}=\sum_{j=1}^{m-p}(a(x_{r-p-j})+a(x_{r+j}))c_jc_{p+j} \tag{2.6}$$

$$+\sum_{j=0}^{p}a(x_{r-j})c_jc_{p-j},\qquad \text{if } p=1,\dots,m-1,$$

$$(A_n)_{r,r-p}=\sum_{j=p-m}^{m}a(x_{r-j})c_jc_{p-j},\qquad \text{if } p=m,\dots,2m. \tag{2.7}$$

2.2. The Spectral properties of $\{A_n(a,m,k)\}_n$. The spectral properties of the FD approximation of the continuous problem in (1.1) can be briefly summarized as follows.

THEOREM 2.3. [36] *Let $S^{n\times n}$ be the (real) linear space of the $n\times n$ symmetric matrices and let $C(\overline{\Omega})$ be the (real) linear space of the continuous functions on $\overline{\Omega}$. Let $\mathcal{G}_n=\{\tilde{x}_i\}$ be the set of equispaced samplings of the coefficient function $a(x)$. Let $\mathbf{c}[i]$ be a vector of $\mathbb{R}^n$ containing in "its middle" the vector $\mathbf{c}$, defining the maximal order FD formula, suitably shifted in accordance with i. The matrices $A_n(a,m,k)$ can be expressed as $A_n(a,m,k)=\sum_i a(\tilde{x}_i)Q_n(\mathbf{c},\mathbf{c},i)$, with $Q_n(\mathbf{c},\mathbf{c},i)=\mathbf{c}[i]\mathbf{c}^T[i]$ being a symmetric nonnegative definite dyad. As a consequence, for any n,k and m, the operator $A_n(\cdot,m,k):C(\overline{\Omega})\to S^{n\times n}$ is linear and positive, i.e. if $a(x)$ is a nonnegative function then $A_n(a,m,k)$ is a nonnegative definite matrix. In addition, the operator is normally positive, i.e. if $a(x)$ is a strictly positive function then $A_n(a,m,k)$ is a positive definite matrix.*

THEOREM 2.4. [36] *Let $\mathcal{G}_n=\{\tilde{x}_i\}$ be the set of equispaced samplings of the coefficient function $a(x)$ and let $I^+(a)=\{i\,:\,a(\tilde{x}_i)>0\}$. Suppose that the $n+q-1$ vectors $\{\mathbf{c}[i]\in\mathbb{R}^n:\,i=1,\dots,n+q-1\}$ strongly generate $\mathbb{R}^n$, i.e. each subset $\{\mathbf{c}[i_k]:\,1\le i_1<i_2\cdots<i_n\le n+q-1\}$ is a basis for $\mathbb{R}^n$. Then $\operatorname{rank}(A_n(a,m,k))=\min\{n,\#I^+(a)\}$.*

ETNA
Kent State University
etna@mcs.kent.edu

Lastly, the following results pertain to the spectral condition number $k_2(\cdot)$ of the considered high–order FD matrices.

THEOREM 2.5. [36] *If the coefficient function $a(x)$ is strictly positive then $k_2(A_n(a, m, k)) \sim k_2(A_n(1, m, k))$, while if $a(x) \geq 0$ then $k_2(A_n(a, m, k)) \geq cn^{2k}$, with $c > 0$ and for n large enough. Here we write $f \sim g$ over the interval I if f and g are nonnegative over I and there exist two positive universal constants c_1 and c_2 so that $c_1 g \leq f \leq c_2 g$ almost everywhere in I.*

2.3. The Toeplitz case $\{\Delta_n(m, k)\}_n$. When the coefficient function $a(x) \equiv 1$, the matrices $A_n(a, m, k)$ enjoy the Toeplitz structure. Hereafter, each Toeplitz matrix $A_n(1, m, k)$ will be denoted by $\Delta_n(m, k)$.
A key role is played by Toeplitz matrices generated by 2π–periodic integrable functions f defined on the fundamental interval $[-\pi, \pi]$, where the entry along the $k^{\rm th}$ diagonal is given by the $k^{\rm th}$ Fourier coefficient of f, i.e.

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

Clearly, if f is real–valued, the quoted definition implies that $a_{-k} = \overline{a_k}$ so that the Toeplitz matrices are Hermitian for any value of the dimension n.

THEOREM 2.6. [36] *The matrices $\Delta_n(m, k)$ are Toeplitz matrices generated by the non-negative real–valued polynomial $p_{\mathbf{w}}(x) = |p_{\mathbf{c}}(x)|^2$, $x \in [-\pi, \pi]$, where $p_{\mathbf{c}}$ is the polynomial related to the maximal order FD formula coefficient vector $\mathbf{c}$ for the discretization of the operator d^k/dx^k, defined as*

$$p_{\mathbf{c}}(x) = \begin{cases} \sum_{j=-m}^{m} c_j e^{\mathbf{i}jx}, & \text{if } q = 2m+1, \\ \sum_{j=-m}^{-1} c_j e^{\mathbf{i}jx} + \sum_{j=1}^{m} c_j e^{\mathbf{i}(j-1)x}, & \text{if } q = 2m. \end{cases} \tag{2.8}$$

Therefore, the matrices $\Delta_n(m, k)$ are symmetric positive definite for any value of the dimension n and the related spectral condition number is asymptotically greater than cn^{2k}, with c a positive universal constant.

Lastly, it can be easily verified [34] that the polynomial $p_{\mathbf{w}}(x) = w_0 + 2\sum_{j=1}^{q-1} w_j \cos(jx)$, with degree $q - 1 \geq k$ with respect to the cosine expansion, can be written as

$$p_{\mathbf{w}}(x) = \begin{cases} 2^{2k} \sin^{2k}(x/2), & \text{if } q = k+1, \\ 2^{2k} \sin^{2k}(x/2) \cdot g^{q,k}(x), & \text{if } q > k+1, \end{cases} \tag{2.9}$$

where $g^{q,k}(0) > 0$; that is, $x = 0$ is always a zero of exactly order $2k$. Therefore, we deduce that the maximal order of the zeros of $p_{\mathbf{w}}(x)$ is $2s = 2k$ for $2k \geq q - 1$ and is $2s \in [2k, 2(q-1) - 2k]$ for $2k < q - 1$. This property is used for analyzing the clustering properties of the proposed Toeplitz based preconditioning matrix sequence (refer to §5).

3. The Toeplitz based preconditioners. Our goal with respect to the preconditioning problem in the conjugate gradient method is stated in the following definition.

DEFINITION 3.1. [1] *Let $\{A_n\}_n$ and $\{P_n\}_n$ be two sequences of $n \times n$ positive definite Hermitian matrices. The sequence $\{P_n\}_n$ is an optimal sequence of preconditioners for the sequence $\{A_n\}_n$ if for any n all the eigenvalues of $P_n^{-1} A_n$ belong to a bounded interval $[d_1, d_2]$, with d_i positive universal constants independent of n.*

The same attention must be paid to the clustering properties [42] of the preconditioned matrix sequence $\{P_n^{-1} A_n\}_n$, where $A_n, P_n \in \mathbb{C}^{n \times n}$. Clearly, if $\{A_n\}_n$ and $\{P_n\}_n$ are

Hermitian and $\{P_n\}_n$ are positive definite, then the following clustering properties are in the sense of the eigenvalues.

Let us denote by $||\cdot||_F$ the Frobenius norm, i.e. the Euclidean norm of the n-dimensional vector formed by the n singular values (refer to [4, Bhatia, p. 92]). The following propositions can be easily obtained as a consequence of Tyrtyshnikov's results [43].

PROPOSITION 3.2. [Strong Clustering Property] *If there exists a sequence $\{D_n\}_n$, where $D_n \in \mathbb{C}^{n\times n}$, so that $\|A_n - P_n - D_n\|_F^2 = O(1)$, with rank$(D_n) = O(1)$ and the minimal singular value of P_n is greater than a fixed constant $\delta > 0$, then for any $\varepsilon > 0$ all the singular values of $P_n^{-1}(A_n - P_n)$ belong to $[0, \varepsilon)$ except for $N_o(\varepsilon, n) = O(1)$ outliers.*

PROPOSITION 3.3. [Weak Clustering Property] *If there exists a sequence $\{D_n\}_n$, where $D_n \in \mathbb{C}^{n\times n}$, so that $\|A_n - P_n - D_n\|_F^2 = o(n)$, with rank$(D_n) = o(n)$ and the minimal singular value of P_n is greater than a fixed constant $\delta > 0$, then for any $\varepsilon > 0$ all the singular values of $P_n^{-1}(A_n - P_n)$ belong to $[0, \varepsilon)$ except for $N_o(\varepsilon, n) = o(n)$ outliers.*

Separate mention has to be made of the following limit case.

DEFINITION 3.4. [Weakest Strong Clustering Property] *If the matrix sequence $\{P_n^{-1}(A_n - P_n)\}_n$ satisfies the Strong Clustering Property, but the quantity $N_o(\varepsilon, n)$ goes to infinity as n goes to infinity and ε goes to zero, then we say that $\{P_n^{-1}(A_n - P_n)\}_n$ shows the Weakest Strong Clustering Property. More precisely, this case occurs if for any $C_\varepsilon > 0$ such that $N_o(\varepsilon, n) \le C_\varepsilon$ holds definitely, then the relation $\sup_\varepsilon C_\varepsilon = \infty$ holds true.*

Notice that the case where $\sup_\varepsilon C_\varepsilon = C \in \mathbb{R}^+$ is characterized by a "true superlinear" behaviour of the PCG method, meaning that for n going to infinity, the number of the iterations decreases to a value close to $\lceil C \rceil$. When the *Weakest Strong Clustering* is obtained, then the PCG method is optimal [2] in the sense that we generally observe a number of iterations which is constant with respect to n (this behaviour also characterizes the case where all the eigenvalues belong to a fixed interval bounded away from zero). Compare [35] and [31, 36], to see these different behaviours.

Finally, it is worthwhile stressing that the assumption on the minimal singular value of P_n being greater than a fixed constant is necessary and cannot be removed [34]. Moreover, it can be easily verified that the minimal singular value of a sequence of matrices discretizing in a consistent way differential operators must tend to zero as n tends to infinity [9, 23]. This fact justifies the following definition.

DEFINITION 3.5. [31, 34] *A sequence of matrices $\{X_n\}_n$, where $X_n \in \mathbb{C}^{n\times n}$, is said to be sparsely vanishing if there exists a nonnegative function $x(s)$ with $\lim_{s\to 0} x(s) = 0$ so that for any $\varepsilon > 0$ there exists $n_\varepsilon \in \mathbb{N}$ such that for any $n \ge n_\varepsilon$*

$$\frac{1}{n}\#\{i : \sigma_i(X_n) \le \varepsilon\} \le x(\varepsilon),$$

where $\{\sigma_i(X_n)\},\ i = 1, \ldots, n$, denotes the complete set of the singular values of X_n.

In fact, if the matrix sequence $\{P_n\}_n$ is *sparsely vanishing* according to Definition 3.5 then the *Weak Clustering Property* can be obtained again.

LEMMA 3.6. [31, 34] *Consider two sequences $\{A_n\}_n$ and $\{P_n\}_n$, where $A_n, P_n \in \mathbb{C}^{n\times n}$. If the sequence $\{P_n\}_n$ is* sparsely vanishing *(with P_n nonsingular at least definitely) and if there exists a sequence $\{D_n\}_n$, where $D_n \in \mathbb{C}^{n\times n}$, so that $\lim_{n\to\infty} \|A_n - P_n - D_n\|_2 = 0$ with* rank$(D_n) \le \varepsilon n$*, then the* Weak Clustering Property *holds.*

In a previous paper [36] we proved that the Toeplitz sequence $\{\Delta_n(m, k)\}_n$ is an optimal preconditioning sequence according to Definition 3.1 for the sequence $\{A_n(a, m, k)\}_n$ with respect to the case of any strictly positive coefficient function $a(x)$. Hereafter, we want to introduce a special improvement of Toeplitz based preconditioners also giving truly effective results in the degenerate elliptic case. This preconditioning sequence can be constructed by coupling the previously considered Toeplitz sequence $\{\Delta_n(m,k)\}_n$ with the sequence of the suitably scaled main diagonal of the matrices $A_n(a,m,k)$, the aim being to introduce more informative content from the original linear system into the preconditioner, while keeping the additional computational cost as low as possible. More precisely, for any fixed n, we consider as a preconditioner the matrix

$$P_n(a,m,k) = D_n^{1/2}(a,m,k)\Delta_n(m,k)D_n^{1/2}(a,m,k),$$

where $D_n(a,m,k) = \Delta^{-1}\mathrm{diag}(A_n(a,m,k))$, with $\Delta > 0$ being the main diagonal entry of the positive definite Toeplitz matrix $\Delta_n(m,k)$, so that, for the limit case of $a(x) \equiv 1$, we obtain $D_n(a,m,k) = I_n$ and $P_n(a,m,k) = A_n(1,m,k) = \Delta_n(m,k)$.

