HOW TO PROVE THAT A PRECONDITIONER CANNOT BE SUPERLINEAR

S. SERRA CAPIZZANO AND E. TYRTYSHNIKOV

Abstract. In the general case of multilevel Toeplitz matrices, we recently proved that any multilevel circulant preconditioner is not superlinear (a cluster it may provide cannot be proper). The proof was based on the concept of quasi-equimodular matrices, although this concept does not apply, for example, to the sine-transform matrices. In this paper, with a new concept of partially equimodular matrices, we cover all trigonometric matrix algebras widely used in the literature. We propose a technique for proving the non-superlinearity of certain frequently used preconditioners for some representative sample multilevel matrices. At the same time, we show that these preconditioners are, in a certain sense, the best among the sublinear preconditioners (with only a general cluster) for multilevel Toeplitz matrices.

1. Motivation

In the past two decades a lot of work has been done to find superlinear preconditioners (see [4] and references therein). While many of them have been devised and studied for the unilevel case, no superlinear preconditioning strategies based on matrix algebras with unitary transforms are known for multilevel Toeplitz-like structures. In [21] we proved that such a search cannot be successful in the case where we use circulant-like approximation spaces (see [15] for an application of our result [21] to image restoration problems). Here we extend this analysis to a wide class of approximation spaces including all the relevant algebras used in the literature, and we prove the same negative result as in [21] (see [22] for a different approach). The contribution (both theoretical and practical) is in the following direction: if we desire to find an optimal preconditioning strategy, then for multilevel structures with polynomial ill-conditioning the only possibility is to use a multigrid approach [9, 5] or spectrally equivalent approximations [17]. Indeed, we recall that no spectrally equivalent matrix algebra preconditioners are known for multilevel Toeplitz-like structures with polynomial ill-conditioning; and, in this respect, we have some preliminary negative results [16].

Concerning direct solvers and Toeplitz structures, an interesting observation is that algorithms based on displacement rank have been successfully devised only for unilevel Toeplitz matrices (see [14]). For multilevel Toeplitz matrices, no fast or superfast procedures are known. We are convinced that a negative answer should be found in this case too, since no displacement operators with rank constant with...
respect to the dimension are available for multilevel Toeplitz structures. Therefore, the mathematical tools developed below for analyzing preconditioned conjugate gradient algorithms could be useful with regard to the analysis of this kind of direct solvers as well.

2. Introduction

Although a preconditioner can be generally regarded as an approximant to a given matrix, the approximation here is understood in a fairly broad sense. In particular, given \( C_n \) and \( A_n \), both of order \( n \), assume that for any \( \varepsilon > 0 \) there exist matrices \( E_n \varepsilon \) and \( R_n \varepsilon \) such that

\[
(*) \quad A_n - C_n = E_n \varepsilon + R_n \varepsilon, \quad ||E_n \varepsilon||_2 \leq \varepsilon, \quad \text{rank } R_n \varepsilon \leq r(n, \varepsilon) = o(n).
\]

Then we say that \( C_n \) and \( A_n \) are \( \varepsilon \)-close by rank with the rank bound \( r(n, \varepsilon) \). If \( \gamma_n(\varepsilon) \) counts how many singular values \( \sigma_{in}(A_n - C_n) \) are greater than \( \varepsilon \), then \( (*) \) amounts to the claim that \( \gamma_n(\varepsilon) = o(n) \); in other words, the singular values of \( A_n - C_n \) have a general cluster at zero. By definition, it becomes a proper cluster if \( \gamma_n(\varepsilon) = O(1) \), which holds equally with \( r(n, \varepsilon) = \varepsilon \) = \( O(1) \), for any \( \varepsilon > 0 \).

