

**A NOTE ON THE ASYMPTOTIC SPECTRA OF FINITE  
DIFFERENCE DISCRETIZATIONS OF SECOND ORDER ELLIPTIC  
PARTIAL DIFFERENTIAL EQUATIONS\***

STEFANO SERRA CAPIZZANO<sup>†</sup>

**Abstract.** We consider Finite Difference discretizations of an elliptic second order PDE as  $-\sum_{i,j=1}^d \frac{\partial}{\partial x_i} \left( a_{i,j}(x) \frac{\partial}{\partial x_j} u(x) \right) = b(x)$  over  $(0, 1)^d$  with Dirichlet boundary conditions, where the  $d \times d$  matrix  $A(x) = (a_{i,j}(x))$  is symmetric, uniformly positive definite and whose entries are Riemann integrable. We choose the discretization so that the resulting matrices  $\{A_n(A)\}_n$  form a sequence of Hermitian positive definite matrices. The eigenvalue distribution has been studied and characterized [23] in terms of weighted multidimensional Szegő formulas. Here by using some tools introduced in a preceding paper [17] we analyze the spectral behaviour of the preconditioned matrix sequences  $\{A_n^{-1}(B)A_n(A)\}_n$  so that  $B(x)$  is symmetric positive definite and with Riemann integrable entries. Some issues on efficient preconditioning strategies are discussed as well.

**1. Introduction.** Let  $A(x) = (a_{i,j}(x))_{i,j=1}^d$  be a  $d \times d$  matrix of functions defined over the hypercube  $\Omega_d = [0, 1]^d$  and let us consider the differential problem

$$(1.1) \quad \begin{aligned} (Lu)(x) &\equiv -\sum_{i,j=1}^d \frac{\partial}{\partial x_i} \left( a_{i,j}(x) \frac{\partial}{\partial x_j} u(x) \right) = b(x) \quad \text{if } x \in \Omega_d^\circ = (0, 1)^d, \\ &+ \quad \text{Dirichlet boundary conditions.} \end{aligned}$$

If  $\circ$  denotes the componentwise Hadamard product [5], then the preceding equation can be conveniently rewritten as follows

$$(1.2) \quad \begin{aligned} -e^T(A(x) \circ H_u(x))e + \text{first order terms} &= b(x) \quad \text{if } x \in \Omega_d^\circ = (0, 1)^d, \\ &+ \quad \text{Dirichlet boundary conditions} \end{aligned}$$

where  $H_u(x)$  is the Hessian matrix of  $u$  evaluated at  $x \in \Omega_d$  and  $e^T = (1, \dots, 1) \in \mathbf{R}^d$ . The discretization of (1.1) by finite differences (FD) over equispaced  $d$ -dimensional grid-sequences  $\mathcal{G}$  leads to a sequence of multilevel linear systems [27] whose sequence of coefficient matrices is denoted by

$$(1.3) \quad \{A_n(A, \mathcal{F}, \mathcal{G})\}.$$

Here each  $A_n(A, \mathcal{F}, \mathcal{G})$  has dimension  $N(n) \times N(n)$  with  $n = (n_1, \dots, n_d)$  and  $N(n) = n_1 \cdots n_d$ . The grid sequence  $\mathcal{G}$  is given by  $\{\mathcal{G}_{n_1} \times \mathcal{G}_{n_2} \times \cdots \times \mathcal{G}_{n_d}\}_n$ ,  $\mathcal{G}_{n_j} = \left\{ \frac{i}{n_j + 1} \right\}_{i=1}^{n_j}$  and  $\mathcal{F}$  is a “symbolic”  $d \times d$  matrix whose entry  $\mathcal{F}_{i,j}$  denotes the FD formula used for the discretization of the operator  $-\frac{\partial^2}{\partial x_i \partial x_j}$ . Finally we must specify the ordering of the unknowns and of the equations. The discretized equation of  $(Lu)(x) = b(x)$  precedes the discretized equation of  $(Lu)(\tilde{x}) = b(\tilde{x})$  for  $x$  and  $\tilde{x}$  belonging to  $\mathcal{G}_n$  if  $x < \tilde{x}$ . Accordingly, the unknown  $u(y)$  precedes  $u(\tilde{y})$  for  $y$  and  $\tilde{y}$  belonging to  $\mathcal{G}_n$  if  $y < \tilde{y}$ . Here we say that  $z < \tilde{z}$  for  $z, \tilde{z} \in \mathcal{G}_n \subset \Omega_d$  if and only if  $\sum_{j=1}^d z_j \prod_{k=j}^d (n_k + 1) < \sum_{j=1}^d \tilde{z}_j \prod_{k=j}^d (n_k + 1)$ . In other words,  $z < \tilde{z}$  iff there exists the minimal index  $j \in \{1, \dots, d\}$  such that  $z_j \neq \tilde{z}_j$  and, in this case, it holds that  $z_j < \tilde{z}_j$ .

\*Received April 22, 1999; accepted for publication September 21, 1999.

<sup>†</sup>Dipartimento di Energetica “S. Stecco”, Via Lombroso 6/17, 50100 Firenze, Dipartimento di Informatica, Corso Italia 40, 56100 Pisa, Italy (serra@mail.dm.unipi.it).

Due to the “shift invariance” property that characterizes any differential operator  $D$  with constant coefficients, it is customary to represent a FD formula  $\phi$  over a sequence of  $n$ -sized equispaced grids as a sequence of Toeplitz matrices  $\{T_n(D, \phi)\}$  related to some (polynomial) symbol.

Here we give the formal definition of multilevel Toeplitz sequences generated by a multivariate symbol.

DEFINITION 1.1. *Let  $f$  be a  $d$  variate complex-valued integrable function, defined over the hypercube  $Q^d$ , with  $Q = (-\pi, \pi)$  and  $d \geq 1$ . From the Fourier coefficients of  $f$*

$$(1.4) \quad f_j = \frac{1}{m\{Q^d\}} \int_{Q^d} f(s) e^{-i(j,s)} ds, \quad \hat{i}^2 = -1, \quad j = (j_1, \dots, j_d) \in \mathbf{Z}^d$$

with  $(j, s) = \sum_{k=1}^d j_k s_k$ ,  $n = (n_1, \dots, n_d)$  and  $N(n) = n_1 \cdots n_d$ , the sequence of Toeplitz matrices  $\{T_n(f)\}$  is defined, where  $T_n(f) = \{f_{j-i}\}_{i,j=e^T}^n \in \mathbf{C}^{N(n) \times N(n)}$ ,  $e^T = (1, \dots, 1) \in \mathbf{N}^d$  is said to be the Toeplitz matrix of order  $n$  generated by  $f$  (see [27]).

For  $D = \frac{d}{dx}$  and a consistent formula  $\phi$  of precision 2 involving three contiguous discretization points  $(u'(x_i) = (u(x_{i+1}) - u(x_{i-1}))/2h + O(h^2), u \in C^2, x_t = th + x_0)$  we have  $2(n+1)^{-1} T_n(D, \phi) = T_n(q(s))$  where  $q(s) = -e^{-is} + e^{is}$  is the generating function [9] in the sense of Definition 1.1 with  $d = 1$ ,  $f_{-1} = -1$ ,  $f_1 = 1$ , and  $f_j = 0$  for any  $|j| \neq 1$ . Among all the possible consistent formulas we choose those such that  $q(s) = -q(\bar{s})$  (antisymmetric formulas according to the terminology of [21]) and such that the points, where we discretize  $u'(x)$  or where  $u(x)$  is evaluated, belong to the same equispaced grid sequence. According to the analysis in [22] these requirements will imply that the resulting FD matrix  $A_n(A, \mathcal{F}, \mathcal{G})$  is nonnegative definite for any multiindex  $n$  whenever  $A(x)$  is nonnegative definite for any  $x$ .