PROPOSITION 3.7. *If $a(x)$ is a strictly positive function then, for any dimension n, the preconditioner $P_n(a,m,k)$ is a well-defined symmetric positive definite matrix. The same holds true, at least for n large enough, in the case where $a(x)$ is a nonnegative function with a finite number of zeros.*

Proof. By recalling relations in (2.2) and (2.5) respectively, the entries of the main diagonal of $A_n(a,m,k)$ are a positive linear combination of samplings of the function $a(x)$ at equispaced mesh points. So, if $a(x)$ shows at most a finite number of zeros, each entry is positive for a mesh spacing fine enough and, therefore, $D_n^{1/2}(a,m,k)$ is a well-defined positive definite diagonal matrix. Now, the positive definiteness of the matrix $P_n(a,m,k)$ is easily proved by defining the vector $\mathbf{y} = D_n^{1/2}(a,m,k)\mathbf{x} \neq 0$ for each $\mathbf{x} \neq 0$. In fact, we have

$$\lambda_{\min}(P_n(a,m,k)) = \min_{\mathbf{x}\neq 0} \frac{\mathbf{x}^T P_n(a,m,k)\mathbf{x}}{\mathbf{x}^T\mathbf{x}} = \min_{\mathbf{y}\neq 0} \frac{\mathbf{y}^T \Delta_n(m,k)\mathbf{y}}{\mathbf{y}^T D_n^{-1}(a,m,k)\mathbf{y}} > 0,$$

since both $D_n^{-1}(a,m,k)$ and $\Delta_n(m,k)$ are positive definite. ☐

4. Numerical experiments. Before giving a rigorous spectral analysis of the preconditioned matrix sequences and before proving optimality and clustering features of our PCG method, we present several numerical experiments both in $1D$ and $2D$ cases that motivated our subsequent work. Therefore, in the following two subsections we considered some cases in which problems (1.1) and (1.2) are strictly elliptic ($\inf a > 0$), semielliptic ($\inf a = 0$), with regular ($a \in C^2(\overline{\Omega})$) and irregular weight function ($a(x)$ piecewise regular, with a countably infinite number of discontinuity points, $a \in L^1(\Omega)$ with zeros and/or poles). The information that we get from all the tables both in the $1D$ and $2D$ cases is that the eigenvalues of the preconditioned matrix $P_n^{-1}(a,m,k)A_n(a,m,k)$ are clustered at unity and that the resulting PCG method is optimal at least when a is regular.

4.1. $1D$ Case. In tables 4.1–4.3 we report the number of PCG iterations required to obtain $||r||_2/||b||_2 \leq 10^{-7}$ when the data vector is made up of all ones with respect to increasing matrix dimensions. The test functions $a(x)$ are listed in the first column, the preconditioners are denoted in the heading by $P = P_n(a,m,k)$, $\Delta = \Delta_n(m,k)$ and $D = D_n(a,m,k)$; the pair (k,m) varies among $(1,2)$, $(1,3)$ and $(2,2)$. We observe that in some definition domains we specify $a(x)$ for x outside $[0,1]$ according to relation (2.1) and this is necessary in order to manage the extra points. Some additional numerical evidences can be found in [37]. With respect to the preconditioner $D_n(a,m,k)$ only the case $n = 100$ is reported since a number of iteration at least equal to the matrix dimension n is in general required in order to reach the convergence. In addition, the notation '—' means that no convergence is reached in 1000 PCG iterations.

ETNA
Kent State University
etna@mcs.kent.edu

TABLE 4.1
Number of PCG iterations - $1D$ case, $k = 1, m = 2$.

$a(x)$	n: 100 P	100 Δ	100 D	200 P	200 Δ	300 P	300 Δ	400 P	400 Δ	500 P	500 Δ	600 P	600 Δ
$1+x$	3	11	101	3	11	3	11	3	11	3	11	3	11
$\exp(x)$	3	14	102	3	14	3	14	3	15	3	15	3	15
$\sin^2(7x)+1$	8	12	101	8	12	8	12	8	12	8	12	8	12
x	6	59	102	7	86	7	107	7	124	7	140	7	153
x^2	4	132	103	4	281	4	433	4	587	4	742	4	896
x^4	9	582	103	10	—	10	—	10	—	10	—	11	—
$\lvert x-\frac{1}{2}\rvert+\frac{1}{2}$	4	11	50	4	11	4	11	4	11	4	11	4	11
$\lvert x-\frac{1}{2}\rvert$	8	39	50	8	57	8	70	9	81	9	91	9	100
$1+\sqrt{x}$ if $x \geq 0$ 1 if $x<0$	4	11	101	4	11	4	12	4	12	4	12	4	12
$\exp(x)$ if $x \leq \frac{2}{3}$ $2-x$ if $x>\frac{2}{3}$	5	11	102	5	11	5	11	6	11	6	11	6	12
$\left\lceil \frac{1}{\sqrt{x}} \right\rceil$ if $x>0$ 1 if $x \leq 0$	9	12	101	10	15	11	17	11	18	13	19	13	20
$\frac{\lceil \frac{1}{x} \rceil}{1+\lceil \frac{1}{x} \rceil}$ if $x>0$ 1 if $x \leq 0$	7	8	101	8	9	8	9	9	9	9	9	9	9
$\frac{x\lceil \frac{1}{x} \rceil}{1+\lceil \frac{1}{x} \rceil}$ if $x>0$ 0 if $x \leq 0$	8	50	102	9	72	10	89	10	103	11	115	11	126

In Tables 4.4 and 4.5 we give evidence of the number of outliers with respect to a cluster at unity with radius $\delta = 0.1$; specifically, we count the total number of the eigenvalues of $P_n^{-1}(a,m,k)A_n(a,m,k)$ not belonging to $(1-\delta, 1+\delta)$, the related percentage with respect to the matrix dimension, and the number of outliers less then $1-\delta$ (the latter quantity is reported in brackets).

4.1.1. Convergence remarks. Some remarks are needed. In Tables 4.1–4.3, the observed number of PCG iterations is constant with respect to increasing matrix dimensions when the preconditioner is $\Delta_n(m,k)$ or $P_n(a,m,k)$ and the coefficient function $a(x)$ is strictly positive. This independence with regard to n fully agrees with the spectral analysis of $\{\Delta_n^{-1}(m,k)A_n(a,m,k)\}_n$ given in [36] and the spectral clustering theorems that will be proved in §5 for the sequence $\{P_n^{-1}(a,m,k)A_n(a,m,k)\}_n$. When the coefficient function $a(x)$ has zeros, it is immediatly observed that the only working preconditioner of the three under consideration is $P_n(a,m,k)$. In this case, as shown in Tables 4.4 and 4.5, the number of outlying eigenvalues grows very slowly (only logarithmically with n).

The presence of jumps or discontinuities of $a(x)$ or of its derivatives up to the first order, does not spoil the performances of the associated PCG methods when $\Delta_n(m,k)$ or $P_n(a,m,k)$ are used as preconditioners (see [31] for more details on this). The case of highly oscillating coefficient function $a(x)$ slightly deteriorates the performance of the second preconditioner $P_n(a,m,k)$ and indeed the simpler preconditioner $\Delta_n(m,k)$ performs as well as $P_n(a,m,k)$. This is obvious since the matrix $D_n(a,m,k)$ (the diagonal part of $A_n(a,m,k)$) is given by an equispaced sampling of $a(x)$. Therefore, $D_n(a,m,k)$ cannot be in general a faithful representation of $a(x)$ when $a(x)$ oscillates too much with regard to the grid parameter h. This phenomenon was also noticed in [16, 31] with regard to similar

ETNA
Kent State University
etna@mcs.kent.edu

TABLE 4.2
Number of PCG iterations - $1D$ case, $k=1, m=3$.

$a(x)$	n = 100			200		300		400		500		600	
	P	Δ	D	P	Δ	P	Δ	P	Δ	P	Δ	P	Δ
$1+x$	3	11	106	3	11	3	11	3	11	3	11	3	11
$\exp(x)$	3	14	106	3	14	3	15	3	15	3	15	3	15
$\sin^2(7x)+1$	8	12	106	8	12	8	12	8	12	8	12	8	12
x	6	60	107	7	86	7	107	7	124	7	140	7	154
x^2	4	132	108	4	282	4	433	4	588	4	742	4	896
x^4	9	583	108	9	—	10	—	10	—	10	—	10	—
$\lvert x-\frac{1}{2}\rvert+\frac{1}{2}$	4	11	51	4	11	4	11	4	11	4	11	4	11
$\lvert x-\frac{1}{2}\rvert$	8	39	51	8	57	8	70	9	81	9	91	9	100
$1+\sqrt{x}$ if $x\geq 0$ 1 if $x<0$	4	11	106	4	11	4	12	4	12	4	12	4	12
$\exp(x)$ if $x\leq\frac{2}{3}$ $2-x$ if $x>\frac{2}{3}$	5	11	106	6	11	6	11	6	11	6	11	6	12
$\left\lceil\frac{1}{\sqrt{x}}\right\rceil$ if $x>0$ 1 if $x\leq 0$	9	12	105	10	15	11	17	11	18	13	19	13	20
$\frac{\lceil\frac{1}{x}\rceil}{1+\lceil\frac{1}{x}\rceil}$ if $x>0$ 1 if $x\leq 0$	7	8	106	8	8	8	9	9	9	9	9	9	9
$\frac{x\lceil\frac{1}{x}\rceil}{1+\lceil\frac{1}{x}\rceil}$ if $x>0$ 0 if $x\leq 0$	8	50	107	9	72	10	89	10	103	11	115	11	126

TABLE 4.3
Number of PCG iterations - $1D$ case, $k=2, m=2$.

$a(x)$	n = 100			200		300		400		500		600	
	P	Δ	D	P	Δ	P	Δ	P	Δ	P	Δ	P	Δ
$1+x$	4	13	399	4	13	4	13	4	13	4	13	4	13
$\exp(x)$	4	16	398	4	17	4	17	4	17	4	17	4	17
$\sin^2(7x)+1$	13	14	401	13	14	13	14	13	14	13	14	13	14
x	10	67	429	10	100	11	124	11	146	12	165	12	182
x^2	10	143	432	11	302	11	468	11	638	12	831	12	—
x^4	6	809	446	6	—	6	—	6	—	6	—	6	—
$\lvert x-\frac{1}{2}\rvert+\frac{1}{2}$	6	13	183	7	14	7	14	7	14	8	14	9	14
$\lvert x-\frac{1}{2}\rvert$	15	43	184	17	65	23	96	22	96	26	109	29	120
$1+\sqrt{x}$ if $x\geq 0$ 1 if $x<0$	6	13	387	6	14	6	14	6	14	6	15	6	15
$\exp(x)$ if $x\leq\frac{2}{3}$ $2-x$ if $x>\frac{2}{3}$	10	13	394	13	13	15	14	15	14	17	14	18	14
$\left\lceil\frac{1}{\sqrt{x}}\right\rceil$ if $x>0$ 1 if $x\leq 0$	21	16	406	31	20	48	22	56	26	77	28	94	30
$\frac{\lceil\frac{1}{x}\rceil}{1+\lceil\frac{1}{x}\rceil}$ if $x>0$ 1 if $x\leq 0$	14	11	388	23	11	29	11	38	12	65	12	68	12
$\frac{x\lceil\frac{1}{x}\rceil}{1+\lceil\frac{1}{x}\rceil}$ if $x>0$ 0 if $x\leq 0$	18	57	430	27	83	48	103	63	120	74	136	89	150

ETNA
Kent State University
etna@mcs.kent.edu

TABLE 4.4
Number of outliers - 1D case, $k = 1, m = 1, 2, 3$.

$a(x)$	n = 75	n = 150	n = 300	n = 600
$1 + x$	0	0	0	0
$\exp(x)$	0	0	0	0
$\sin^2(7x) + 1$	4 (2) 5.3%	4 (2) 2.6%	4 (2) 1.3%	4 (2) 0.6%
x	2 (2) 2.6%	2 (2) 1.3%	3 (3) 1%	3 (3) 0.5%
x^2	0	0	0	0
x^4	6 (1) 8%	7 (1) 4.6%	8 (1) 2.6%	9 (1) 1.5%
$\left\|x - \frac{1}{2}\right\| + \frac{1}{2}$	1 1.3%	1 0.6%	1 0.3%	1 0.1%
$\left\|x - \frac{1}{2}\right\|$	5 (4) 6.6%	6 (5) 4%	6 (5) 2%	6 (5) 1%
$1 + \sqrt{x}$ if $x \geq 0$, 1 if $x < 0$	0	0	0	0
$\exp(x)$ if $x \leq \frac{2}{3}$, $2 - x$ if $x > \frac{2}{3}$	2 (1) 2.6%	2 (1) 1.3%	2 (1) 0.6%	2 (1) 0.3%
$\lceil 1/\sqrt{x} \rceil$ if $x > 0$, 1 if $x \leq 0$	5 (2) 6.6%	6 (2) 4%	8 (3) 2.6%	10 (4) 1.6%
$\lceil \frac{1}{x} \rceil / \left(1 + \lceil \frac{1}{x} \rceil\right)$ if $x > 0$, 1 if $x \leq 0$	2 (1) 2.6%	3 (1) 2%	4 (2) 1.3%	5 (2) 0.8%
$x \lceil \frac{1}{x} \rceil / \left(1 + \lceil \frac{1}{x} \rceil\right)$ if $x > 0$, 0 if $x \leq 0$	4 (3) 5.3%	4 (3) 2.6%	6 (4) 2%	7 (5) 1.1%

problems where $(k, m) = (1, 1)$.

4.1.2. Spectral remarks. Figures 4.1 and 4.2 represent the spectra of $A_n(a, m, k)$ and of the two preconditioned matrices, with preconditioners $\Delta_n(m, k)$ or $P_n(a, m, k)$ respectively, for $a(x) = \exp(x)$ and $a(x) = x$ in the case $n = 300$. The clustering properties of $\{P_n^{-1}(a, m, k)A_n(a, m, k)\}_n$ are impressive as well as the fact that the preconditioner $\Delta_n(m, k)$ "removes" all the small eigenvalues of $A_n(a, m, k)$ if and only if the function $a(x)$ does not vanish.

4.1.3. Further observations. A further remark concerns the polynomial $p_{\mathbf{w}} \equiv p_{q(m),k}$, $q(m) = 2m + 1$ associated in the Toeplitz sense to $\Delta_n(m, k)$. For $k = 1$ and $m = 1, 2, 3$ in Figure 4.3.a, we observe that $x = 0$ is the unique zero of order two (the order of the zero at $x = 0$ was predicted in (2.9)). Moreover, for $k = 2$ and $m = 1, 2$ in Figure 4.3.b, we observe that $x = 0$ is the unique zero of order four (the order of the zero at $x = 0$ was predicted in (2.9)).