To estimate the convergence rate of the cg-like methods, one should be interested in the \( \varepsilon \)-closeness of \( C_n^{-1}A_n \) and \( I_n \) rather than \( C_n \) and \( A_n \). In the former case, \( C_n \) is called superlinear for \( A_n \) if \( r(n, \varepsilon) = O(1) \), and sublinear if \( r(n, \varepsilon) = o(n) \). For the cg-like methods, superlinear preconditioners provide the superlinear convergence (see [1] [23] [27]).

However, it is still useful and easier to work with the \( \varepsilon \)-closeness of \( C_n \) and \( A_n \). In this case, \( C_n \) is superlinear for \( A_n \) in the above sense so long as \( r(n, \varepsilon) = O(1) \) and a term of rank bounded by \( \varepsilon \) and a term of constant rank, and this contradicts the assumption that \( A_n \) and \( C_n \) are not \( \varepsilon \)-close by \( O(1) \) rank.
Proposition 2.3 gives a criterion to establish the existence of a superlinear preconditioner by analyzing the difference $A_n - C_n$. However, we have to suppose the uniform boundedness of $C_n$ and $C_n^{-1}$, and this request is not practical since $C_n$ is unknown. In Proposition 2.2 we give a simpler condition to establish the non-superlinearity of a preconditioner $C_n$, but again we impose an assumption on $C_n$. In the next proposition we eliminate any assumption on $C_n$, so the related statement is truly useful to decide the non-superlinearity of a preconditioner $C_n$.

**Proposition 2.3.** Let $A_n$ and $C_n$ be nonsingular. If $A_n$ is bounded uniformly in $n$ and if $A_n$ and $C_n$ are not $\varepsilon$-close by $O(1)$ rank, then $C_n$ is not superlinear for $A_n$.

**Proof.** By contradiction, if $C_n$ is a superlinear preconditioner for $A_n$, then $C_n^{-1}A_n - I_n$ can be written as a term of norm bounded by $\varepsilon$ and a term of rank bounded by a constant independent of $n$. Therefore, by using the Sherman-Morrison-Woodbury formula we have that $A_n^{-1}C_n - I_n$ can be written as a term of norm bounded by $O(\varepsilon)$ and a term of rank bounded by a constant independent of $n$. Therefore $-(A_n - C_n) = A_n(A_n^{-1}C_n - I_n)$ is the sum of a term of norm bounded by $O(\varepsilon)$ and a term of constant rank, and this contradicts the assumption that $A_n$ and $C_n$ are not $\varepsilon$-close by $O(1)$ rank.

In this paper we propose a technique for proving that

$$r(n, \varepsilon) \neq o(\rho(n)),$$

for certain unbounded functions $\rho(n)$, so long as $A_n$ and $C_n$ possess some special properties. In fact, we generalize our previous results from [21]. In that paper we considered $C_n$ of the form $C_n = U_nD_nV_n$, where the matrices $U_n$ are unitary, the $D_n$ are diagonal, and the $V_n$ are unitary quasi-equimodular. For brevity, we refer to the $C_n$ as QE-based matrices.

We call a matrix *equimodular* if all its entries are equal in modulus. It is clear that any unitary equimodular matrix of order $n$ has all its entries equal to $1/\sqrt{n}$ in modulus. If there exist two positive constants $0 < c_1 \leq c_2 < \infty$, independent of $n$, so that the entries of a sequence of matrices $V_n$ belong to $[c_1\sqrt{n}, c_2/\sqrt{n}]$, then the sequence is called *quasi-equimodular*.

The foremost result proved in [21] is the following.

**Theorem 2.1.** Assume that $A_n$ and $C_n$ are $\varepsilon$-close by rank with the rank bound $r(n, \varepsilon)$ and, for any $\varepsilon > 0$, let $A_n$ have $\rho(n, \varepsilon)$ columns of Euclidean length not greater than $\varepsilon$. Let $||A_n||_2$ be bounded uniformly in $n$. Assume also that

(a) the singular values of $A_n$ are not clustered at zero, and

(b) $C_n = U_nD_nV_n$ are QE-based matrices.