Suppose now that  $\exists a_1, a_2, \dots, a_d \in \mathbf{N}^+$  such that  $n_j + 1 = va_j$  with  $v^{-1}$  being the “finesse” parameter. For  $D = \frac{\partial^2}{\partial x_i \partial x_j}$  and by considering a formula  $\phi$  obtained by composition of unidimensional formulas, we infer that  $v^{-2} T_n(D, \phi) = T_n(q(s))$  with  $s = (s_1, \dots, s_d)$ ,  $n = (n_1, \dots, n_d)$  with  $n_j + 1 = va_j$  and

$$q(s_1, \dots, s_d) = -a_i a_j p(s_i) \overline{p(s_j)}$$

where  $p(s_t) = -\overline{p(s_t)}$  is the generating polynomial of a formula for  $\frac{\partial}{\partial x_t}$ . We observe that, taking the half step formula i.e.  $u'(x_{i+1/2}) = (u(x_{i+1}) - u(x_i))/h + O(h^2)$  we can reduce the bandwidth of the Toeplitz matrices discretizing  $D = \frac{\partial^2}{\partial x_i^2}$  for  $i = 1, \dots, d$ , but the discretization of  $D = \frac{\partial^2}{\partial x_i \partial x_j}$  for  $j \neq i$  would be such that the function  $u$  is evaluated on the wrong half step grid sequence.

Of course from the definition of multilevel Toeplitz matrices is easy to see that  $T_n(f_1(s_1) \cdot f_2(s_2) \cdots f_d(s_d)) = T_{n_1}(f_1(s_1)) \otimes T_{n_2}(f_2(s_2)) \otimes \cdots \otimes T_{n_d}(f_d(s_d))$ . Therefore for the operator  $D$  obtained from (1.1) with  $A(x) = I_d$  we have

$$(1.5) \quad D = e^T H_u e \implies \phi \begin{matrix} \implies \\ \longleftarrow \end{matrix} q(s_1, \dots, s_d) = -e^T \left[ \left( p(s_i) \overline{p(s_j)} \right)_{i,j=1}^d \circ W_{\mathbf{a}} \right] e$$

where the matrix  $W_{\mathbf{a}} = (a_1, \dots, a_d)^T (a_1, \dots, a_d)$  is the nonnegative definite dyad constructed by using the weight numbers  $a_j$  related to the stepsizes in the different

directions. Here  $\phi \xLeftrightarrow{q} q(s_1, \dots, s_d)$  means that the formula  $\phi$  is equivalently represented by the polynomial  $q(s_1, \dots, s_d)$  and  $D = e^T H_u e \implies \phi$  means that  $\phi$  is one possible formula discretizing the operator  $D$ .

Owing to (1.5), it is clear that the matrix sequence  $\{A_n(A, \mathcal{F}, \mathcal{G})\}$  considered in (1.3) can be equivalently represented as  $\{A_n(A, P, W_{\mathbf{a}})\}$  where  $(P \circ W_{\mathbf{a}})_{i,j} = a_i a_j p(s_i) \overline{p(s_j)}$  is the polynomial of the variables  $s_i$  and  $s_j$  uniquely associated to the formula  $\mathcal{F}_{i,j}$  discretizing the operator  $-\frac{\partial^2}{\partial x_i \partial x_j}$  over the (sub)sequence of grids  $\mathcal{G}$  with  $n_j + 1 = a_j v$ .

An interesting and nice correspondence between the differential equation (1.2) and the sequence  $\{A_n(A, P, W_{\mathbf{a}})\}$  is given in the following result.

**THEOREM 1.2.** [23] *Let  $a_{i,j}$  be Riemann integrable for any  $i$  and  $j$ . Suppose that the multi-index  $n = (n_1, \dots, n_d)$  is such that  $n_j + 1 = v a_j$  and  $a_j \in \mathbf{N}^+$  for any  $j$ . For any continuous function  $F$  with bounded support it holds*

$$(1.6) \quad \lim_{v \rightarrow \infty} \frac{1}{N(n)} \sum_{j=1}^{N(n)} F[\sigma_j(v^{-2} A_n(A, P, W_{\mathbf{a}}))] = \frac{1}{m\{\Omega_d \times Q^d\}} \int_{\Omega_d \times Q^d} F(|e^T [A(x) \circ P(s) \circ W_{\mathbf{a}}] e|) dx ds, \\ x \in \Omega_d = [0, 1]^d, \quad s \in Q^d = (-\pi, \pi)^d.$$

We write in short  $\{A_n(A, P, W_{\mathbf{a}})\}_n \sim_{\sigma} |e^T [A(x) \circ P(s) \circ W_{\mathbf{a}}] e|$ .

Notice that the “discrete operator”  $e^T [A(x) \circ P(s) \circ W_{\mathbf{a}}] e$  formally adheres to the continuous operator  $-e^T [A(x) \circ H_u] e$  given in (1.2) since  $P(s) \circ W_{\mathbf{a}}$  is a finite difference (functional) representation of the matrix operator  $-H_u$ .

If  $A(x)$  is symmetric, then we arrange the choice of  $P_{i,j}$  so that  $P(s)$  is a Hermitian, nonnegative definite dyad of functions (notice that  $H_u = [\nabla \cdot \nabla]^T u$  is a sort of dyad with respect to composition of operators): we observe that the nonnegativity of the dyad  $P(s)$  is a consequence of the choice of antisymmetric formulas for the discretization of the first derivatives (refer to [22]). Consequently, in the light of the analysis in [23], formula (1.6) can be also stated in the sense of the eigenvalues.

We proceed as follows. We leave the differential operator in divergence form as in (1.1) and for any derivative appearing in the differential operator we use the same finite difference formula [21] over  $q$  contiguous equispaced points belonging to the mesh  $\{x_t = th, t \in \mathbf{N}, h = (n+1)^{-1}\}$ .

Let  $E_{i,j}$  be the dyad obtained as the product of the  $i$ -th vector of the canonical basis times the transpose of the  $j$ -th vector of the canonical basis. Then the discretized matrix  $A_n(A, P, W_{\mathbf{a}})$  can be written as

$$(1.7) \quad A_n(A, P, W_{\mathbf{a}}) = v^2 \sum_{i,j=1}^d \hat{A}_n(a_{i,j} E_{i,j}, P_{i,j} E_{i,j}, a_i a_j E_{i,j})$$

where  $\hat{A}_n(B, P, W_{\mathbf{a}})$  denotes the discretization of the operator

$$-\sum_{i,j=1}^d \frac{\partial}{\partial x_i} \left( b_{i,j}(x) \frac{\partial}{\partial x_j} u(x) \right)$$

for  $B(x) = (b_{i,j}(x))_{i,j=1}^d$ ,  $b_{i,j} : \Omega_d \rightarrow \mathbf{R}$  and suitably scaled by  $v^{-2}$ .

By [22], the following facts are true:

- F1.**  $\hat{A}_n(a_{i,j}E_{i,j}, P_{i,j}E_{i,j}, a_i a_j E_{i,j}) = \left[ \hat{A}_n(a_{j,i}E_{j,i}, P_{j,i}E_{j,i}, a_i a_j E_{j,i}) \right]^T$  if  $a_{i,j} = a_{j,i}$ .
- F2.** The matrix  $\hat{A}_n(a_{i,i}E_{i,i}, P_{i,i}E_{i,i}, a_i^2 E_{i,i})$  is symmetric semidefinite if  $a_{i,i} \geq 0$  and the matrix  $\hat{A}_n(a_{i,i}E_{i,i}, P_{i,i}E_{i,i}, a_i^2 E_{i,i})$  is a positive definite multilevel Toeplitz matrix generated by a nonnegative polynomial  $a_i^2 |p(s_i)|^2$  if  $a_{i,i}$  is the constant function 1 (see Theorems 3.4 and 3.6 of [21]).
- F3.** The operator  $A_n(\cdot, P, W_{\mathbf{a}})$  is positive in the sense that it maps nonnegative definite matrices of functions  $A(x)$  into nonnegative definite matrices  $A_n(A, P, W_{\mathbf{a}})$  (see the dyadic decomposition of  $A_n(A, P, W_{\mathbf{a}})$  proved in [22]).