Now define the quantity

$$\alpha(q(m), k) = \max_x \, p_{q(m),k}(x).$$

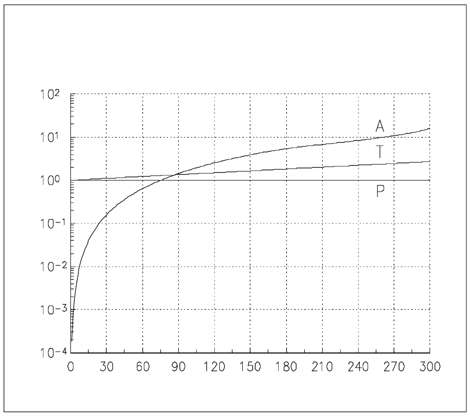
We notice that, for $k = 1, 2$, the numerical experiments tell us that $\alpha(q(m), k)$ is an increasing function of m. In addition, due to the *Locally Toeplitz* structure [41] of the sequence $\{A_n(a, m, k)\}_n$, we have that the eigenvalues distribute as $a(x)p_{q(m),k}(y)$ over $[0, 1] \times [-\pi, \pi]$ in the sense that, for any real–valued continuous function with bounded support $F \in C(\mathbb{R})$, the asymptotic formula

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} F\left(\lambda_i(A_n(a, m, k))\right) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \int_0^1 F(a(x)p_{q(m),k}(y))dxdy \tag{4.1}$$

ETNA
Kent State University
etna@mcs.kent.edu

TABLE 4.5
Number of Outliers - $1D$ case, $k = 2, m = 2$.

$a(x)$	n = 75	150	300	600
$1 + x$	0	0	0	0
$\exp(x)$	0	0	0	0
$\sin^2(7x) + 1$	8 (4) 10.6%	8 (4) 5.3%	8 (4) 2.6%	8 (4) 1.3%
x	4 (4) 5.3%	5 (5) 3.3%	6 (6) 2%	7 (7) 1.1%
x^2	5 (5) 6.6%	6 (6) 4%	7 (7) 2.3%	8 (8) 1.3%
x^4	1 1.3%	1 0.6%	1 0.3%	1 0.16%
$\lvert x - \frac{1}{2}\rvert + \frac{1}{2}$	3 (1) 4%	3 (1) 2%	3 (1) 1%	3 (1) 0.5%
$\lvert x - \frac{1}{2}\rvert$	9 (7) 12%	10 (8) 6.6%	12 (10) 4%	14 (12) 2.3%
$1 + \sqrt{x}$ if $x \geq 0$, 1 if $x < 0$	0	0	0	0
$\exp(x)$ if $x \leq \frac{2}{3}$, $2 - x$ if $x > \frac{2}{3}$	4 (2) 5.3%	4 (2) 2.6%	4 (2) 1.3%	4 (2) 0.6%
$\lceil 1/\sqrt{x} \rceil$ if $x > 0$, 1 if $x \leq 0$	9 (5) 12%	12 (6) 8%	16 (8) 5.3%	21 (11) 3.5%
$\lceil \frac{1}{x} \rceil / \left(1 + \lceil \frac{1}{x} \rceil\right)$ if $x > 0$, 1 if $x \leq 0$	7 (4) 9.3%	10 (5) 6.6%	13 (7) 4.3%	17 (9) 2.8%
$x \lceil \frac{1}{x} \rceil / \left(1 + \lceil \frac{1}{x} \rceil\right)$ if $x > 0$, 0 if $x \leq 0$	10 (7) 13.3%	14 (9) 9.3%	17 (11) 5.6%	21 (13) 3.5%



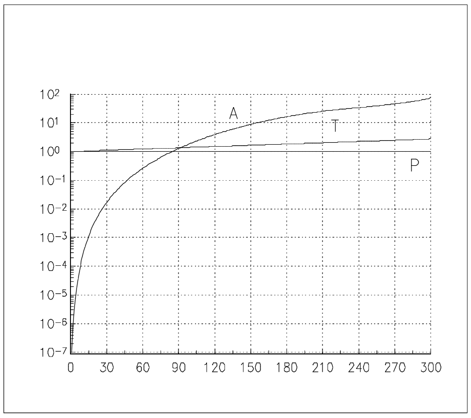


FIG. 4.1. *Complete sets of the ordered eigenvalues of $A = A_n(a, m, k)$, $T = \Delta_n^{-1}(m, k)A_n(a, m, k)$ and $P = P_n^{-1}(a, m, k)A_n(a, m, k)$ for $n = 300$, $k = 1$, $m = 3$, and $k = 2$, $m = 2$, $a(x) = \exp(x)$.*

holds true. Therefore, owing to (4.1) and Theorem 2.3, a simple verification is that

$$\lim_{n\to\infty} \lambda_{\max}(A_n(a, m, k)) = \alpha(q(m), k) \cdot \max_x \ a(x).$$

Moreover, with a monotonicity argument, it follows that

$$\lambda_{\min}(A_n(a, m, k)) \geq \min_x \ a(x) \cdot \lambda_{\min}(A_n(1, m, k)),$$

ETNA
Kent State University
etna@mcs.kent.edu

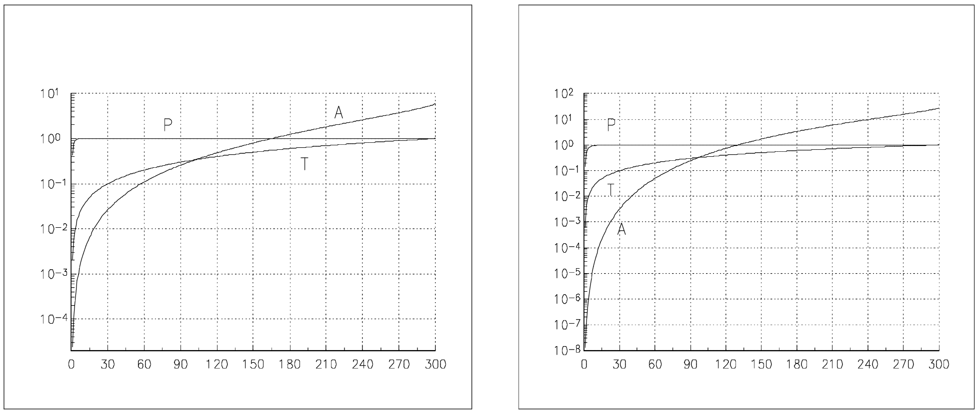


FIG. 4.2. *Complete sets of the ordered eigenvalues of* $A = A_n(a, m, k)$*,* $T = \Delta_n^{-1}(m, k)A_n(a, m, k)$ *and* $P = P_n^{-1}(a, m, k)A_n(a, m, k)$ *for* $n = 300$*,* $k = 1$*,* $m = 3$*, and* $k = 2$*,* $m = 2$*,* $a(x) = x$*.*

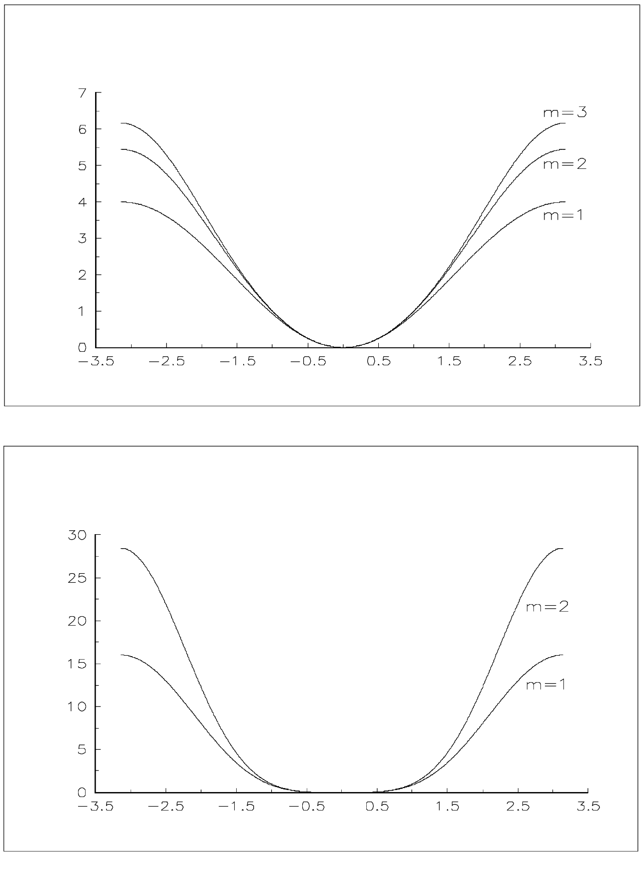


FIG. 4.3. *Polynomials* $p_{q,k}(x)$ *for* $k = 1$*,* $m = 1, 2, 3$ *and for* $k = 2$*,* $m = 1, 2$ *with* $x \in [-\pi, \pi]$*.*

ETNA
Kent State University
etna@mcs.kent.edu

TABLE 4.6
Number of PCG iterations - 2D case, $k = 2, m = 1$.

	$n_1 = n_2$								
	10			20			30		
$a(x,y)$	D	Δ	P	D	Δ	P	D	Δ	P
$1+x+y$	49	13	3	171	14	4	367	14	4
$\exp(x+y)$	49	19	3	172	22	3	365	24	4
$\sin^2(7(x+y))+1$	50	10	11	172	11	13	367	12	14
$x+y$	51	22	5	177	34	5	373	43	5
$(x+y)^2$	52	41	5	179	95	5	381	148	6
$(x+y)^4$	53	88	4	181	444	4	386	–	4
$\lvert x-\frac{1}{2}\rvert+\lvert y-\frac{1}{2}\rvert+\frac{1}{2}$	15	10	5	61	13	6	126	14	7
$\lvert x-\frac{1}{2}\rvert+\lvert y-\frac{1}{2}\rvert$	15	14	7	60	25	9	125	33	11
$1+\sqrt{x+y}$ if $x+y\geq 0$ 1 if $x+y<0$	49	9	4	171	10	4	366	11	4
$\exp(x+y)$ if $x+y\leq\frac{2}{3}$ $2-(x+y)$ if $x+y>\frac{2}{3}$	51	23	7	175	35	10	379	42	12
$\left\lceil\frac{1}{\sqrt{x}}\right\rceil+\left\lceil\frac{1}{\sqrt{y}}\right\rceil$ if $x,y>0$ 1 if x or $y\leq 0$	52	11	10	180	15	14	379	18	20
$\frac{\lceil\frac{1}{x}\rceil}{1+\lceil\frac{1}{x}\rceil}+\frac{\lceil\frac{1}{y}\rceil}{1+\lceil\frac{1}{y}\rceil}$ if $x,y>0$ 1 if x or $y\leq 0$	49	8	7	172	10	9	369	10	11
$(x+y)\left(\frac{\lceil\frac{1}{x}\rceil}{1+\lceil\frac{1}{x}\rceil}+\frac{\lceil\frac{1}{y}\rceil}{1+\lceil\frac{1}{y}\rceil}\right)$ if $x,y>0$; 1 if x or $y\leq 0$	50	18	7	176	25	9	371	31	11
$\exp(\frac{\lceil\frac{1}{x}\rceil}{(1.+\lceil\frac{1}{x}\rceil)})\left\lceil\frac{1}{\sqrt{y}}\right\rceil+y$ if $x,y>0$ 1 if x or $y\leq 0$	72	13	11	261	19	17	554	24	24

where $\lambda_{\min}(A_n(1,m,k)) = c_k p^{(2k)}_{q(m),k}(0)n^{-2k}(1+o(1))$, with c_k a positive constant independent of m and n (see [24, 29]). However, $p_{q(m),k}(x) = x^{2k} + O(x^{2k+\nu})$, where $\nu = \nu(m,k)$ is the accuracy of the FD formula reported in Lemma 2.1, and then we deduce that $p^{(2k)}_{q(m),k}(0) = (2k)!$, which depends on k but is independent of m (refer to [36]). Consequently, for a fixed value of $k \in \{1,2\}$, the conditioning of the sequence of matrices $\{A_n(a,m,k)\}_n$ worsens as m increases. This is a theoretical explanation of the fact that the number of iterations of the PCG increases with m when the simple diagonal preconditioner $D_n(a,m,k)$ is used and in fact for $k \geq 1$ and $m \geq 2$, n iterations are not sufficient for the convergence when the diagonal preconditioning is applied.

On the other hand, when the Toeplitz preconditioner $\Delta_n(m,k)$ is used, the eigenvalues of the related preconditioned matrices behave as the function $a(x)$ in the sense that for any real–valued continuous function with bounded support $F \in C(\mathbb{R})$, we have [36]

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}F\left(\lambda_i(\Delta_n^{-1}(m,k)A_n(a,m,k))\right) = \int_0^1 F(a(x))dx.$$

Therefore, the dependency on $p_{q(m),k}$ is completely lost and this explains why, for fixed $a(x)$ and by varying the parameters m and k, we have substantial stability in the number of the PCG iterations when the preconditioners $\Delta_n(m,k)$ and $P_n(a,m,k)$ are used.

ETNA
Kent State University
etna@mcs.kent.edu

TABLE 4.7
Number of Outliers - $2D$ case, $k = 2, m = 1$.

$a(x,y)$	$n_1 = n_2$: 10	20	30
$1+x+y$	0	0	0
$\exp(x+y)$	0	0	0
$\sin^2(7(x+y))+1$	24 (12) 24%	51 (23) 12.75%	64 (28) 7.1%
$x+y$	0	0	0
$(x+y)^2$	0	0	0
$(x+y)^4$	0	0	0
$\lvert x-\frac{1}{2}\rvert + \lvert y-\frac{1}{2}\rvert + \frac{1}{2}$	4 4%	5 (1) 1.25%	7 (1) 0.7%
$\lvert x-\frac{1}{2}\rvert + \lvert y-\frac{1}{2}\rvert$	7 (1) 7%	15 (4) 3.75%	23 (7) 2.5%
$1+\sqrt{x+y}$ if $x+y \geq 0$ 1 if $x+y<0$	0	0	0
$\exp(x+y)$ if $x+y \leq \frac{2}{3}$ $2-(x+y)$ if $x+y > \frac{2}{3}$	5 (3) 5%	11 (6) 2.75%	15 (8) 1.6%
$\left\lceil \frac{1}{\sqrt{x}} \right\rceil + \left\lceil \frac{1}{\sqrt{y}} \right\rceil$ if $x,y>0$ 1 if x or $y \leq 0$	11 (8) 11%	32 (24) 8%	56 (34) 6.2%
$\frac{\lceil \frac{1}{x} \rceil}{1+\lceil \frac{1}{x} \rceil} + \frac{\lceil \frac{1}{y} \rceil}{1+\lceil \frac{1}{y} \rceil}$ if $x,y>0$ 1 if x or $y \leq 0$	5 (5) 5%	11 (10) 2.75%	23 (18) 2.5%
$(x+y)\left(\frac{\lceil \frac{1}{x} \rceil}{1+\lceil \frac{1}{x} \rceil} + \frac{\lceil \frac{1}{y} \rceil}{1+\lceil \frac{1}{y} \rceil} \right)$ if $x,y>0$ 1 if x or $y \leq 0$	4 (1) 4%	14 (6) 3.5%	21 (7) 2.3%
$\exp(\lceil \frac{1}{x} \rceil / (1. + \lceil \frac{1}{x} \rceil)) \left\lceil \frac{1}{\sqrt{y}} \right\rceil + y$ if $x,y>0$ 1 if x or $y \leq 0$	11 (9) 11%	36 (26) 9%	56 (34) 6.2%

4.2. $2D$ **Case.** Here we consider a set of numerical tests for the numerical solution of problem (1.2) where $k = 2$ and where each derivative in the x direction is discretized by a FD formula of order of accuracy 2 ($m_1 = m = 1$) and each derivative in the y direction is discretized by a FD formula of order 2 of accuracy ($m_2 = m = 1$). By ordering the unknowns in the classical manner, the obtained coefficient matrix denoted by $A_n(a, m, k)$ is a two level matrix of external dimension $n_2 \times n_2$ where each block has dimension $n_1 \times n_1$ so that the global size is $N(n) \times N(n)$ with $N(n) = n_1 n_2$ and $n_1 \sim n_2$. Moreover, by the structure of each discretizing formula we deduce that the matrix $A_n(a, m, k)$ is banded at each level.

In Table 4.6 we report the number of PCG iterations where $n_1 = n_2$, with the test functions $a(x, y)$ listed in the first column and the preconditioners given in the heading. The data vector $\mathbf{b}$ is made up of all ones. In Table 4.7 we give the total number of outliers with respect to a cluster at unity with radius $\delta = 0.1$, the related percentage and the number of outliers less then $1 - \delta$.

It is interesting that the only noteworthy remark is that there are no new phenomena when comparing to the $1D$ case, so that the remarks given in 4.1.1 and 4.1.2 can be repeated almost verbatim in the present $2D$ case.

In conclusion, we observe that the results recorded in the tables as well as the displayed figures give evidence of the goodness of the proposed approach so that in the following sections we are motivated to derive a structural and spectral analysis of the preconditioned matrix sequences in order to explain the observed numerical behaviour.

Since the analysis is rather technical, only the $1D$ case is reported in detail here. The multidimensional case is analyzed in [38] and is heavily based on the results of §5 and §6. Consequently, the contribution of this paper is also aimed at creating the necessary tools in order to manipulate the same kind of problems in a higher dimensional setting in a simple way.

5. The clustering properties of the Toeplitz based preconditioner. In order to analyze in depth the asymptotic spectral properties of the sequence of preconditioned matrices $\{P_n^{-1}(a,m,k)A_n(a,m,k)\}_n$, it is better to formulate the problem by taking into account the following Lemma.

LEMMA 5.1. *The preconditioned matrix $P_n^{-1}(a,m,k)A_n(a,m,k)$ is similar to the matrix $\Delta_n^{-1}(m,k)A_n^*(a,m,k)$, where $A_n^*(a,m,k)$ is the symmetric positive definite matrix defined as*

$$A_n^*(a,m,k) = D_n^{-1/2}(a,m,k)A_n(a,m,k)D_n^{-1/2}(a,m,k),$$

under the assumption that $a(x)$ possesses at most a finite number of zeros and n is large enough.

Proof. The positive definiteness of the matrix $A_n^*(a,m,k)$ can be proved in the same manner as in Proposition 3.7, by recalling that $A_n(a,m,k)$ is positive definite by virtue of Theorems 2.3 and 2.4. The matrix defining the similarity property can clearly be chosen as $D_n^{1/2}(a,m,k)$. □

The computational problem of solving a linear system with matrix $A_n(a,m,k)$ is now reduced to the solution of a linear system with a coefficient matrix given by $A_n^*(a,m,k)$. In the following we will show that the latter matrix for large n can be written as the Toeplitz matrix $\Delta_n(m,k)$ plus terms of infinitesimal spectral norm. We point out that this fact is essentially based on the asymptotic expansion of the entries of the matrices $A_n(a,m,k)$ given in Appendix **A**. Finally, the clustering properties proved in the following are a consequence of the asymptotic expansion of $A_n^*(a,m,k)$ and of the second order results reported in Appendix **B**.

5.1. The strictly elliptic case.

5.1.1. Asymptotic expansion of preconditioned matrices.

Starting from the asymptotic expansion of the matrices $A_n^*(a,m,k)$ defined in Proposition 5.2, we obtain an interesting asymptotic expansion for the preconditioned matrices.

PROPOSITION 5.2. *If the coefficient function $a(x)$ is strictly positive and belongs to $C^2(\overline{\Omega})$, then the matrices $A_n^*(a,m,k)$ can be expanded as*

$$A_n^*(a,m,k) = \Delta_n(m,k) + h^2\Theta_n(a,m,k) + o(h^2)E_n(a,m,k),$$

where $\Theta_n(a,m,k)$ and $E_n(a,m,k)$ are bounded symmetric band matrices. If $a(x)$ belongs to $C^1(\overline{\Omega})$ then $A_n^*(a,m,k) = \Delta_n(m,k) + \Theta_n(a,m,k)$, where $\Theta_n(a,m,k)$ is a band matrix whose elements are $O(h\omega_{a_x}(h))$. Finally, if $a(x)$ belongs to $C(\overline{\Omega})$ then $A_n^*(a,m,k) = \Delta_n(m,k) + \Theta_n(a,m,k)$, where $\Theta_n(a,m,k)$ is a band matrix whose elements are $O(\omega_a(h))$. Here the matrices $\Theta_n(a,m,k)$ and $E_n(a,m,k)$ always possess the same bandwidth as $\Delta_n(m,k)$ and $\omega_f(\cdot)$ denotes the modulus of continuity of the function f.

Proof. It is enough to consider the nonzero coefficients of the lower triangular part of $A_n^*(a,m,k)$ due to the symmetry property. Let Δ be the constant entry along the main diagonal and let Δ_{r-p} be the constant entry along the p^{th} subdiagonal in the Toeplitz matrix

$\Delta_n(m,k)$. The coefficients $(A_n^*)_{r,r-p}$ are defined as

$$(A_n^*)_{r,r-p} = \Delta \frac{(A_n)_{r,r-p}}{\sqrt{(A_n)_{r,r}(A_n)_{r-p,r-p}}}.$$

Clearly, for $p=0$, we have $(A_n^*)_{r,r-p} = \Delta$, so that $(\Theta_n)_{r,r} = (E_n)_{r,r} = 0$.

Case k odd: According to Proposition A.1 and to the assumption of the strictly positiveness of the function coefficient $a(x)$, we can also prove that, for each $p = 1, \ldots, 2m-1$,

$$\begin{aligned}
(A_n^*)_{r,r-p} &= \\
&\frac{\Delta\left(a_{r-\frac{p}{2}}\Delta_{r-p}+h^2 a_{xx,r-\frac{p}{2}}\delta_p+o(h^2)\right)}{\sqrt{\left(a_{r-\frac{p}{2}}\Delta+h a_{x,r-\frac{p}{2}}\beta_p+h^2 a_{xx,r-\frac{p}{2}}\gamma_p+o(h^2)\right)\left(a_{r-\frac{p}{2}}\Delta-h a_{x,r-\frac{p}{2}}\beta_p+h^2 a_{xx,r-\frac{p}{2}}\gamma_p+o(h^2)\right)}} \\
&= \frac{\Delta_{r-p}\left(1+h^2 a_{xx,r-\frac{p}{2}}\delta_p^*\right)+o(h^2)}{\sqrt{\left(1+h a_{x,r-\frac{p}{2}}\beta_p^*+h^2 a_{xx,r-\frac{p}{2}}\gamma_p^*+o(h^2)\right)\left(1-h a_{x,r-\frac{p}{2}}\beta_p^*+h^2 a_{xx,r-\frac{p}{2}}\gamma_p^*+o(h^2)\right)}} \\
&= \frac{\Delta_{r-p}\left(1+h^2 a_{xx,r-\frac{p}{2}}\delta_p^*\right)+o(h^2)}{\sqrt{1+\left(2a_{xx,r-\frac{p}{2}}\gamma_p^*-(\beta_p^*)^2 a^2_{x,r-\frac{p}{2}}\right)h^2+o(h^2)}} \\
&= \Delta_{r-p} + \left(\delta_p^* a_{xx,r-p/2} - \Delta_{r-p}\left(2a_{xx,r-\frac{p}{2}}\gamma_p^* - a^2_{x,r-\frac{p}{2}}(\beta_p^*)^2\right)/2\right)h^2 + o(h^2),
\end{aligned}$$

so that

$$(\Theta_n)_{r,r-p} = \delta_p^* a_{xx,r-p/2} - \Delta_{r-p}\left(2a_{xx,r-p/2}\gamma_p^* - a^2_{x,r-p/2}(\beta_p^*)^2\right)/2,$$

and the claimed thesis follows.

Case k even ($q=2m$): Since the same type of asymptotic expansions holds true by virtue of the Proposition A.2, the thesis follows in the same manner where

$$(\Theta_n)_{r,r-p} = \tilde{\delta}_p^* a_{xx,r-p/2} - \Delta_{r-p}\left(2a_{xx,r-p/2}\tilde{\gamma}_p^* - a^2_{x,r-p/2}(\tilde{\beta}_p^*)^2\right)/2.$$

When the function $a(x)$ has less regularity, the claimed thesis is an easy consequence of the preceding steps where the Taylor expansions are halted according to the regularity of $a(x)$. □

5.1.2. Clustering properties of preconditioned matrices. We start with the following preliminary clustering result.

THEOREM 5.3. *If the coefficient function $a(x)$ is strictly positive and belongs to $C(\overline{\Omega})$, then for any $\varepsilon > 0$ all the eigenvalues of the preconditioned matrix $P_n^{-1}(a,m,k)\,A_n(a,m,k)$ lie in the open interval $(1-\varepsilon, 1+\varepsilon)$ except for $o(n)$ outliers* [Weak Clustering Property].

Proof. Due to the similarity between $P_n^{-1}(a,m,k)A_n(a,m,\,k)$ and $\Delta_n^{-1}(m,k)A_n^*(a,m,k)$, we can analyze the spectra of the latter. By virtue of Proposition B.1 and Theorem 2.6, the sequence $\{\Delta_n(m,k)\}_n$ is *sparsely vanishing* and the considered matrices are nonsingular for any n. Therefore, by recalling Proposition 5.2 and setting $D_n \equiv 0$ for any n, Lemma 3.6 applies, so that the claimed thesis follows. □

Moreover, the spectral analysis of the preconditioned matrix sequence $\{P_n^{-1}(a,m,k)A_n(a,m,k)\}_n$ can be improved in the case of a coefficient function $a(x)$ belonging to $C^2(\overline{\Omega})$, with respect to the optimality and the clustering properties.

THEOREM 5.4. *If $k = 1$, the coefficient function $a(x)$ is strictly positive and belongs to $C^2(\overline{\Omega})$ and if the maximal order of the zeros of the related Toeplitz generating polynomial $p_{\mathbf{w}}(x) = |p_{\mathbf{c}}(x)|^2$ equals 2, then the spectra of the sequence of preconditioned matrices $\{P_n^{-1}(a,m,k)A_n(a,m,k)\}_n$ belong to the interval $[d_1, d_2]$, with d_i universal positive constants independent of n and well separated from zero.*

Proof. Due to a similarity argument, we analyze the sequence $\{\Delta_n^{-1}(m,k)A_n^*(a,m,k)\}_n$. Now, according to the assumptions and Proposition 5.2 we have

$$A_n^*(a,m,k) = \Delta_n(m,k) + h^2\Theta_n(a,m,k) + o(h^2)E_n(a,m,k),$$

so that

$$\Delta_n^{-1}(m,k)A_n^*(a,m,k) = I_n + \Delta_n^{-1}(m,k)(h^2\Theta_n(a,m,k) + o(h^2)E_n(a,m,k)).$$

By the hypothesis on the order of the zeros of $|p_{\mathbf{c}}(x)|^2$, we infer that there exists a constant C so that $\|\Delta_n^{-1}(m,k)\|_2 \le Ch^{-2}$ ([7, 32]). Therefore by standard linear algebra we know that

$$\begin{aligned}\lambda_{\max}(\Delta_n^{-1}(m,k)A_n^*(a,m,k)) &\le \|\Delta_n^{-1}(m,k)A_n^*(a,m,k)\|_2 \\ &\le \|I_n\|_2 + \|\Delta_n^{-1}(m,k)(h^2\Theta_n(a,m,k) + o(h^2)E_n(a,m,k))\|_2 \\ &\le 1 + C\|\Theta_n(a,m,k)\|_2 + o(1).\end{aligned}$$

Conversely, for obtaining a bound from below for $\lambda_{\min}(\Delta_n^{-1}(m,k)A_n^*(a,m,k))$ we consider the inverse matrix $[A_n^*(a,m,k)]^{-1}\Delta_n(m,k)$ and we apply Proposition 5.2 obtaining

$$[A_n^*(a,m,k)]^{-1}\Delta_n(m,k) = I_n - [A_n^*(a,m,k)]^{-1}(h^2\Theta_n(a,m,k) + o(h^2)E_n(a,m,k)).$$

Since $a(x)$ is positive, as a consequence of Theorem 2.3 (refer to [36]) we deduce that $\|[A_n^*(a,m,k)]^{-1}\|_2 \le (\max a)(\min a)^{-1}\|\Delta_n^{-1}(m,k)\|_2 \le C(\max a)(\min a)^{-1}h^{-2}$, so that

$$\lambda_{\min}(\Delta_n^{-1}(m,k)A_n^*(a,m,k)) \ge \left[1 + C(\max a)(\min a)^{-1}\|\Theta_n(a,m,k)\|_2 + o(1)\right]^{-1}.$$

□

In addition, by making use of the following spectral characterization, the *Weakest Strong Clustering Property* can be proved.