**THEOREM 1.3.** [23] *Let  $A(x) = (a_{i,j}(x))_{i,j=1}^d$  be symmetric and let  $a_{i,j}$  be Riemann integrable for any  $i$  and  $j$ . Suppose that the multi-index  $n = (n_1, \dots, n_d)$  is such that  $n_j + 1 = va_j$  and  $a_j \in \mathbf{N}^+$  for any  $j$  and suppose that any derivative appearing in the differential operator (1.1) is discretized by using the same finite difference formula over  $q$  contiguous equispaced points. Then for any continuous function  $F$  with bounded support it holds*

$$(1.8) \quad \lim_{v \rightarrow \infty} \frac{1}{N(n)} \sum_{j=1}^{N(n)} F[\lambda_j(v^{-2}A_n(A, P, W_{\mathbf{a}}))] = \frac{1}{m\{\Omega_d \times Q^d\}} \int_{\Omega_d \times Q^d} F(e^T [A(x) \circ P(s) \circ W_{\mathbf{a}}] e) dx ds, \\ x \in \Omega_d = [0, 1]^d, \quad s \in Q^d = (-\pi, \pi)^d.$$

We write in short  $\{A_n(A, P, W_{\mathbf{a}})\}_n \sim_{\lambda} e^T [A(x) \circ P(s) \circ W_{\mathbf{a}}] e$ .

**REMARK 1.1.** In Theorems 1.2 and 1.3 we have supposed the Riemann integrability of each  $a_{i,j}$ . The statement still holds if, for any  $m \leq M$  real numbers and any  $i$  and  $j$ , the functions  $\max\{\min\{a_{i,j}, M\}, m\}$  are Riemann integrable. However, in order to have (regular) solution to the differential problem (1.1), it is convenient to recall that we must require more regularity for the coefficients of the matrix  $A(x)$ .

**REMARK 1.2.** When we deal with systems of PDEs with  $k$  equations (the solution  $u$  is a  $k$ -dimensional vector and for any  $x$  each entry of the block Hermitian matrix  $A(x)$  is  $k \times k$ ), the case of constant coefficients leads to multilevel block Toeplitz matrices generated by Hermitian matrix-valued polynomials so that the ergodic results proved in [26] can be used to build up a theory for *multilevel block Locally Toeplitz sequences*. In that case, denoting by  $A_n[k](A, P, W_{\mathbf{a}})$  the corresponding discretization matrix, the statement will read as follows: for any continuous function  $F$  with bounded support it holds

$$(1.9) \quad \lim_{v \rightarrow \infty} \frac{1}{N(n)} \sum_{j=1}^{kN(n)} F[\lambda_j(v^{-2}A_n[k](A, P, W_{\mathbf{a}}))] = \frac{1}{m\{\Omega_d \times Q^d\}} \int_{\Omega_d \times Q^d} \sum_{t=1}^k F(\lambda_t (E^T [A(x) \circ P(s) \otimes I_k \circ W_{\mathbf{a}} \otimes I_k] E)) dx ds, \\ x \in \Omega_d = [0, 1]^d, \quad s \in Q^d = (-\pi, \pi)^d$$

where  $E$  is  $e \otimes I_k$ .

Suppose now that the differential problem is defined over a bounded subset  $\Omega'$  of  $\mathbf{R}^d$  with  $\Omega'$  Peano-Jordan measurable [12, Jordan, pp. 28-29], not necessarily hyperrectangular. Without loss of generality we can always suppose that  $\Omega' \subset \Omega_d$ . If it is not the case a linear change of coordinates leads to the desired inclusion.

Given  $A(x)$  defined over  $\Omega'$  we consider its extension  $\hat{A}(x)$  over  $\Omega_d$  in the following way:  $\hat{A}(x) = A(x)$  if  $x \in \Omega'$  and zero otherwise.

With these premise it is easy to see that  $A_n(A, P, W_{\mathbf{a}}, \Omega')$  is a submatrix of dimension  $d_n(\Omega') \times d_n(\Omega')$  of  $A_n(\hat{A}, P, W_{\mathbf{a}})$  while the other rows and columns are zero since the function  $\hat{A}(x)$  vanishes outside  $\Omega'$ . As proved in [23] the following formula holds true:

$$\lim_{v \rightarrow \infty} \frac{1}{d_n(\Omega')} \sum_{j=1}^{d_n(\Omega')} F[\sigma_j(v^{-2} A_n(A, P, W_{\mathbf{a}}, \Omega'))] = \frac{1}{m\{\Omega' \times Q^d\}} \int_{\Omega' \times Q^d} F(|e^T [A(x) \circ P(s) \circ W_{\mathbf{a}}] e|) dx ds$$

$$x \in \Omega', \quad s \in Q^d = (-\pi, \pi)^d.$$

REMARK 1.3. These results have a theoretical interest in its own in order to understand the spectral behaviour of the considered matrix-sequences and its relationships with the properties of the continuous problems (1.1). However another important aspect is of practical nature. The considered analysis provides a theoretical tool in order to devise and to analyze optimal and superlinear preconditioners for the numerical solution of linear systems arising from the discretization of PDEs as (1.1) by preconditioned conjugate gradient methods (for more details concerning special instances of (1.1) see [16, 21, 17]). Some aspects of the general case are analyzed and discussed in the following.

What the previous discussion revealed is that the operator  $A_n(\cdot, P, W_{\mathbf{a}}) = A_n(\cdot)$  is *linear and positive*. The linearity is evident. The positivity stated in **F3** is in the sense of the partial ordering defined over the real space of the Hermitian matrices. In fact if  $A \geq B$  that is  $A \equiv A(x)$  and  $B \equiv B(x)$  are  $d \times d$  Hermitian matrices of functions and  $A - B$  is nonnegative definite for any  $x \in \Omega_d$  then  $A_n(A) \geq A_n(B)$  in the same sense and for any  $n$ .

The second keystone is the “distributional” property. In fact in light of Theorem 1.3 we know that the matrix sequence  $\{A_n(A)\}_n$  is *spectrally distributed as the symbol* [27]  $G(A) = e^T [A(x) \circ P(s) \circ W_{\mathbf{a}}] e$ .

Now we are ready for stating the new results that are proved in this paper. More specifically, we deduce localization and distribution results for the spectra of the preconditioned matrices by using tools developed in [17].

DEFINITION 1.4. Let  $B(x)$  be a symmetric nonnegative definite matrix of Riemann integrable functions. The preconditioned matrix  $P_n(A, B)$  is defined as  $A_n^+(B)A_n(A)$  where the preconditioner  $A_n(B)$  is Hermitian and nonnegative definite and  $X^+$  denotes the usual pseudo-inverse of Moore-Penrose [13, 15] of a generic matrix  $X$ .

In the following section, we will prove the following results.

1. Let  $\ker(X)$  be the null space of a matrix  $X$ . If  $\ker(B(x)) \equiv K_x \subseteq \ker(A(x))$ , then each nonzero eigenvalue of  $P_n(A, B)$  is contained in the set  $[r, R]$  with

$$r = \inf_{x \in \Omega_d} \lambda_{\min}((B|(K_x)^\perp)^{-1}(x)(A|(K_x)^\perp)(x))$$

and

$$R = \sup_{x \in \Omega_d} \lambda_{\max}((B|(K_x)^\perp)^{-1}(x)(A|(K_x)^\perp)(x)).$$

Here  $(B|(K_x)^\perp)(x)$  (resp.  $(A|(K_x)^\perp)(x)$ ) denotes for any  $x$  the projection of  $B(x)$  (resp.  $A(x)$ ) over the orthogonal to the kernel of  $B(x)$ .

2. If the assumption in 1. is violated, then at least one of the values  $r$  and  $R$  is unbounded.
3. If the functions  $G(B)$  is sparsely vanishing (sv) i.e.  $m\{(x, s) \in \Omega_d \times Q^d : G(B)(x, s) = 0\} = 0$ , then the sets of the eigenvalues  $\{\Lambda_n\}_n$ ,  $\Lambda_n = \{\lambda_i^{(n)}\}_{1 \leq i \leq N(n)}$  of the matrices  $\{P_n(A, B)\}_n$  satisfy the subsequent ergodic formula:

for any function  $F \in C_0(\mathbf{R})$ , it holds

$$(1.10) \quad \lim_{v \rightarrow \infty} \frac{1}{N(n)} \sum_{i=1}^{N(n)} F(\lambda_i^{(n)}) = \frac{1}{m\{\Omega_d \times Q^d\}} \int_{\Omega_d \times Q^d} F(G(A)/G(B)) dx ds.$$

The interesting fact is that we start with the families of matrices  $\{A_n(A)\}$  and  $\{A_n(B)\}$  where  $G(A)$ ,  $G(B)$  are Riemann integrable and we obtain a formula for the family of preconditioned matrices involving  $\frac{G(A)}{G(B)}$  which is measurable but may fail to be Riemann integrable and even Lebesgue integrable. However, the right-hand side of the proved relation (1.10) makes sense because  $F(G(A)/G(B))$  is Lebesgue measurable and bounded and consequently it belongs to  $L^1(\Omega_d \times Q^d)$ .