LEMMA 5.5. *Let $\{\varepsilon_n\}_n$ be a sequence decreasing to zero (as slowly as wanted) and let us assume that the maximum order of zeros of the polynomial $p_{\mathbf{w}}(x) = |p_{\mathbf{c}}(x)|^2$ generating the Toeplitz sequence $\{\Delta_n(m,1)\}_n$ equals 2. Then,*

$$\#\left\{i : \lambda_i\left(\Delta_n(m,1)\right) < \left\lceil \varepsilon_n^{-1}\right\rceil h^2\right\} = O\left(\left\lceil \varepsilon_n^{-1/2}\right\rceil\right). \tag{5.1}$$

Proof. It is a simple consequence of Proposition B.1. □

THEOREM 5.6. *Let us consider $k = 1$ and any choice of m such that the maximal order of zeros of the polynomial $p_{\mathbf{w}}(x) = |p_{\mathbf{c}}(x)|^2$ generating the Toeplitz sequence $\{\Delta_n(m,1)\}_n$ equals 2. If the coefficient function $a(x)$ is strictly positive and belongs to $C^2(\overline{\Omega})$ then the* Weakest Strong Clustering Property *holds, i.e. for any sequence $\{\varepsilon_n\}_n$ decreasing to zero (as slowly as wanted), for each $\varepsilon > 0$ there exists $\bar{n}$ such that for any $n > \bar{n}$, then $n - O\left(\left\lceil \varepsilon_n^{-1/2}\right\rceil\right)$ eigenvalues of the preconditioned matrix $P_n^{-1}(a,m,k)A_n(a,m,k)$ belong to the open interval $(1-\varepsilon, 1+\varepsilon)$.*

Proof. By calling X_n the symmetric matrix

$$I_n + h^2 \Delta_n^{-1/2}(m,1)\Theta_n(a,m,k)\Delta_n^{-1/2}(m,1) + o(h^2)\Delta_n^{-1/2}(m,1)E_n(a,m,k)\Delta_n^{-1/2}(m,1)$$

similar to the matrix $\Delta_n^{-1}(m,1)A_n^*(a,m,1)$ and by considering the $n \times \left(n - O\left(\left\lceil \varepsilon_n^{-1/2} \right\rceil\right)\right)$ matrix U, whose columns are made up by considering the orthonormal eigenvectors of $\Delta_n(m,1)$ corresponding to the eigenvalues $\lambda_j(\Delta_n(m,1)) \geq \left\lceil \varepsilon_n^{-1} \right\rceil h^2$, we have

$$\begin{aligned} U^H X_n U &= I_{n-O\left(\left\lceil \varepsilon_n^{-1/2} \right\rceil\right)} \\ &\quad + h^2 \mathrm{diag}(\lambda_j^{-1/2}(\Delta_n(m,1)))U^H \Theta_n(a,m,k) U \mathrm{diag}(\lambda_j^{-1/2}(\Delta_n(m,1))) \\ &\quad + o(h^2)\mathrm{diag}(\lambda_j^{-1/2}(\Delta_n(m,1)))U^H E_n(a,m,k) U \mathrm{diag}(\lambda_j^{-1/2}(\Delta_n(m,1))) \\ &= I_{n-O\left(\left\lceil \varepsilon_n^{-1/2} \right\rceil\right)} + Y_n + Z_n. \end{aligned}$$

Now, since

$$\begin{aligned} \|Y_n\|_2 &\leq \lceil \varepsilon_n \rceil \|\Theta_n(a,m,k)\|_2, \\ \|Z_n\|_2 &\leq o(1) \lceil \varepsilon_n \rceil \|E_n(a,m,k)\|_2, \end{aligned}$$

where $\Theta_n(a,m,k)$ and $E_n(a,m,k)$ are bounded matrices according to Proposition 5.2, it follows that for each $\varepsilon > 0$ there exists $\bar{n}$ such that for any $n > \bar{n}$, then

$$-\varepsilon < \lambda_i(Y_n + Z_n) < \varepsilon,$$

or equivalently,

$$1 - \varepsilon < \lambda_i(U^H X_n U) < 1 + \varepsilon.$$

Lastly, by applying the Cauchy interlacing theorem [18], it directly follows that at least $n - O\left(\left\lceil \varepsilon_n^{-1/2} \right\rceil\right)$ eigenvalues of the preconditioned matrix $P_n^{-1}(a,m,k)A_n(a,m,1)$, similar to the matrix $\Delta_n^{-1}(m,1)A_n^*(a,m,1)$, belong to the open interval $(1-\varepsilon, 1+\varepsilon)$. □

Notice that, for $k = 1$ and $m = 1,2,3$ in Figure 4.3.a we observed that $x = 0$ is the unique zero of $p_{\mathbf{w}}(x)$ whose order equals 2 according to relation (2.9). Therefore, in light of Theorem 5.6 we deduce the *Weakest Strong Clustering Property*.

5.1.3. How many outliers?. In the case where $k > 1$ or $2s > 2$, we lose the *Weakest Strong Clustering Property*, although the *Weak Clustering Property* always holds according to Theorem 5.3. Consequently, the main task is the characterization of the goodness of the cluster; that is, for any positive fixed ε we want to know how many outliers do not belong to the interval $(1-\varepsilon, 1+\varepsilon)$.

THEOREM 5.7. *If $A_n^*(a,m,k) = \Delta_n(m,k) + O(h^t)$, with t positive real valued and the coefficient function $a(x)$ being strictly positive, then for each $\varepsilon_n = o(h^{1-\frac{t}{2s}})$ decreasing to zero and for each $\varepsilon > 0$ there exists $\bar{n}$ such that for each $n > \bar{n}$ at least $n - O(\lceil \varepsilon_n^{-1} \rceil)$ eigenvalues of $P_n^{-1}(a,m,k)A_n(a,m,k)$ fall in $(1-\varepsilon, 1+\varepsilon)$. Here, $2s$ denotes the maximal order of the zeros of the related Toeplitz generating polynomial $p_{\mathbf{w}}$, where $2s = 2k$ if $2k \geq q-1$ and otherwise belongs to $[2k, 2(q-1)-2k]$.*

Proof. We first observe that the i-th eigenvalue of $\Delta_n(m,k)$ behaves like $p_{\mathbf{w}}(x_i)$, $x_i = \pi i/(n+1)$ (refer to [5] and Proposition B.1). This implies that, for $i = i(n) = o(n)$, the i-th eigenvalue goes to zero as $(i(n)/n)^{2s}$. Since the matrix error $A_n^*(a,m,k) - \Delta_n(m,k) = O(h^t)$, it follows that, in any subspace generated by the eigenvectors $v_{i(n)}$ of $\Delta_n(m,k)$ related to eigenvalues $l_{i(n)}$ with

$$h^t = o\left([i(n)/n]^{2s}\right), \tag{5.2}$$

ETNA
Kent State University
etna@mcs.kent.edu

the Rayleigh quotient of $\Delta_n^{-1/2}(m,k)A_n^*(a,m,k)\Delta_n^{-1/2}(m,k) - I_n$ is infinitesimal. By referring to equation (5.2), this is equivalent to saying that for any $i(n)$ such that $(i(n))^{-1} = o(h^{1-\frac{t}{2s}})$, the preceding Rayleigh quotient is infinitesimal in the subspace generated by all the eigenvectors v_j, $j \geq i(n)$. Now, by setting $\varepsilon_n = (i(n))^{-1}$ and by invoking the Cauchy interlacing theorem [18], the proof is concluded. □

By using the general information on the error matrix $A_n^*(a,m,k) - \Delta_n(m,k)$ with spectral norm bounded by $O(h^t)$, we infer an estimate concerning the number of outlying eigenvalues as a function of t, but also depending on the "spectral difficulty" of the problem represented by the parameter k. In fact, the growth of the order $2k$ ($2s \geq 2k$) of the differential problem leads to a deterioration of the "strength" of the cluster. Therefore, in order to obtain a better clustering for higher order problems, it is necessary to increase the order of approximation of $\Delta_n(m,k)$ by $A_n^*(a,m,k)$; a possible proposal is the use of bidiagonal uniformly well conditioned matrices $\{B_n(a,m,k)\}_n$ in place of $\{D_n(a,m,k)\}_n$. A suitable choice of its entries can be used for obtaining a higher order of approximation in the difference $A_n^*(a,m,k) - \Delta_n(m,k)$ with $A_n^*(a,m,k) = B_n^{-1/2}(a,m,k)A_n(a,m,k)(B_n^{-1/2}(a,m,k))^T$.

It should be noticed that the growth of the number of outliers predicted by Theorem 5.7 for $P_n^{-1}(a,m,k)A_n(a,m,k)$ in the case where $k > 1$ and $a > 0$, is not observed in Table 4.5. It is probably possible to prove something more.

On the other hand, the improvement given by the *Strong Clustering Property* is just a theoretical one since, for the case where $a(x)$ is positive, regular and $k = 1$, the optimality of our PCG iterations follows from the fact that each eigenvalue of $P_n^{-1}(a,m,k)A_n(a,m,k)$ belongs to $[d_1, d_2]$ with d_i universal positive constants independent of n (Theorem 5.4).

5.2. The degenerate elliptic case. First, the following theorem explains why the Toeplitz sequence $\{\Delta_n(m,k)\}_n$ is not an optimal preconditioning sequence in the case where $a(x)$ has zeros. In fact, the condition defining the optimality in Definition 3.1 is violated.

THEOREM 5.8. *The spectra of the preconditioned matrix sequences* $\{\Delta_n^{-1}(m,k)A_n(a,m,k)\}_n$ *are contained in the interval* $[\bar{a}, \bar{A}]$, *with* $\bar{a}$ *and* $\bar{A}$ *being the infimum and the supremum of the function* $a(x)$ *respectively. If the coefficient function* $a(x)$ *has a finite number of zeros,* $\#I^+(a) \geq n$, $\inf a = 0$ *and the assumption of Theorem 2.4 is fulfilled, then the spectra of the preconditioned matrices are contained in the interval* $(0, \bar{A}]$. *In addition, the lower bound is tight in the sense that the smallest eigenvalue* $\lambda_{\min}(\Delta_n^{-1}(m,k)A_n(a,m,k))$ *of the preconditioned matrix tends to zero as* n *tends to infinity as long as* $\inf a = 0$.

Proof. The two localization results are immediate consequences of Theorem 2.3 and Theorem 2.4 respectively.

From Theorem 4.8 in [36] we know that the eigenvalues of $\{\Delta_n^{-1}(m,k)A_n(a,m,k)\}_n$ distribute as the function $a(x)$; a simple argument taken from measure theory implies that the minimal eigenvalue $\lambda_{\min}(\Delta_n^{-1}(m,k)A_n(a,m,k))$ tends to $\bar{a} = 0$.

Finally, by using the same argument of Theorem 4.1 in [31] based on special choices of the Rayleigh quotients, it follows that $\lambda_{\min}(\Delta_n^{-1}(m,k)A_n(a,m,k)) = O(h^{\alpha})$ if α is the maximal order of the zeros of $a(x)$. □

5.2.1. Clustering properties of preconditioned matrices. The reason for the very fast PCG convergence observed when we consider the preconditioner $P_n(a,m,k)$ with respect to the case of the basic Toeplitz preconditioner $\Delta_n(m,k)$ is explained in the following theorem.

THEOREM 5.9. *If the nonnegative coefficient function* $a(x)$ *has a finite number of zeros and belongs to* $C(\overline{\Omega})$, *then for any* $\varepsilon > 0$ *all the eigenvalues of the preconditioned matrix* $P_n^{-1}(a,m,k)A_n(a,m,k)$ *lie in the open interval* $(1-\varepsilon, 1+\varepsilon)$ *except for* $o(n)$ *outliers*

[Weak Clustering Property].

Proof. Due to the similarity between $P_n^{-1}(a,m,k)A_n(a,\ m,k)$ and $\Delta_n^{-1}(m,k)A_n^*(a, m,k)$, we can analyze the spectrum of the latter matrix. First, for the sake of simplicity, we consider the case when $a(x)$ has a unique zero of order α located at $x=0$.
For any index i such that $x=\tilde{x}_i$ is well separated from $x=0$, we find that

$$(A_n^*(a,m,k))_{i,i\pm p} = (\Delta_n(m,k))_{i,i\pm p} + (\Theta_n(a,m,k))_{i,i\pm p}, \quad p=0,\dots,q-1, \tag{5.3}$$

where $\|\Theta_n(a,m,k)\|_2 = O(\omega_a(h))$. More specifically, for any $\varepsilon>0$, let us consider the indices i such that the distance of the point $\tilde{x}_i$ from the point $x=0$ is greater or equal to ε. For all these indices i the relationship (5.3) holds true. Consequently, we can consider an asymptotic expansion

$$A_n^*(a,m,k) = \Delta_n(m,k) + \Theta_n(a,m,k,\varepsilon) + D_n(\varepsilon),$$

where $D_n(\varepsilon)$ is the null matrix except for the north-west corner of dimension εn. Clearly, the rank of D_n is bounded by εn and, by virtue of Theorem 2.6 and Proposition B.1, the sequence $\{\Delta_n(m,k)\}_n$ is sparsely vanishing.
Therefore, by setting $A_n = A_n^*(a,m,k)$ and $P_n = \Delta_n(m,k)$, the hypotheses of Lemma 3.6 are fulfilled and the claimed thesis follows.

Notice that the presence of a zero in a different position moves the position of the nonzero part of D_n along the diagonal, but does not change the size of its asymptotic rank. Moreover, the proof is unchanged in the case of the presence of a finite number of zeros. □

From Theorem 5.9, we deduce that almost all the eigenvalues of the preconditioned matrices are in a small neighbourhood of unity except for $o(n)$ outliers; this is not completely satisfactory but is nonetheless very good when compared with the Toeplitz preconditioning alone. In order to see this, consider the distributional results found in [36] where we prove that the number of PCG iterations to reach the solution within a fixed accuracy, $\Delta_n(m,k)$ being the preconditioner, is linear in the dimension n. Indeed, the latter is a consequence of the fact that the eigenvalues of the sequence $\{\Delta_n^{-1}(m,k)A_n(a,m,k)\}_n$ are distributed as the function $a(x)$ and of the fact that $a(x)$ has zeros.

Moreover, the behaviour in the numerical experiments is much better when compared with the quoted theoretical results. It is most likely that the analysis presented in Theorem 5.9 can be substantially refined. In particular, the theoretical tools introduced in [42] and [11] could be used in this context.

6. General results on distribution and clustering.

The aim of this section is to give general results on the approximation of the sequence $\{A_n^*(a,m,k)\}_n$ by the sequence $\{\Delta_n(m,k)\}_n$ in the spirit of the ergodic Theorems proved by Szegő, Widom, etc [19, 45]. Let us denote the trace norm by $||\cdot||_{trace}$ (refer to [4, Bhatia, p. 92]).

DEFINITION 6.1. [19, 43] *Two real sequences $\{a_i^{(n)}\}_{i\le n}$, $\{b_i^{(n)}\}_{i\le n}$ are equally distributed if and only if, for any real-valued continuous function F with bounded support, the following relation holds:*

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\left(F\left(a_i^{(n)}\right)-F\left(b_i^{(n)}\right)\right)=0. \tag{6.1}$$

When the previous limit goes to zero as $O(n^{-1})$ and F is Lipschitz continuous, we say that there is strong equal distribution.

THEOREM 6.2. *If the coefficient function $a(x)$ is strictly positive and belongs to $C(\overline{\Omega})$, then*

$$\begin{aligned}\|A_n^*(a,m,k) - \Delta_n(m,k)\|_F^2 &\le nO(\omega_a^2(n^{-1})),\\ \|A_n^*(a,m,k) - \Delta_n(m,k)\|_2 &\le O(\omega_a(n^{-1})),\\ \|A_n^*(a,m,k) - \Delta_n(m,k)\|_{\mathrm{trace}} &\le nO(\omega_a(n^{-1})),\end{aligned}$$

where ω_a denotes the modulus of continuity of the function $a(x)$. If the coefficient function $a(x)$ is nonnegative with a finite number r of zeros and belongs to $C(\overline{\Omega})$, then there exists a matrix sequence $\{D_n\}_n$, $D_n = D_n^{[1]} + \ldots + D_n^{[r]}$, with $\mathrm{rank}(D_n) = o(n)$, such that

$$\begin{aligned}\|A_n^*(a,m,k) - D_n - \Delta_n(m,k)\|_F^2 &= o(n),\\ \|A_n^*(a,m,k) - D_n - \Delta_n(m,k)\|_2 &= o(1).\end{aligned}$$

The latter theorem, in view of Tyrtyshnikov's results [43], tells us that the eigenvalues of the two sequences $\{A_n^*(a,m,k)\}_n$ and $\{\Delta_n(m,k)\}_n$ of symmetric matrices are equally distributed. But each matrix $\Delta_n(m,k)$ is the $n\times n$ Toeplitz matrix generated by $p_{\mathbf{w}}(x) = |p_{\mathbf{c}}(x)|^2$ and therefore, by taking into account the ergodic Szegő Theorem, it follows that for any real-valued continuous function F with bounded support

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n} F(\lambda_i(A_n^*(a,m,k))) = \frac{1}{2\pi}\int_{-\pi}^{\pi} F(p_{\mathbf{w}}(x))dx. \tag{6.2}$$

Moreover, we have also the following Widom–like second order result.

COROLLARY 6.3. *If the coefficient function $a(x)$ is Lipschitz-continuous, then the sequences $\{A_n^*(a,m,k)\}_n$ and $\{\Delta_n(m,k)\}_n$ are strongly equally distributed and for any real-valued Lipschitz-continuous function F with bounded support we find*

$$\left|\frac{1}{n}\sum_{i=1}^{n} F(\lambda_i(A_n^*(a,m,k)) - \frac{1}{2\pi}\int_{-\pi}^{\pi} F(p_{\mathbf{w}}(x))dx\right| = O(n^{-1}). \tag{6.3}$$

Proof. If $a(x)$ is Lipschitz-continuous, then $nO(\omega_a(n^{-1})) = O(1)$. Consequently, from the latter theorem we deduce that

$$\|A_n^*(a,m,k) - \Delta_n(m,k)\|_{\mathrm{trace}} = O(1).$$

The application of the third part of Lemma 4.3 in [34] yields the strong equal distribution of $\{A_n^*(a,m,k)\}_n$ and $\{\Delta_n(m,k)\}_n$. Moreover, we remark that the generating function of the Toeplitz matrices $\{\Delta_n(m,k)\}_n$ is a trigonometric polynomial and we simply point out that all the trigonometric polynomials are from the Krein algebra $\mathcal{K}$ [25]. Finally, we use the second-order result of Widom [45] concerning the spectral distribution of Toeplitz matrices with symbol belonging to $\mathcal{K}$ and this concludes the proof. ☐

Notice that Theorem 6.2 can be extended in the same way to the case of p dimensions, i.e. to differential problems on p dimentional domains (recall that the Szegő - Tyrtyshnikov Theorem holds generically in p dimensions).

The fact $A_n^*(a,m,k) = \Delta_n(m,k) + O(h^2)$ proved in Propositions A.1, A.2 and 5.2, with $a(x) \in C^2(\overline{\Omega})$, is in some sense exceptional because it is produced by the cancellation of $O(h)$ terms in the expression $(A_n^*(a,m,k) - \Delta_n(m,k))_{r,r\pm p}$. Moreover, if the coefficient function $a(x)$ is strictly positive and belongs to $C^1(\overline{\Omega})$ and $a_x(x) \in Lip_1$, then

$A_n^*(a,m,k) = \Delta_n(m,k) + O(h^2)$. Nevertheless, it is worth pointing out that equation (1.1) imposes that the coefficient function $a(x)$ belongs to $C^k(\overline{\Omega})$, so it would appear that a refined analysis be just an academic exercise. However, when we consider the "weak formulation" [9], the problem (1.1) is transformed into an integral problem. Therefore, in this sense, the given analysis becomes again meaningful. Concerning this fact, we are able to prove something more. If $a(x) \in L^\infty(\Omega)$, the application of the Lusin Theorem allows one to prove the following result.

THEOREM 6.4. *Let $A_n^*(a,m,k) = D_n^{-1/2}(a,m,k)A_n(a,m,k)D_n^{-1/2}(a,m,k)$ be the symmetrically scaled matrix of $A_n(a,m,k)$, FD discretization matrix of the continuous problem* (1.1) *and let $\Delta_n(m,k)$ be the related Toeplitz matrix. Here the coefficients $a(x_i)$ should be replaced by mean values on the interval $I_i = [x_i, x_{i+1}]$ in the sense that $a(x_i)$ means $n\int_{I_i} a(t)$. Then, when the coefficient function $a(x)$ is nonnegative and, at most, sparsely vanishing and belongs to $L^\infty(\Omega)$, there exists a matrix sequence $\{D_n\}_n$, with* $\mathrm{rank}(D_n) = o(n)$*, such that*

$$\|A_n^*(a,m,k) - \Delta_n(m,k) - D_n\|_F^2 = o(n),$$
$$\|A_n^*(a,m,k) - \Delta_n(m,k) - D_n\|_2 = o(1).$$

In addition, the number of outliers of the sequence of preconditioned matrices $\{P_n^{-1}(a,m,k)A_n(a,m,k)\}_n$ is generically $o(n)$, while if $a(x)$ is not sparsely vanishing then the preconditioners $P_n(a,m,k)$ are not well defined or the spectra of the preconditioned matrices are not clustered.

Proof. This is a simple generalization of Theorem 5.3 in [34]. □

7. Computational costs and comparison with the literature.

7.1. Computational costs. In conclusion, we have reduced the asymptotic cost of these band systems (which are Locally Toeplitz [41]) to the cost of band-Toeplitz systems for which the recent literature provides very sophisticated algorithms (for the most efficient see [6]) whose cost is lower than that of the classical band solvers [18]: (**a**) Multigrid methods requiring $O(ln)$ arithmetic operations (ops) and $O(\log n)$ parallel steps with $O(ln)$ processors [13, 14] in the parallel PRAM model of computation; (**b**) a recursive displacement rank based technique [6] requiring $O(n\log(l) + l\log^2(l)\,\log(n))$ ops and $O(\log n)$ parallel steps with $O(ln)$ processors. Here n is the matrix dimension and $l = 2q - 1$ is the matrix bandwidth.

So, in order to obtain the total computational cost needed to compute the solution of a system $A_n\mathbf{u} = \mathbf{f}$, with $A_n = A_n(a,m,k)$, using the PCG method, the preceding costs must be multiplied by the PCG iteration number which is constant with respect to n and added to the cost of few matrix-vector multiplications (recall the PCG algorithm). The final cost is $O(n\log(l))$ ops and $O(\log(nl))$ parallel steps with $O(ln)$ processors in the PRAM model of computation.

Concerning the computational cost in the $2D$ case we point out that the "displacement rank technique" developed in [6] is no longer "optimal" in the multilevel Toeplitz case (i.e. in the case where a Toeplitz matrix has block Toeplitz structure and each block has Toeplitz structure and so on recursively for a finite number $d \geq 2$ of levels). The same remark also holds for the classical band-solvers [18]. Nevertheless, in the multilevel case optimal iterative solvers are those based on the multigrid methods [14] or on mixed methods (PCG + multigrid) [33]; with the latter multigrid-type choices, the computational cost is linear as the size $N(n)$ of the linear system and linear as the sum of the internal and external bandwidths of the coefficient matrix.

7.2. Comparison with the literature. Hereafter, we compare our technique with the iterative and direct solvers used in the relevant literature. In the following N denotes the dimension of the algebraic system, that is, $N = n$ for the $1D$ case and $N = N(n) = n_1 n_2$ for the $2D$ case with $n_1 \sim n_2$. The case under study is that of boundary value problems such as (1.1) and (1.2) over a rectangle Ω of $\mathbf{R}^d$, $d = 1$ or $d = 2$, discretized by uniformly spaced FD schemes. We analyze the following three situations:
a.1) $a(x) > 0$ and $a(x) \in C^2(\overline{\Omega})$,
a.2) $a(x) > 0$ and $0 < \inf a(x) \leq \sup a(x) < \infty$,
a.3) $a(x) \geq 0$ with at most isolated zeros and $a(x)$ belonging to $L^1(\Omega)$.

The PCG methods based on preconditioners from incomplete LU factorizations [27, 10, 20] (for $d = 2$) and from the circulant algebra [8, 22, 26] (for $d = 1$ and $d = 2$) are sublinear, i.e. they require a number of iterations $O(N^\beta)$ with positive β and an overall cost of at least $O(N^{1+\beta})$. This is true even in the case **a.1** where $a(x)$ is positive and smooth. In particular, the Strang preconditioner is singular when N is big enough due to the consistency condition and the T. Chan preconditioner suggested in [8] leads to preconditioned matrix sequences whose condition number grows at least as $N^{(2k-1)/d}$, where d is the dimension of the definition domain and, in the simplest case, where $a(x)$ is equal to a positive constant. As a consequence, it is clear that the value of β is $\min\{1, (2k-1)/2d\}$. Moreover, in light of the analysis given in [11], there is a sub-cluster of eigenvalues to zero and this leads to a substantial slowdown in the performances of the associated PCG method.

On the other hand, the PCG methods defined by using separable preconditioners [12] and the multigrid algorithms [21, 3] are optimal in the sense of Definition 3.1 in the cases **a.1** and **a.2**, but not in the case **a.3**. On the contrary, our technique is superlinear in the case **a.1** (therefore also optimal) and assures a "weak" clustering in the cases **a.2** and **a.3**; this property does not theoretically guarantee optimality, but the numerous numerical experiments performed here and in [34, 36] suggest a convergence rate independent of the size N. Furthermore, we recall that the "weak" clustering property also holds in the case where $a \in L^1(\Omega)$.