The paper is organized as follows. In Section 2 we derive the main results and in Section 3 we discuss some related preconditioning strategies.

## 2. Main results.

**2.1. Localization results.** The following Lemma 2.1 is a preparatory result for Theorem 2.2.

LEMMA 2.1. [17] *Let  $A$  and  $B$  be two  $n \times n$  Hermitian matrices with  $B$  nonnegative definite and  $\mathbf{u}$  be a nonzero vector of  $\mathbf{C}^n$ . Let us denote by  $\ker(X)$  the null space of a matrix  $X$ . The following three statements hold true.*

1. *If  $\mathbf{u}^H B \mathbf{u} = 0$ , then  $\mathbf{u} \in \ker(B)$ .*
2. *In general, from equation  $\mathbf{u}^H A \mathbf{u} = 0$  does not follow that  $\mathbf{u} \in \ker(A)$ .*
3. *If there exists  $r \in \mathbf{R}$  such that, for any  $\mathbf{x} \in \mathbf{C}^n$ ,  $r \mathbf{x}^H B \mathbf{x} \leq \mathbf{x}^H A \mathbf{x}$  and  $\mathbf{u}^H A \mathbf{u} = \mathbf{u}^H B \mathbf{u} = 0$ , then  $\mathbf{u} \in \ker(A)$ .*

THEOREM 2.2. *Let  $r$  and  $R$  be two constants such that  $A(x) - rB(x)$  is nonnegative definite for any  $x$  and  $A(x) - RB(x)$  is nonpositive definite for any  $x$ . Each nonzero eigenvalue of  $P_n(A, B)$  belongs to  $[r, R]$ . In addition if  $r$  and  $R$  are finite, then the kernel of the matrices  $A_n^+(B)$  and  $A_n(B)$  is contained into the kernel of  $A_n(A)$ . Finally if  $0 < r \leq R < \infty$  then the kernels of  $A_n(A)$  and  $A_n(B)$  coincide.*

*Proof.* Let us consider  $\mathbf{u} \in \mathbf{C}^n$ . Then, by the assumption we infer that

$$r\mathbf{u}^H A_n(B)\mathbf{u} \leq \mathbf{u}^H A_n(A)\mathbf{u} \leq R\mathbf{u}^H A_n(B)\mathbf{u}.$$

From this the claimed thesis follows as in Theorem 3.1 of [8].

If  $A_n(B)$  is invertible then each eigenvalue of the preconditioned matrices belongs to  $[r, R]$ .  $\square$

REMARK 2.1. If  $B(x)$  is invertible for any  $x$ , then it is easy to give a characterization of the best constants  $r$  and  $R$ . Indeed, by using the Sylvester inertia law, we have  $A(x) - rB(x)$  nonnegative definite iff  $B^{-1/2}(x)A(x)B^{-1/2}(x) - rI$  nonnegative definite. The latter is clearly equivalent to require that the minimal eigenvalue of  $B^{-1/2}(x)A(x)B^{-1/2}(x)$  is not less than  $r$  for any  $x$ . Since  $B^{-1/2}(x)A(x)B^{-1/2}(x)$  is similar to  $B^{-1}(x)A(x)$  it follows that best constant  $r$  is  $\inf_{x \in \Omega_d} \lambda_{\min}(B^{-1}(x)A(x))$ . Analogously the best constant  $R$  is  $\sup_{x \in \Omega_d} \lambda_{\max}(B^{-1}(x)A(x))$ .

REMARK 2.2. If  $B(x)$  is not invertible for some  $x$  the characterization of  $r$  and  $R$  is a bit more complicated.

First assume that  $\exists x$  such that  $B(x)$  is singular and  $\exists \mathbf{v} = \mathbf{v}(x)$  for which  $B(x)\mathbf{v} = 0$ ,  $\|\mathbf{v}\|_2 = 1$  and  $t(x) = \mathbf{v}^H A(x)\mathbf{v} < 0$ . Then  $r = -\infty$ . The proof is obvious: the statement " $\exists r \in \mathbf{R}$  so that  $A(x) - rB(x) \geq 0$ " is contradicted by  $\mathbf{v}^H (A(x) - rB(x))\mathbf{v} = t(x) < 0$ .

Of course, if  $t(x) > 0$  then the conclusion is that  $R = \infty$ .

From the point of view of the applications (in particular the numerical solution of the linear systems arising in the discretization [1, 11] of the given PDE), in light of Theorem 2.2, it is evident that we are interested in the case where  $r$  and  $R$  are finite (more specifically finite and strictly positive). In this case the application of the PCG method [3] with  $A_n(B)$  being the preconditioner [3] that we suppose invertible leads to an iterative method converging to the solution within a preassigned accuracy in a number of steps independent of  $n$  [4]. Therefore, by recalling Remark 2.2 and Lemma 2.2 in [17], we have to impose that for any  $x$  the relation  $\ker(B(x)) \subseteq \ker(A(x))$ . By using these preliminary remarks the following result holds

THEOREM 2.3. *Assume that  $K_x = \ker(B(x)) \subseteq \ker(A(x))$  for any  $x$ . Then each nonzero eigenvalue of  $P_n(A, B)$  is contained in the set  $[r, R]$  with*

$$r = \inf_{x \in \Omega_d} \lambda_{\min}((B|(K_x)^\perp)^{-1}(x)(A|(K_x)^\perp)(x))$$

and

$$R = \sup_{x \in \Omega_d} \lambda_{\max}((B|(K_x)^\perp)^{-1}(x)(A|(K_x)^\perp)(x)).$$

Here  $(B|(K_x)^\perp)(x)$  (resp.  $(A|(K_x)^\perp)(x)$ ) denotes for any  $x$  the projection of  $B(x)$  (resp.  $A(x)$ ) over the orthogonal to the kernel of  $B(x)$ .

*Proof.* Let  $\mathbf{y}$  be a generic nonzero vector of  $\mathbf{C}^d$  and let us write  $\mathbf{y}$  as  $\mathbf{v} + \mathbf{w}$  where  $\mathbf{v} \in K_x$  and  $\mathbf{w} \in K_x^\perp$ . Let  $N$  be the  $d \times q$  matrix with  $q = \dim(K_x^\perp)$  whose columns constitute a basis for  $K_x^\perp$  so that  $\mathbf{w} = N\hat{\mathbf{w}}$  (for a certain  $\hat{\mathbf{w}} \in \mathbf{C}^d$ ). Then it holds that

$$\begin{aligned} \mathbf{y}^H (A(x) - rB(x))\mathbf{y} &= \mathbf{w}^H (A(x) - rB(x))\mathbf{w} \\ &= \hat{\mathbf{w}}^H N^H (A(x) - rB(x))N\hat{\mathbf{w}} \\ &= \hat{\mathbf{w}}^H (N^H A(x)N - rN^H B(x)N)\hat{\mathbf{w}}. \end{aligned}$$

Therefore, by using the same argument as in Remark 2.1, the claimed thesis is proved.  $\square$

**2.2. Ergodic and distribution results.** Now we localize (asymptotically and up to  $o(N(n))$  elements) the position of the eigenvalues of  $A_n(A)$ .