Finally, some remarks concerning direct methods are needed. In the $1D$ case it is clear that Gaussian elimination is also optimal with regard to the dimension n and with a constant growing as l^2, where $l = 2q - 1$ is the matrix bandwidth. However, the matrices $A_n(a, m, k)$ are very ill-conditioned if k is large or if $a(x)$ has zeros so that possible numerical instabilities can be observed and the proposed PCG technique can be a good alternative method. The advantage of our proposal is especially evident in the $2D$ case where Gaussian elimination is no longer optimal requiring $O(N^2)$ ops against $O(N)$ of our PCG method. Moreover, this advantage is stronger when the number of dimensions d increases ($O\left(N^{2(d-1)/d+1}\right)$ ops against $O(N)$).

Finally, concerning the matrix algebra approach, it is useful to recall a negative result stated in [39, 40] regarding multilevel structures in d dimensions for $d \geq 2$: at least $cN^{(d-1)/d}$ outliers with $c > 0$ are present when we use a matrix algebra preconditioning sequence and the matrix algebra is "partially equimodular" [40]. We recall that all the known trigonometric, Hartley and ω circulant algebras ($|\omega| = 1$) are "partially equimodular". The bound of at least $cN^{(d-1)/d}$ outliers is realized by both Strang and T. Chan preconditioners so that important information contained in the negative result is that their unsatisfactory performances cannot be substantially improved by changing preconditioning sequence in any algebra of "partially equimodular" type.

8. Concluding remarks. To conclude, in this paper we have discussed the asymptotic distributional properties of the spectra of Toeplitz–based preconditioned matrices arising from FD discretization of the differential problems of the form (1.1). We have proved that the weak clustering of the spectra holds for $a(x)$ ranging from the good case in which it is

ETNA
Kent State University
etna@mcs.kent.edu

regular and strictly positive to the bad case where $a(x)$ is only $L^\infty(\Omega)$ and sparsely vanishing. Moreover, the results indicate that a possible deterioration of the convergence properties of the associated PCG methods occurs when the parameter k and/or the order of the zeros of $a(x)$ increases.

Appendix A. Asymptotic expansions of $A_n(a, m, k)$. This section is devoted to the evaluation of the asymptotic expansion of the matrix $A_n(a, m, k)$ with respect to the coefficient function $a(x)$, this being the basic step in the analysis of the clustering properties of our Toeplitz based preconditioners. Let Δ be the constant entry along the main diagonal and let Δ_{r-p} be the constant entry along the p^{th} subdiagonal in the Toeplitz matrix $\Delta_n(m, k)$.

PROPOSITION A.1. *Let k be odd and let $A_n(a, m, k)$ be the $n \times n$ symmetric band matrix according to Definition 2.2. If the coefficient function $a(x)$ belongs to $C^2(\overline{\Omega})$, then the following asymptotic expansions hold true*

$$\begin{aligned} (A_n)_{r,r} &= a_{r-p/2}\Delta + h a_{x,r-p/2}\beta_p + h^2 a_{xx,r-p/2}\gamma_p + o(h^2), \\ (A_n)_{r-p,r-p} &= a_{r-p/2}\Delta - h a_{x,r-p/2}\beta_p + h^2 a_{xx,r-p/2}\gamma_p + o(h^2), \\ (A_n)_{r,r-p} &= a_{r-p/2}\Delta_{r-p} + h^2 a_{xx,r-p/2}\delta_p + o(h^2), \end{aligned}$$

where $p = 1, \ldots, 2m-1$, $a_{r-p/2} = a\left(x_{r-p/2}\right)$, $a_{x,r-p/2} = (da(x)/dx)_{|x=x_{r-p/2}}$, $a_{xx,r-p/2} = (d^2a(x)/dx^2)_{|x=x_{r-p/2}}$, and β_p, γ_p and δ_p are constant numbers. In the case where the coefficient function $a(x)$ belongs to $C^1(\overline{\Omega})$, then the related expansion takes the form

$$\begin{aligned} (A_n)_{r,r} &= a_{r-p/2}\Delta + h a_{x,r-p/2}\beta_p + O(h\omega_{a_x}(h)), \\ (A_n)_{r-p,r-p} &= a_{r-p/2}\Delta - h a_{x,r-p/2}\beta_p + O(h\omega_{a_x}(h)), \\ (A_n)_{r,r-p} &= a_{r-p/2}\Delta_{r-p} + O(h\omega_{a_x}(h)). \end{aligned}$$

Finally, if $a(x)$ belongs to $C(\overline{\Omega})$ we find that

$$\begin{aligned} (A_n)_{r,r} &= a_{r-p/2}\Delta + O(\omega_a(h)), \\ (A_n)_{r-p,r-p} &= a_{r-p/2}\Delta + O(\omega_a(h)), \\ (A_n)_{r,r-p} &= a_{r-p/2}\Delta_{r-p} + O(\omega_a(h)). \end{aligned}$$

Here the symbol $\omega_f(\cdot)$ denotes the modulus of continuity of the function f.

Proof. We consider the Taylor expansions centered at $x = x_{r-p/2}$ with respect to the coefficient function $a(x)$. According to (2.2) we have the main diagonal entry in the r^{th} row given by

$$\begin{aligned} (A_n)_{r,r} &= \sum_{j=1}^{m} \left(2a_{r-p/2} + h a_{x,r-p/2} p + \frac{h^2}{2} a_{xx,r-p/2} \left(\left(j + \frac{(p-1)}{2} \right)^2 \right.\right. \\ &\qquad \left.\left. + \left(j - \frac{(p+1)}{2} \right)^2 \right) \right) c_j^2 + o(h^2) \\ &= a_{r-p/2}\Delta + h a_{x,r-p/2}\beta_p + h^2 a_{xx,r-p/2}\gamma_p + o(h^2). \end{aligned}$$

ETNA
Kent State University
etna@mcs.kent.edu

As in the previous case,

$$
\begin{aligned}
(A_n)_{r-p,r-p} &= \sum_{j=1}^{m} \left(a\left(x_{r-p-j+1/2}\right) + a\left(x_{r-p+j-1/2}\right)\right) c_j^2 \\
&= \sum_{j=1}^{m} \Bigg(2a_{r-p/2} - ha_{x,r-p/2}p + \frac{h^2}{2} a_{xx,r-p/2} \left(\left(j + \frac{(p-1)}{2}\right)^2 \right. \\
&\quad \left. + \left(j - \frac{(p+1)}{2}\right)^2\right)\Bigg) c_j^2 + o(h^2) \\
&= a_{r-p/2}\Delta - ha_{x,r-p/2}\beta_p + h^2 a_{xx,r-p/2}\gamma_p + o(h^2).
\end{aligned}
$$

Now, for p ranging from 1 to $m-1$, according to equation (2.3), we have

$$
\begin{aligned}
(A_n)_{r,r-p} = a_{r-p/2}\Delta_{r-p} &+ ha_{x,r-p/2} \left(\sum_{j=1}^{p} \left(j - (p+1)/2\right) c_j c_{p+1-j}\right) + \\
&+ \frac{h^2}{2} a_{xx,r-p/2} \left(\sum_{j=1}^{m-p} \left((j + (p-1)/2)^2 + (j-(p+1)/2)^2\right) c_j c_{j+p} \right. \\
&\left. - \sum_{j=1}^{p} \left(j-(p+1)/2\right)^2 c_j c_{p+1-j}\right) + o(h^2) \\
&= a_{r-p/2}\Delta_{r-p} + h^2 a_{xx,r-p/2}\delta_p + o(h^2),
\end{aligned}
$$

since, for each $p = 1, \ldots, m-1$,

$$
\begin{aligned}
&\sum_{j=1}^{p} \left(j - \frac{(p+1)}{2}\right) c_j c_{p+1-j} = \\
&\quad \sum_{j=1}^{\lfloor p/2 \rfloor} \left(j - \frac{(p+1)}{2}\right) c_j c_{p+1-j} + \sum_{j=\lceil p/2 \rceil + 1}^{p} \left(j - \frac{(p+1)}{2}\right) c_j c_{p+1-j} \\
&\quad = \sum_{j=1}^{\lfloor p/2 \rfloor} \left(j - \frac{(p+1)}{2}\right) c_j c_{p+1-j} - \sum_{j=1}^{\lfloor p/2 \rfloor} \left(j - \frac{(p+1)}{2}\right) c_j c_{p+1-j} = 0.
\end{aligned}
$$

For p ranging from m to $2m-1$, according to equation (2.4), we have

$$
\begin{aligned}
(A_n)_{r,r-p} = -&\left(\sum_{j=p+1-m}^{\lfloor p/2 \rfloor} a\left(x_{r-j+1/2}\right) c_j c_{p+1-j} + F_{\mathbb{N}}\left((p+1)/2\right) a\left(x_{r-p/2}\right) c_{(p+1)/2}^2 \right. \\
&\left. + \sum_{j=\lceil p/2 \rceil + 1}^{m} a\left(x_{r-j+1/2}\right) c_j c_{p+1-j}\right) \\
&= -\sum_{j=1}^{m-\lceil p/2 \rceil} \left(a\left(x_{r-j-p+m+1/2}\right) + a\left(x_{r+j-m-1/2}\right)\right) c_{m+1-j} c_{p+j-m} + \\
&\quad -F_{\mathbb{N}}\left((p+1)/2\right) a\left(x_{r-p/2}\right) c_{(p+1)/2}^2 \\
&= a_{r-p/2}\Delta_{r-p} + h^2 a_{xx,r-p/2}\delta_p + o(h^2),
\end{aligned}
$$

where $F_{\mathbb{N}}$ is the characteristic function over the set of integer numbers. Notice that, for $p = 2m-1$, we simply infer that $(A_n)_{r,r-(2m-1)} = a_{r-m+1/2}\Delta_{r-(2m-1)}$.

Finally, when the function $a(x)$ shows less regularity, it is enough to stop the preceding asymptotic expansions at a lower degree. □

PROPOSITION A.2. *Let k be even and let $A_n(a,m,k)$ be the $n \times n$ symmetric band matrix according to Definition 2.2. If the coefficient function $a(x)$ belongs to $C^2(\overline{\Omega})$, then the following asymptotic expansions hold true*

$$\begin{aligned}
(A_n)_{r,r} &= a_{r-p/2}\Delta + h a_{x,r-p/2}\tilde{\beta}_p + h^2 a_{xx,r-p/2}\tilde{\gamma}_p + o(h^2),\\
(A_n)_{r-p,r-p} &= a_{r-p/2}\Delta - h a_{x,r-p/2}\tilde{\beta}_p + h^2 a_{xx,r-p/2}\tilde{\gamma}_p + o(h^2),\\
(A_n)_{r,r-p} &= a_{r-p/2}\Delta_{r-p} + h^2 a_{xx,r-p/2}\tilde{\delta}_p + o(h^2),
\end{aligned}$$

where $p = 1,\ldots,2m$ and where $a_{r-p/2} = a\left(x_{r-p/2}\right)$, $a_{x,r-p/2} = (da(x)/dx)_{|x=x_{r-p/2}}$, $a_{xx,r-p/2} = (d^2a(x)/dx^2)_{|x=x_{r-p/2}}$, and $\tilde{\beta}_p, \tilde{\gamma}_p$ and $\tilde{\delta}_p$ are constant numbers. If $a(x)$ belongs to $C^1(\overline{\Omega})$, then the related expansion takes the form

$$\begin{aligned}
(A_n)_{r,r} &= a_{r-p/2}\Delta + h a_{x,r-p/2}\tilde{\beta}_p + O(h\omega_{a_x}(h)),\\
(A_n)_{r-p,r-p} &= a_{r-p/2}\Delta - h a_{x,r-p/2}\tilde{\beta}_p + O(h\omega_{a_x}(h)),\\
(A_n)_{r,r-p} &= a_{r-p/2}\Delta_{r-p} + O(h\omega_{a_x}(h)).
\end{aligned}$$

Finally, if $a(x)$ just belongs to $C(\overline{\Omega})$, then we have

$$\begin{aligned}
(A_n)_{r,r} &= a_{r-p/2}\Delta + O(\omega_a(h)),\\
(A_n)_{r-p,r-p} &= a_{r-p/2}\Delta + O(\omega_a(h)),\\
(A_n)_{r,r-p} &= a_{r-p/2}\Delta_{r-p} + O(\omega_a(h)).
\end{aligned}$$

Proof. We consider the Taylor expansions centered at $x = x_{r-p/2}$. According to equation (2.5), the main diagonal entry in the r^{th} row is given by

$$\begin{aligned}
(A_n)_{r,r} &= a_{r-p/2}\Delta + h a_{x,r-p/2}\left(pc_0^2/2 + p\sum_{j=1}^{m} c_j^2\right) +\\
&\quad + \frac{h^2}{2} a_{xx,r-p/2}\left(p^2c_0^2/4 + \sum_{j=1}^{m}\left((j+p/2)^2 + (j-p/2)^2\right)c_j^2\right) + o(h^2)\\
&= a_{r-p/2}\Delta + h a_{x,r-p/2}\tilde{\beta}_p + h^2 a_{xx,r-p/2}\tilde{\gamma}_p + o(h^2).
\end{aligned}$$

As in the previous case, we find that

$$\begin{aligned}
(A_n)_{r-p,r-p} &= a(x_{r-p})c_0^2 + \sum_{j=1}^{m}\left(a\left(x_{r-p-j}\right) + a\left(x_{r-p+j}\right)\right)c_j^2\\
&= a_{r-p/2}\Delta - h a_{x,r-p/2}\left(pc_0^2/2 + p\sum_{j=1}^{m} c_j^2\right) +\\
&\quad + \frac{h^2}{2} a_{xx,r-p/2}\left(p^2c_0^2/4 + \sum_{j=1}^{m}\left((j+p/2)^2 + (j-p/2)^2\right)c_j^2\right) + o(h^2)\\
&= a_{r-p/2}\Delta - h a_{x,r-p/2}\tilde{\beta}_p + h^2 a_{xx,r-p/2}\tilde{\gamma}_p + o(h^2).
\end{aligned}$$

With respect to the coefficients $(A_n)_{r,r-p}$, we must distinguish between the case of p ranging from 1 to $m-1$ and the case of p ranging from m to $2m$. More precisely, for p ranging from 1 to $m-1$, according to equation (2.6), we have

$$\begin{aligned}(A_n)_{r,r-p} &= a_{r-p/2}\Delta_{r-p} - h a_{x,r-p/2}\left(\sum_{j=0}^{p}\left(j-\frac{p}{2}\right)c_j c_{p-j}\right) + \\ &\quad + \frac{h^2}{2}a_{xx,r-p/2}\left(\sum_{j=0}^{p}\left(j-\frac{p}{2}\right)^2 c_j c_{p-j} + 2\sum_{j=1}^{m-p}\left(j+\frac{p}{2}\right)^2 c_j c_{p+j}\right) + o(h^2) \\ &= a_{r-p/2}\Delta_{r-p} + h^2 a_{xx,r-p/2}\tilde{\delta}_p + o(h^2),\end{aligned}$$

since, for each $p = 1,\ldots,m-1$,

$$\begin{aligned}\sum_{j=0}^{p}(j-p/2)\,c_j c_{p-j} &= \sum_{j=0}^{\lceil p/2\rceil-1}(j-p/2)\,c_j c_{p-j} + \sum_{\lfloor p/2\rfloor+1}^{p}(j-p/2)\,c_j c_{p-j} \\ &= \sum_{j=0}^{\lceil p/2\rceil-1}(j-p/2)\,c_j c_{p-j} - \sum_{j=0}^{\lceil p/2\rceil-1}(j-p/2)\,c_j c_{p-j} = 0.\end{aligned}$$

For p ranging from m to $2m$, according to equation (2.7), we obtain that

$$\begin{aligned}(A_n)_{r,r-p} &= \sum_{j=p-m}^{\lceil p/2\rceil-1} a\left(x_{r-j}\right)c_j c_{p-j} + F_{\mathbb{N}}\left(p/2\right)a\left(x_{r-p/2}\right)c_{p/2}^2 + \sum_{j=\lfloor p/2\rfloor+1}^{m} a\left(x_{r-j}\right)c_j c_{p-j} \\ &= \sum_{j=1}^{m-\lfloor p/2\rfloor}\left(a\left(x_{r+m+1-j-p}\right) + a\left(x_{r+j-m-1}\right)\right)c_{m+1-j}c_{p+j-m-1} \\ &\quad + F_{\mathbb{N}}\left(p/2\right)a\left(x_{r-p/2}\right)c_{p/2}^2 \\ &= a_{r-p/2}\Delta_{r-p} + h^2 a_{xx,r-p/2}\tilde{\delta}_p + o(h^2),\end{aligned}$$

where $F_{\mathbb{N}}$ is the characteristic function over the set of integer numbers. Notice that, for $p = 2m$, we simply deduce that $(A_n)_{r,r-2m} = a_{r-m}\Delta_{r-2m}$.

In the case of a lower degree of regularity of the function $a(x)$, the preceeding asymptotic expansions are stopped at a lower degree. □

Appendix B. A second order spectral result.

PROPOSITION B.1. *Let p be an even trigonometric polynomial. Let $\{T_n(p)\}_n$ be the related family of Toeplitz matrices generated by p. Let us suppose that p is nonnegative and not identically zero. Then the sequence $\{T_n(p)\}_n$ is* sparsely vanishing*, i.e. for any $\varepsilon > 0$ it holds that*

$$\lim_{\varepsilon\to 0}\lim_{n\to\infty}\frac{1}{n}\#\left\{i:\ \lambda_i(T_n(p)) < \varepsilon\right\} = 0,$$

with $\lambda_i(X_n)$ denoting the i-th eigenvalue of the $n\times n$ matrix X_n. In particular, if q is the degree with regard to the cosine expansion of p ($2q$ being the degree as ordinary complex polynomial) and if the maximal order of the zeros is $2 \le 2s \le 2q$, then we find the relation

$$\frac{1}{n}\#\left\{i:\ \lambda_i(T_n(p)) < \varepsilon\right\} - \frac{1}{\pi}\cdot\mu\{x\in[0,\pi]:\ p(x) < \varepsilon\} = O(n^{-1}),$$

with

$$\mu\{x \in [0,\pi] :\ p(x) < \varepsilon\} \sim \varepsilon^{\frac{1}{2s}}.$$

Proof. Let τ be the algebra of all the matrices simultaneously diagonalized by sine transforms [5]. Let $\tau(T_n(p))$ be the canonical τ representation of $T_n(p)$ (see [17, 13] for details). Then we have $\tau(T_n(p)) = T_n(p) - H_n(p)$, where $H_n(p)$ is the persymmetric Hankel matrix whose first row is $(a_2, \ldots, a_q, 0, \ldots, 0)$, $(a_0, \ldots, a_q, 0, \ldots, 0)$ being the first row of $T_n(p)$. Therefore, according to [5, 13], the eigenvalues of $\tau(T_n(p))$ are $\hat{\lambda}_i^{(n)}(p) = p(x_i^{(n)})$, $x_i^{(n)} = \pi i/(n+1)$ and $\mathrm{rank}(H_n(p)) \le 2(q-1)$. Due to the fact that $p \in C^1$, it follows that

$$\#\{i :\ \hat{\lambda}_i^{(n)}(p) < \varepsilon\} = \frac{n}{\pi} \cdot \mu\{x \in [0,\pi] :\ p(x) < \varepsilon\} + O(1),$$

where μ denotes the Lebesgue measure in $\mathbb{R}$. Since $\{H_n\}_n$ have rank uniformly bounded by a constant, by the Cauchy interlacing theorem [18] we infer that

$$\#\{i :\ \lambda_i(T_n(p)) < \varepsilon\} = \frac{n}{\pi} \cdot \mu\{x \in [0,\pi] :\ p(x) < \varepsilon\} + O(1),$$

where the term $O(1)$ now contains a factor proportional to q. Therefore, we find that

$$\lim_{n\to\infty} \frac{1}{n} \#\{i :\ \lambda_i(T_n(p)) < \varepsilon\} = \frac{1}{\pi} \mu\{x \in [0,\pi] :\ p(x) < \varepsilon\}.$$

In particular, if the maximal order of the zeros of p is $2s$ ($2 \le 2s \le 2q$) a simple analytic argument shows that

$$\mu\{x \in [0,\pi] :\ p(x) < \varepsilon\} \sim \varepsilon^{\frac{1}{2s}},$$

and, since $\varepsilon^{1/2s}$ goes to zero as ε tends to zero, the proof is concluded. □

REFERENCES

[1] O. Axelsson and V. Barker, *Finite Element Solution of Boundary Value Problems, Theory and Computation*, Academic Press Inc., New York, 1984.
[2] O. Axelsson and G. Lindskog, *On the rate of convergence of the preconditioned conjugate gradient method*, Numer. Math., 48 (1986), pp. 499–523.
[3] O. Axelsson and M. Neytcheva, *The algebraic multilevel iteration methods – theory and applications*, Proc. 2nd Internat. Colloq. on Numerical Analysis, D. Bainov Ed., Plovdiv (Bulgaria), August 1993, pp. 13–23.
[4] R. Bhatia, *Matrix Analysis*, Springer-Verlag, New York, 1997.
[5] D. Bini and M. Capovani, *Spectral and computational properties of band symmetric Toeplitz matrices*, Linear Algebra Appl., 52/53 (1983), pp. 99–126.
[6] D. Bini and B. Meini, *Effective methods for solving banded Toeplitz systems*, SIAM J. Matrix Anal. Appl., 20-3 (1999), pp. 700–719.
[7] A. Böttcher and S. Grudsky, *On the condition numbers of large semi-definite Toeplitz matrices*, Linear Algebra Appl., 279 (1998), pp. 285–301.
[8] R. Chan and T. Chan, *Circulant preconditioners for elliptic problems*, J. Numer. Linear Algebra Appl., 1 (1992), pp. 77–101.
[9] P. Ciarlet, *The Finite Element Method for Elliptic Problems*, North-Holland Publishing Co., Amsterdam, 1978.
[10] P. Concus, G. Golub, and G. Meurant, *Block preconditioning for the conjugate gradient method*, TR nr. LBL-14856, Lawrence-Berkeley Laboratory, UCLA, USA, (1982).
[11] F. Di Benedetto and S. Serra Capizzano, *A unifying approach to matrix algebra preconditioning*, Numer. Math., 82-1 (1999), pp. 57–90.

[12] H. ELMAN AND M. SHULTZ, *Preconditioning by fast direct methods for nonself-adjoint nonseparable elliptic equations*, SIAM J. Numer. Anal., 23 (1986), pp. 44–57.

[13] G. FIORENTINO AND S. SERRA, *Multigrid methods for Toeplitz matrices*, Calcolo, 28 (1991), pp. 283–305.

[14] ———*Multigrid methods for symmetric positive definite block Toeplitz matrices with nonnegative generating functions*, SIAM J. Sci. Comput., 17 (1996), pp. 1068–1081.

[15] ———*Tau preconditioners for elliptic problems*, TR nr. 50, Dept. of Mathematics - Univ. of Milano, (1994).

[16] ———*Tau preconditioners for (high order) elliptic problems*, Proc. 2nd IMACS conference on Iterative Methods in Linear Algebra, Vassilevski ed., Blagoevgrad (Bulgaria), June 1995, pp. 241–252.

[17] ———*Fast parallel solvers for elliptic problems*, Comput. Math. Appl., 32 (1996), pp. 61–68.

[18] G. GOLUB AND C. VAN LOAN, *Matrix Computations*, The Johns Hopkins Univ. Press, Baltimore, 1983.

[19] U. GRENANDER AND G. SZEGŐ, *Toeplitz Forms and Their Applications*, Second Edition, Chelsea Publishing Co., New York, 1984.

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

[21] W. HACKBUSH, *Multigrid Methods and Applications*, Springer-Verlag, Berlin, Germany, 1985.

[22] S. HOLMGREN AND K. OTTO, *Iterative solution methods and preconditioners for block-tridiagonal systems of equations*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 863–886.

[23] C. JOHNSON, *Numerical Solutions of Partial Differential Equations by the Finite Elements Methods*, Cambridge Univ. Press, Cambridge, UK, 1988.

[24] M. KAC, W.L. MURDOCH, AND G.SZEGŐ, *On the eigenvalues of certain Hermitian forms*, J. Rational Mech. Anal., 2 (1953), pp. 767–800.

[25] M.G. KREIN, *On some new Banach algebras and Wiener-Levy theorems for Fourier Series and integrals*, Amer. Math. Soc. Transl., 93 (1970), pp. 177–199.

[26] I. LIRKOV, S. MARGENOV, AND P. VASSILEVSKY, *Circulant block factorization for elliptic problems*, Computing, 53 (1994), pp. 59–74.

[27] J. MEIJERINK AND H. VAN DER VORST, *An iterative solution method for linear systems of which the coefficient matrix is a symmetric M-matrix*, Math. Comp., 31 (1977), pp. 148–162.

[28] ———*Guidelines for the usage of incomplete decompositions in solving sets of linear equations as they occur in practical problems*, J. Comput. Phys., 14 (1981), pp. 134–155.

[29] S. PARTER, *On the extreme eigenvalues of truncated Toeplitz matrices*, Bull. Amer. Math. Soc., 67 (1961), pp. 191–196.

[30] W. RUDIN, *Real and Complex Analysis*, McGraw-Hill Book Co., New York, 1985.

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

[32] ———*On the extreme eigenvalues of Hermitian (block) Toeplitz matrices*, Linear Algebra Appl., 270 (1998), pp. 109–129.

[33] ———*Preconditioning strategies for asymptotically ill–conditioned block Toeplitz systems*, BIT, 34 (1994), pp. 579–594.

[34] ———*Spectral analysis of Toeplitz based preconditioned matrices for boundary value problems*, TR nr. 31, LAN - Dept. of Mathematics - Univ. of Calabria, (1998).

[35] ———*Optimal, quasi-optimal and superlinear band-Toeplitz preconditioners for asymptotically ill-conditioned positive definite Toeplitz systems*, Math. Comp., 66 (1997), pp. 651–665.

[36] S. SERRA CAPIZZANO AND C. TABLINO POSSIO, *Spectral and structural analysis of high precision Finite Difference matrices for elliptic Operators*, Linear Algebra Appl. 293 (1999), pp. 85–131.

[37] ———*High–precision Finite Difference schemes and Toeplitz based preconditioners for Elliptic Problems*, Tr. 1-1999, Dipartimento di Scienza dei Materiali, Università di Milano Bicocca.

[38] ———*Preconditioning strategies for 2D Finite Difference matrix sequences*, Tr. 2-1999, Dipartimento di Scienza dei Materiali, Università di Milano Bicocca, submitted.

[39] S. SERRA CAPIZZANO AND E. TYRTYSHNIKOV, *Any circulant-like preconditioner for multilevel matrices is not superlinear*, SIAM J. Matrix Anal. Appl., 21-2 (1999), pp. 431–439.

[40] ———*Any preconditioner belonging to partially equimodular spaces is not optimal for multilevel matrices*, TR nr. 30, Dept. of Mathematics (LAN) - Univ. of Calabria, (1998).

[41] P. TILLI, *Locally Toeplitz matrices: spectral theory and applications*, Linear Algebra Appl., 278 (1998), pp. 91–120.

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

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

[44] R.S. VARGA, *Matrix Iterative Analysis*, Prentice-Hall, Inc., Englewood Cliffs, NJ, 1962.

[45] H. WIDOM, *On the singular values of Toeplitz matrices*, Z. Anal. Anwendungen, 8 (1989), pp. 221–229.