LEMMA 2.4. *Let us suppose that  $A(x)$  is symmetric for any  $x$  and*

$$(2.1) \quad m\{(x, s) \in \Omega_d \times Q^d : G(A)(x, s) = z \text{ or } G(A)(x, s) = t\} = 0.$$

*Then the number  $N(z, t, n)$  of eigenvalues of  $A_n(A)$  belonging to  $(z, t)$ ,  $z < t$ , is asymptotical to  $c(z, t)N(n)$  with  $c(z, t) = m\{(x, s) \in \Omega_d \times Q^d : G(A)(x, s) \in (z, t)\}$ . The same is true if the special cases  $(z, t) = (-\infty, 0)$  or  $(z, t) = (0, \infty)$  are considered.*

*Proof.* Since  $\{A_n(A)\}_n \sim_\sigma G(A)$ , it follows that the claimed thesis is equivalent to consider equation (1.6) with  $F$  being the characteristic function of  $[z, t]$ . But  $F$  is not continuous so that relation (1.6) is not automatically satisfied. However the function  $F$  can be approximated in  $L^1$  norm by continuous symbols and, under the assumption (2.1), the eigenvalues of  $\{A_n(A)\}_n$  cannot cluster at  $z$  neither at  $t$ : the latter remark ends the proof.  $\square$

For the main results (Theorems 2.9 and 2.10) about the distribution of the eigenvalues of the preconditioned matrix we need some definitions and the preliminary Lemma 2.8.

DEFINITION 2.5. *A function  $f \in \mathcal{M}(\Omega_d \times Q^d)$  is sparsely vanishing (sv) if the set of its zeros has zero Lebesgue measure [7]. Here  $\mathcal{M}(\Omega_d \times Q^d)$  is the space of the measurable functions of  $\Omega_d \times Q^d$ .*

DEFINITION 2.6. *Let  $A$  and  $B$  be two symmetric matrices of measurable functions with nonnegative definite  $B$ . Let us define  $S(A; B)$  as the set of the simple real valued functions of the form*

$$\sum_{i \in K} \alpha_i Ch(I_i), \quad z_i, t_i \in \mathbf{R}, \quad z_i \leq t_i, \\ I_i = (z_i, t_i) \text{ or } I_i = [z_i, t_i) \text{ or } I_i = (z_i, t_i] \text{ or } I_i = [z_i, t_i]$$

*where  $K$  is a finite set of indices,  $Ch(X)$  denotes the characteristic function of the set  $X$  and  $m\{(x, s) \in \Omega_d \times Q^d : G(A)/G(B) = z_i\} = m\{x \in \Omega_d \times Q^d : G(A)/G(B) = t_i\} = 0$ .*

DEFINITION 2.7. *Let  $A$  and  $B$  be two symmetric matrices of measurable functions with nonnegative definite  $B$ . The symbol  $R(A; B)$  indicates the topological closure with respect to the infinity norm of the simple functions  $S(A; B)$ .*

LEMMA 2.8. [17] *Let  $A$  and  $B$  be two  $n \times n$  Hermitian matrices with  $B$  nonnegative definite and whose null space has dimension  $k$ . Let us denote by  $N_0(X)$ ,  $N_+(X)$  and  $N_-(X)$  the number of zero, positive and negative eigenvalues of the matrix  $X$ , respectively. The following two statements hold true.*

1. *The matrices  $B^+A$  and  $(B^+)^{1/2}A(B^+)^{1/2}$  have the same spectrum (the same eigenvalues with the same algebraic multiplicities).*
2. *The matrices  $B^+A$  and  $A$  have "almost" the same inertia (if  $k \ll n$ ): more precisely,  $N_0(A) \leq N_0(B^+A) \leq N_0(A) + k$ ,  $N_+(A) - k \leq N_+(B^+A) \leq N_+(A)$  and  $N_-(A) - k \leq N_-(B^+A) \leq N_-(A)$ .*

**THEOREM 2.9.** *Let us assume that the function  $G(B)$  is nonnegative and sparsely vanishing, then the eigenvalues  $\{\Lambda_n\}_n$ ,  $\Lambda_n = \{\lambda_i^{(n)}\}_{1 \leq i \leq N(n)}$  of the matrices  $\{P_n(A, B)\}_n$  enjoy an ergodic formula, i.e, for any simple function  $F \in S(A; B)$ , it follows*

$$(2.2) \quad \lim_{v \rightarrow \infty} \frac{1}{N(n)} \sum_{i=1}^{N(n)} F(\lambda_i^{(n)}) = \frac{1}{m\{\Omega_d \times Q^d\}} \int_{\Omega_d \times Q^d} F(G(A)/G(B)) dx ds.$$

*Proof.* Since  $F$  is a finite linear combination of characteristic functions and both the left-hand and right-hand sides of (2.2) are linear with respect to the argument  $F$ , we can prove the result for a single characteristic function  $Ch(z, t)$  with  $z < t$ . In addition, we have

$$Ch(z, t) = Ch(z, \infty) - Ch([t, \infty))$$

and  $m\{(x, s) \in \Omega_d \times Q^d : G(A)/G(B) = t\} = 0$ , therefore it is enough to prove the result for the characteristic functions of the half-lines like  $(z, \infty)$ . Notice that, due to the boundedness of  $Ch(z, \infty)$ , it follows that the right-hand side of (2.2) is well defined and makes sense. Therefore we have to prove the equality of the two quantities.

First let us consider  $F = Ch(z, \infty)$ . Then

$$\sum_{i=1}^{N(n)} F(\lambda_i^{(n)}) = \#\{i : \lambda_i^{(n)} > z\}.$$

Because the function  $G(B)$  is nonnegative and sv, in the light of Lemma 2.4 and by virtue of the positivity of the matrix operator  $A_n(\cdot)$ , it follows that the preconditioner  $A_n(B)$  is nonnegative definite and its null eigenvalues are at most  $o(N(n))$ . Consequently, by **part I** of Lemma 2.8, we deduce that  $P_n(A, B) = A_n^+(B)A_n(A)$  and  $(A_n^+(B))^{1/2}A_n(A)(A_n^+(B))^{1/2}$  have the same characteristic polynomial and therefore this is true for

$$P_n(A, B) - zI \quad \text{and} \quad T_z = (A_n^+(B))^{1/2}A_n(A)(A_n^+(B))^{1/2} - zI,$$

with  $I = I_{N(n)}$  being the identity matrix of order  $N(n)$ . Now by **part II** of Lemma 2.8, the matrices  $T_z$  and  $Y_z = A_n^{1/2}(B)T_zA_n^{1/2}(B)$  have the same inertia up to within an error of  $o(N(n))$  eigenvalues. Since

$$Y_z = P(A_n(A) - zA_n(B))P,$$

where the orthogonal projector  $P$  equals  $A_n^{1/2}(B)(A_n^+(B))^{1/2}$ , again by the combined application of **part I** and **part II** of Lemma 2.8, we infer that  $Y_z$  and  $A_n(A) - zA_n(B)$  have the same ‘‘asymptotical’’ inertia. To summarize, up to within an error of  $o(N(n))$  eigenvalues, the matrix  $A_n(A) - zA_n(B)$  has the same inertia as the matrix  $P_n(A, B) - zI$ . In other words, counting the eigenvalues of the preconditioned matrix, which are greater than  $z$ , is equivalent, up to within an error whose magnitude is  $o(N(n))$ , to count the positive eigenvalues of the matrix  $A_n(A) - zA_n(B)$ . Now we use the linearity of the matrices  $A_n(A)$  with respect to  $A$  and then, up to within an error of  $o(N(n))$ , the number of positive eigenvalues of  $P_n(A, B) - zI$  coincides with the number of positive eigenvalues of

$$A_n(A - zB).$$

Now  $m\{(x, s) \in \Omega_d \times Q^d : G(A) - zG(B) = 0\} = m\{(x, s) \in \Omega_d \times Q^d : G(A)/G(B) = z\} + m\{(x, s) \in \Omega_d \times Q^d : G(A) = G(B) = 0\}$  and, by the assumptions,  $m\{(x, s) \in \Omega_d \times Q^d : G(A) = G(B) = 0\} = 0$  ( $G(B)$  is sv) and  $m\{x \in \Omega : G(A)/G(B) = z\} = 0$  since, by definition of  $S(A; B)$ ,  $z$  does not belong to the set where the image-measure via  $G(A)/G(B)$  accumulates. Finally, we find that

$$m\{(x, s) \in \Omega_d \times Q^d : G(A) - zG(B) = 0\} = 0$$

and, in the light of Lemma 2.4, we have that the number  $N(0, \infty, n)$  of the positive eigenvalues of

$$A_n(A - zB)$$

is asymptotical to  $c(0, \infty)N(n)$  with  $c(0, \infty) = m\{(x, s) \in \Omega_d \times Q^d : G(A) - zG(B) > 0\}$ . This fact, in formulas, is equivalent to write

$$\begin{aligned} \#\{i : \lambda_i^{(n)} > z\} &= m\{(x, s) \in \Omega_d \times Q^d : G(A) - zG(B) > 0\}N(n) \\ &\quad + o(N(n)) = \\ &= m\{(x, s) \in \Omega_d \times Q^d : G(A)/G(B) > z\}N(n) \\ &\quad + o(N(n)). \end{aligned}$$

Finally, by recalling that  $F = Ch(z, \infty)$ , it is trivial to recognize that the latter equation is equivalent to the claimed thesis

$$\lim_{v \rightarrow \infty} \frac{1}{N(n)} \sum_{i=1}^{N(n)} F(\lambda_i^{(n)}) = \frac{1}{m\{\Omega_d \times Q^d\}} \int_{\Omega_d \times Q^d} F(G(A)/G(B)) dx ds.$$

□

**THEOREM 2.10.** *Under the assumptions of the preceding theorem, the eigenvalues  $\lambda_i^{(n)}$  of the matrices  $\{P_n(A, B)\}_n$  satisfy a more general ergodic formula, i.e., for any function  $F \in R(A; B)$ , we have*

$$(2.3) \quad \lim_{v \rightarrow \infty} \frac{1}{N(n)} \sum_{i=1}^{N(n)} F(\lambda_i^{(n)}) = \frac{1}{m\{\Omega_d \times Q^d\}} \int_{\Omega_d \times Q^d} F(G(A)/G(B)) dx ds.$$

*Proof.* By definition of  $R(A; B)$ , it follows that  $F$  can be approximated in infinity norm, by simple functions belonging to the set  $S(A; B)$ . More specifically, for any positive  $\epsilon$  there exists a finite collection of indices  $I_\epsilon$  and points  $z_i < t_i$  so that

$$(2.4) \quad \begin{aligned} m\{(x, s) \in \Omega_d \times Q^d : G(A)/G(B) = z_i\} = \\ m\{(x, s) \in \Omega_d \times Q^d : G(A)/G(B) = t_i\} = 0 \end{aligned}$$

and  $F_\epsilon = \sum_{i \in I_\epsilon} \alpha_i Ch(z_i, t_i)$  with

$$\|F - F_\epsilon\|_\infty \leq \epsilon.$$

From the crucial equation (2.4), it follows that we do not find clusters of eigenvalues around the points  $z_i$  or  $t_i$ . Therefore

$$\frac{1}{N(n)} \sum_{i=1}^{N(n)} F(\lambda_i^{(n)}) = \frac{1}{N(n)} \sum_{i=1}^{N(n)} F_\epsilon(\lambda_i^{(n)}) + h_n(\epsilon)$$

with  $|h_n(\epsilon)| \leq \epsilon$ . Moreover, since  $\|F - F_\epsilon\|_\infty \leq \epsilon$ , it simply follows that

$$\left| \frac{1}{m\{\Omega_d \times Q^d\}} \int_{\Omega_d \times Q^d} (F - F_\epsilon)(G(A)/G(B)) dx ds \right| \leq \frac{1}{m\{\Omega_d \times Q^d\}} \int_{\Omega_d \times Q^d} \epsilon dx ds = \epsilon.$$

which completes the proof.  $\square$

Some remarks are needed:

REMARK 2.3. Concerning the assumption that  $G(B)(x, s)$  is sv in Theorems 2.9 and 2.10 we observe that  $G(B)(x, s)$  vanishes at  $(x, 0)$  for any  $x \in \Omega_d$  due to the consistency condition of the underlying FD formulas. Moreover if  $B(x)$  is sv in the sense that its minimal eigenvalue is sv according to Definition 2.5, then it is really trivial to check that  $G(B)(x, s)$  is sv since

$$\begin{aligned} G(B)(x, s) &\geq \lambda_{\min}(B(x)) e^T (I_d \circ P(s) \circ W_{\mathbf{a}}) e \\ &= \lambda_{\min}(B(x)) \sum_{j=1}^d a_j^2 |p_j(s_j)|^2. \end{aligned}$$

On the other hand, surprisingly enough, the former assumption on  $B(x)$  can be substantially relaxed. Indeed we only suppose that the maximal eigenvalue of  $B(x)$  is sv and we call  $\mathbf{v}(x)$  the corresponding normalized eigenvector that is  $\sum_{j=1}^d |\mathbf{v}_j(x)|^2 = 1$  almost everywhere. Then

$$\begin{aligned} G(B)(x, s) &\geq \lambda_{\max}(B(x)) e^T (\mathbf{v}(x) \mathbf{v}^T(x) \circ P(s) \circ W_{\mathbf{a}}) e \\ &= \lambda_{\max}(B(x)) \left| \sum_{j=1}^d a_j \mathbf{v}_j(x) p_j(s_j) \right|^2. \end{aligned}$$

Since  $\theta(x, s) = \left| \sum_{j=1}^d a_j \mathbf{v}_j(x) p_j(s_j) \right|^2$  is a not identically zero multivariate trigonometric polynomial for almost every  $x \in \Omega_d$ , it follows that each section of the zeros of  $\theta(x, s)$ , with respect to almost every  $x$ , is an algebraic manifold in the variable  $s \in (-\pi, \pi)^d$  and consequently has zero Lebesgue measure in  $(-\pi, \pi)^d$ . The application of the Fubini-Tonelli Theorem (see e.g. [10, Theorem A, p. 147]) on multivariate integration yields the claimed thesis. Therefore we can conclude that the assumptions of the above mentioned Theorems 2.9 and 2.10 are in reality very mild.

REMARK 2.4. If  $d = 1$ , then  $A(x) = a(x)$ ,  $B(x) = b(x)$ , and the matrix sequence  $\{P_n(A, B)\}_n$  distributes as the function  $\eta = a/b$  which depends only on the variable  $x \in \Omega_d$ . Conversely, if  $d \geq 2$ , then the matrix sequence  $\{P_n(A, B)\}_n$  distributes as the function  $\eta = G(A)/G(B)$  which depends on the variables  $s \in Q^d$  as well except when  $A(x) = \theta(x)B(x)$  with  $\theta(x)$  scalar function.

REMARK 2.5. All the ergodic results for Toeplitz sequences like the Szegő formula [9] and its extensions [2, 14, 30, 27, 28, 24] are stated by assuming  $F$  continuous with bounded support. Actually all the results are still valid if no assumptions are made

on the support of  $F$  but we suppose that  $F$  is continuous and with finite limits at  $+\infty$  and at  $-\infty$ . In this way we require that  $F \in C(\mathbf{R})$  where  $\mathbf{R}$  is the two points compactification of  $\mathbf{R}$  (observe that  $R(A; B) \supset C(\mathbf{R})$  for any pair  $(A, B)$ ). Notice that this is a bit less general than the assumptions on the test functions given in [26] where the author proved ergodic formulas concerning multilevel block Toeplitz sequences with  $F$  being uniformly continuous and bounded.

**3. Applications to the preconditioning.** Suppose that  $A(x)$  is elliptic that is  $\exists r, R$  with  $0 < r \leq R < \infty$  for which  $rI_d \leq A(x) \leq RI_d$  uniformly over  $\Omega_d$ .

The idea is to approximate  $A(x)$  by a matrix-valued function  $B(x)$  such that the corresponding FD coefficient matrix  $A_n(B)$  is easy to invert. If the ratio  $\sqrt{R/r}$  is not too large the convergence theory of the PCG method [3, 4] and Theorems 2.2-2.3 suggest one that the multilevel symmetric and positive definite matrix  $A_n(I_d)$  is a good preconditioner ( $B(x) = I_d$ ). On the other hand this basic approximation can be improved: see [16, 21] for other proposals and the comprehensive survey [6] for efficient PCG-based Toeplitz solvers.

For notational simplicity we first assume that  $A(x) = a(x)I_d$  where  $a(x)$  is a scalar function whose range is contained in  $[r, R]$ .

Consider a value  $\epsilon > 0$  small enough and  $N \equiv N_\epsilon = \lceil \frac{R-r}{\epsilon} \rceil$ . Construct the piecewise constant approximant  $a_\epsilon(x) = r + (j + 1/2)N^{-1}$  if  $x \in \{a(x) \in [r + jN^{-1}, r + (j + 1)N^{-1}]\}$  for  $j = 0, \dots, N - 1$ . Therefore we have

$$\min_{j=0, \dots, N-1} \frac{r + jN^{-1}}{r + (j + 1/2)N^{-1}} \leq \frac{a}{a_\epsilon} \leq \max_{j=0, \dots, N-1} \frac{r + (j + 1)N^{-1}}{r + (j + 1/2)N^{-1}}$$

where

$$1 > \frac{r + jN^{-1}}{r + (j + 1/2)N^{-1}} = 1 - \frac{N^{-1}/2}{r + (j + 1/2)N^{-1}} \geq 1 - \frac{N^{-1}/2}{r} \equiv c_{\min}(\epsilon)$$

$$1 < \frac{r + (j + 1)N^{-1}}{r + (j + 1/2)N^{-1}} = 1 + \frac{N^{-1}/2}{r + (j + 1/2)N^{-1}} \leq 1 + \frac{N^{-1}/2}{R} \equiv c_{\max}(\epsilon).$$

Therefore the spectral condition number of  $A_n^{-1}(a_\epsilon I_d)A_n(A)$  is bounded by  $c_{\max}(\epsilon)/c_{\min}(\epsilon)$  which tends to 1 as  $\epsilon$  tends to zero.

Now the spectral condition number of  $A_n^{-1}(I_d)A_n(a_\epsilon I_d)$  is approximately the same as the one  $A_n^{-1}(I_d)A_n(A)$  so that the proposed preconditioning step seems useless. However, by Theorem 2.10, the sequence  $\{A_n^{-1}(I_d)A_n(a_\epsilon I_d)\}_n$  is spectrally distributed as the function  $a_\epsilon$  i.e.

$$(3.1) \quad \{A_n^{-1}(I_d)A_n(a_\epsilon I_d)\}_n \sim_\lambda a_\epsilon.$$

Since the range of  $a_\epsilon$  is constituted by  $N$  points it follows by the subsequent Proposition 3.3 that the sequence  $\{A_n^{-1}(I_d)A_n(a_\epsilon I_d)\}_n$  has exactly  $N$  subclusters [19] so that we expect that the PCG method applied to the preceding sequence converges in a number of steps proportional to  $N$  but substantially independent of the size  $n$  (see [4]).

Finally, for the sake of completeness and in order to properly formulate Proposition 3.3, we report the definition of clusters and subclusters and some relevant related properties (For a slightly different definition of subcluster see [29]).

**DEFINITION 3.1.** [27] *Consider a sequence of  $d_n \times d_n$  complex matrices  $\{A_n\}_n$  (with  $d_n < d_{n-1}$ ) and a set  $M$  in the nonnegative real line. Denote by  $M_\epsilon$  the  $\epsilon$ -extension of  $M$ , which is the union of all balls of radius  $\epsilon$  centered at points of  $M$ .*

For any  $n$ , let  $\gamma_n(\epsilon) \equiv \gamma_n(A_n, M, \epsilon)$  count those singular values of  $A_n$  that do not belong to  $M_\epsilon$ .

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

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

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

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

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

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

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

DEFINITION 3.2. [19] Let  $\{A_n\}_n$ ,  $M$ ,  $M_\epsilon$  and  $\gamma_n(\epsilon)$  be as in the preceding definition.

- The set  $M$  is a subcluster if

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

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

With regard to the terminology of the preceding definitions, when the eigenvalues/singular values of  $\{P_n^{-1}A_n - I\}_n$ ,  $I = I_{N(n)}$  are *properly clustered* at zero (and the minimal eigen/singular value of  $P_n^{-1}A_n$  does not go to zero too fast) or when the sequence of the spectral condition numbers  $\kappa(P_n^{-1}A_n)$  of  $\{P_n^{-1}A_n\}_n$  is upperbounded by a constant independent of  $n$ , we know [4] that a constant number of iterations is required by the PCG method in order to solve a linear system with coefficient matrix  $A_n$  within a preassigned accuracy. In particular, if  $\{P_n^{-1}A_n - I\}_n$  is *properly clustered* then the related PCG method is optimal and, after a suitable constant number of iterations, the convergence is of superlinear type (refer to [4] for more details).

PROPOSITION 3.3. [19] Consider a sequence of  $d_n \times d_n$  complex matrices  $\{A_n\}_n$  and suppose that  $\{A_n\}_n \sim_\sigma f$  with  $f$  measurable function over a finite-measure domain  $K \subset \mathbf{R}^d$ . Then  $p$  is a (singular value) subcluster point for  $\{A_n\}_n$  iff  $m\{x \in K : |f(x)| = p\} > 0$ . If  $f$  is real-valued then  $p$  is a (eigenvalue) subcluster point for  $\{A_n\}_n$  iff  $m\{x \in K : f(x) = p\} > 0$ . Here  $m\{\cdot\}$  is the Lebesgue measure on  $\mathbf{R}^d$ . Finally if  $f$  is piecewise constant then  $\{A_n\}_n$  has a finite number of subcluster points.

Since  $a_\epsilon$  is piecewise constant and equation (3.1) holds true, it follows that the latter proposition proves that the sequence  $\{A_n^{-1}(I_d)A_n(a_\epsilon I_d)\}_n$  possesses a finite number of subcluster points.

Now we give some basic numerical evidences. We choose two simple examples namely equation (1.1) with  $d = 2$ ,  $A(x_1, x_2) = (\alpha + x_1 + x_2)I_2$ ,  $b(x_1, x_2) = 1$  and  $\alpha \in \{0, 1\}$ .

The approximating matrix  $A_\epsilon(x_1, x_2) = a_\epsilon(x_1, x_2)I_2$  is determined as follows:

$$a_\epsilon(x_1, x_2) = \frac{1}{2} \left[ \sup_{(x_1, x_2) \in J_i \times J_k} (\alpha + x_1 + x_2) + \inf_{(x_1, x_2) \in J_i \times J_k} (\alpha + x_1 + x_2) \right]$$

Table 1:  $\alpha = 1$ ,  $X_n = A_n(A)$  and  $P_n = A_n(I_2)$ 

$n = (n_1, n_2)$ , $N(n)$	(10,10), 100	(30,30), 900
$N_{\text{it}}$	12	13

Table 2:  $\alpha = 1$ ,  $X_n = A_n(A)$  and  $P_n = A_n(A_\epsilon)$ 

$n = (n_1, n_2)$ , $N(n)$	(10,10), 100	(30,30), 900
$N_{\text{it}}$	9	9

for  $(x_1, x_2) \in J_i \times J_k$ ,  $J_q \in ((q-1)/4, q/4] \cap (0, 1)$  and  $i, j, q = 1, 2, 3, 4$ .

The matrix  $A_n(A, P, W_{\mathbf{a}})$  is obtained by leaving the operator in (1.1) in divergence form, by using the same double step formula for each first derivative ( $v'(x_i) = (v(x_{i+1}) - v(x_{i-1}))/2h + O(h^2)$ ,  $v \in C^2$ ,  $x_t = th + x_0$ ) and by assuming that  $\mathbf{a} = e$ . The same process is performed in order to define the matrices  $A_n(A_\epsilon, P, W_{\mathbf{a}})$  and  $A_n(I_2, P, W_{\mathbf{a}})$ .

For any  $\alpha \in \{0, 1\}$  we consider three tables. The first concerns the case where  $X_n = A_n(A, P, W_{\mathbf{a}})$  is preconditioned by  $P_n = A_n(I_2, P, W_{\mathbf{a}})$ . The second concerns the case where  $X_n = A_n(A, P, W_{\mathbf{a}})$  is preconditioned by  $P_n = A_n(A_\epsilon, P, W_{\mathbf{a}})$  and the third is the case where  $X_n = A_n(A_\epsilon, P, W_{\mathbf{a}})$  is preconditioned by  $P_n = A_n(I_2, P, W_{\mathbf{a}})$ .

All the experiments are done in MATLAB on a PC 486 and in all the tables we report the number of the PCG iterations in order to reach a residual error whose Euclidean norm is bounded by  $10^{-7}$  and in the case of  $n_1 = n_2 = 10$  and  $n_1 = n_2 = 30$ .

Some remarks are useful. When  $\alpha = 1$  (Tables 1, 2 and 3), all the preconditioners are optimal because the related spectral condition numbers of the preconditioned matrices are bounded from above by absolute constants not depending on  $n = (n_1, n_2)$ . To prove this, refer to Theorems 2.2–2.3 and to the following inequalities

$$\kappa^{(1)} \equiv \kappa(P_n(A, I_2)) \leq \sup_x \kappa(A(x)) = 3,$$

$$\kappa^{(2)} \equiv \kappa(P_n(A_\epsilon, I_2)) \leq \kappa^{(1)},$$

$$\kappa^{(3)} \equiv \kappa(P_n(A, A_\epsilon)) \ll \kappa^{(1)}$$

where  $\kappa(X)$  denotes the spectral condition number of a matrix  $X$ .

Observe that, for large  $n$ , the condition number of  $P_n(A, A_\epsilon)$  is substantially less than the condition number of  $P_n(A, I_2)$  and this explains why we have a lower number of iterations. Conversely, the condition numbers of  $P_n(A, I_2)$  and  $P_n(A_\epsilon, I_2)$  are practically the same but we have again a substantial improvement. The explanation of this behavior relies in the fact that the sequence  $\{P_n(A_\epsilon, I_2)\}$  has a finite number of subclusters that cover all its spectra since  $\{P_n(A_\epsilon, I_2)\} \sim_\lambda a_\epsilon$ . The importance of the subcluster structure is even more evident in the case where  $\alpha = 0$  where  $a(x)$  vanishes at  $x = 0$ . Indeed, as shown by Tables 3, 4 and 5, the only case of “practical optimality” is the one of  $\{P_n(A_\epsilon, I_2)\}$  and this is explained by the presence of subclusters since

$$\sup_{\epsilon > 0} \left\{ \lim_{n=(n_1, n_2) \rightarrow (\infty, \infty)} \kappa(P_n(A_\epsilon, I_2)) \right\} = \infty.$$

Table 3:  $\alpha = 1$ ,  $X_n = A_n(A_\epsilon)$  and  $P_n = A_n(I_2)$ 

$n = (n_1, n_2)$ , $N(n)$	(10,10), 100	(30,30), 900
$N_{it}$	9	9

Table 4:  $\alpha = 0$ ,  $X_n = A_n(A)$  and  $P_n = A_n(I_2)$ 

$n = (n_1, n_2)$ , $N(n)$	(10,10), 100	(30,30), 900
$N_{it}$	21	38

**4. Acknowledgement.** I thank the referees for the very pertinent comments and suggestions. A special thank goes to Paolo Tilli for stimulating conversations.

## REFERENCES

- [1] W. AMES, *Numerical Methods for Partial Differential Equations*. Third Edition, Academic Press Inc., New York, 1992.
- [2] F. AVRAM, *On bilinear forms on Gaussian random variables and Toeplitz matrices*, Probab. Th. Related Fields, 79 (1988), pp. 37–45.
- [3] O. AXELSSON AND V. BARKER, *Finite Element Solution of Boundary Value Problems, Theory and Computation*, Academic Press Inc., New York, 1984.
- [4] O. AXELSSON AND G. LINDSKÖG, *The rate of convergence of the preconditioned conjugate gradient method*, Numer. Math., 52 (1986), pp. 499–523.
- [5] R. BHATIA, *Matrix Analysis*, Springer-Verlag, New York, 1997.
- [6] R. H. CHAN AND M. NG, *Conjugate gradient methods for Toeplitz systems*, SIAM Rev., 38 (1996), pp. 427–482.
- [7] F. DI BENEDETTO AND S. SERRA CAPIZZANO, *A unifying approach to matrix algebra preconditioning*, Numer. Math., 82:1 (1999), pp. 57–90.
- [8] A. FRANGIONI AND S. SERRA CAPIZZANO, *Matrix-valued linear positive operators and applications to graph optimization*, TR nr. 04-99, Dept. Informatica, Univ. of Pisa, 1999.
- [9] U. GRENANDER AND G. SZEGÖ, *Toeplitz Forms and Their Applications*, Second Edition, Chelsea, New York, 1984.
- [10] P. HALMOS, *Measure Theory*, Graduate Texts in Math. 18, Springer, New York, 1974.
- [11] E. ISAACSON AND H. KELLER, *Analysis of Numerical Methods*, John Wiley and Sons, New York, 1966.
- [12] C. JORDAN, *Cours d'Analyse de l'Ecole Polytechnique: Vol. I*, Gauthier-Villars, Paris, France, 1909.
- [13] E. H. MOORE, *General Analysis. Part I*, Amer. Phil. Society, Philadelphia, 1935.
- [14] S. V. PARTER, *On the distribution of singular values of Toeplitz matrices*, Linear Algebra Appl., 80 (1986), pp. 115–130.
- [15] R. PENROSE, *A generalized inverse for matrices*, Proc. Cambridge Phil. Soc., 51 (1955), pp. 406–413.
- [16] 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.
- [17] S. SERRA CAPIZZANO, *An ergodic theorem for classes of preconditioned matrices*, Linear Algebra Appl., 282 (1998), pp. 161–183.
- [18] S. SERRA CAPIZZANO, *Locally X matrices, spectral distributions, preconditioning and applications*, SIAM J. Matrix Anal. Appl., 21:4 (2000), pp. 1354–1388.
- [19] S. SERRA CAPIZZANO, *Spectral behavior of matrix-sequences and discretized boundary value problems*, Linear Algebra Appl., to appear. A preliminary version in TR nr. 31, LAN - Dept. Math., Univ. of Calabria, 1998.
- [20] S. SERRA CAPIZZANO, *Some theorems on linear positive operators and functionals and their applications*, Comput. Math. Applic., 39:3-4 (2000), pp. 139–167.
- [21] S. SERRA CAPIZZANO AND C. TABLINO POSSIO, *Spectral and structural analysis of high order finite differences matrices discretizing elliptic operators*, Linear Algebra Appl., 293 (1999), pp. 85–131.
- [22] S. SERRA CAPIZZANO AND C. TABLINO POSSIO, *Positive representation formulas for finite difference discretizations of (elliptic) second order PDEs*, Contemporary Math., in press.
- [23] S. SERRA CAPIZZANO AND P. TILLI, *From partial differential equations to generalized Locally*

Table 5:  $\alpha = 0$ ,  $X_n = A_n(A)$  and  $P_n = A_n(A_\epsilon)$ 

$n = (n_1, n_2)$ , $N(n)$	(10,10), 100	(30,30), 900
$N_{it}$	13	20

Table 6:  $\alpha = 0$ ,  $X_n = A_n(A_\epsilon)$  and  $P_n = A_n(I_2)$ 

$n = (n_1, n_2)$ , $N(n)$	(10,10), 100	(30,30), 900
$N_{it}$	12	14

*Toeplitz sequences*, TR nr. 12-99, Dept. Informatica, Univ. of Pisa, 1999.

- [24] P. TILLI, *Singular values and eigenvalues of non Hermitian block Toeplitz matrices*, Linear Algebra Appl., 272 (1998), pp. 59–89.
- [25] P. TILLI, *Locally Toeplitz sequences: spectral properties and applications*, Linear Algebra Appl., 278 (1998), pp. 91–120.
- [26] P. TILLI, *A note on the spectral distribution of Toeplitz matrices*, Linear Multilin. Algebra, 45 (1998), pp. 147–159.
- [27] E. TYRTYSHNIKOV, *A unifying approach to some old and new theorems on distribution and clustering*, Linear Algebra Appl., 232 (1996), pp. 1–43.
- [28] E. TYRTYSHNIKOV AND N. ZAMARASHKIN, *Spectra of multilevel Toeplitz matrices: advanced theory via simple matrix relationships*, Linear Algebra Appl., 270 (1998), pp. 15–27.
- [29] E. TYRTYSHNIKOV AND N. ZAMARASHKIN, *Thin structure of eigenvalue clusters for non-Hermitian matrices*, Linear Algebra Appl., 292 (1999), pp. 297–310.
- [30] H. WIDOM, *Szegő limit Theorem: the higher dimensional matrix case*, J. Funct. Anal., 39 (1980), pp. 182–198.