Then $r(n, \varepsilon) \neq o(\rho(n, \varepsilon))$.

With this theorem we come up with some interesting negative results about superlinear preconditioners for multilevel matrices.

Recall that $A$ is a $p$-level matrix of multiorder $n = (n_1, \ldots, n_p)$ (and of order $N(n) \equiv n_1 \cdots n_p$), if it encompasses $n_1 \times n_1$ blocks, each of them has $n_2 \times n_2$ blocks, and so on. The rows and columns of $A$ can be pointed to through multi-indices $i = (i_1, \ldots, i_p)$ and $j = (j_1, \ldots, j_p)$ as follows: $A = [a_{ij}]$, $1 \leq i, j \leq n$, where $1 = (1, \ldots, 1)$, and inequalities between multi-indices are understood in the entrywise sense. By definition, $o(n) = o(N(n))$ as $n \to \infty$, and the latter means that every component of $n$ tends to infinity.
A$_n$ is called a $p$-level Toeplitz matrix if $A_n = [a_{i,j}]$, and it is a $p$-level circulant if $a_{i,j} = a_{i,j} \text{mod}(n)$, where $i \text{mod}(n) = (i_1 \text{mod}(n_1), \ldots, i_p \text{mod}(n_p))$. Toeplitz matrices $A_n = A_n(f)$ are said to be generated by a symbol $f \in L_1(\Pi)$, $z \in \Pi \equiv [-\pi, \pi]^p$, if their elements come from the formal Fourier expansion $f(z) \sim \sum_k a_k e^{i(k,z)}$, $r^2 = -1$ where $k = (k_1, \ldots, k_p)$, $(k, z) = k_1 z_1 + \cdots + k_p z_p$.

For matrices $A_n(f)$, assumption (a) of Theorem 2.1 can be reformulated in terms of properties of the symbol $f$. We may use the following result: if a symbol $f \in L_1(\Pi)$ is a complex-valued function of $z \in \Pi = [-\pi, \pi]^p$, then the singular values of $A_n(f)$ are distributed as $|f(z)|$ with $z \in \Pi$, that is, for any $F$ continuous with a bounded support,

$$\frac{1}{N(n)} \sum_{1 \leq t \leq n} F(\sigma_t(A_n)) \to \frac{1}{(2\pi)^p} \int_{\Pi} F(|f(z)|)dz.$$ 

This is a Szegö-like theorem obtained in [28]. For the unilevel Hermitian Toeplitz case, the classical distribution results can be found in [11]. It is easy to see that the above assumption (a) is fulfilled so long as $f(z)$ is not identically zero almost everywhere. In particular (a) holds if $f(z)$ does not vanish on a subset of positive Lebesgue measure in $\Pi$: the latter is equivalent to the demand that $f(z)$ is at most sparsely vanishing [24][7].

In the next example we take a concrete sequence of matrices that shows the applicability of Theorem 2.1. Consider 2-level Toeplitz matrices $A_n(f)$ for $f(x_1, x_2) = \exp\{ix_1\}$. In line with the above,

$$A_n = \begin{bmatrix} 0 & I \\ \vdots & \ddots & \ddots \\ & \ddots & \ddots & I \\ 0 & & 0 & I \\ 0 & & & & \end{bmatrix},$$

where 0 and $I$ are of order $n_2$ as $n = (n_1, n_2)$. Consider any QE-based matrices $C_n$ that are $\varepsilon$-close to $A_n$ by rank. From Theorem 2.1 since $A_n$ has $\rho(n) = n_2$ zero columns, we deduce that $r(n, \varepsilon) \neq 0(\rho(n))$. Consequently, since $\rho(n) = n_2 \to \infty$ as $n \to \infty$, we conclude that $r(n, \varepsilon) \neq O(1)$. Now, choose any real value $s$ and take up matrices $I_n + sA_n$. With QE-based $C_n$ with $U_n = V_n^*$, it is practically obvious to infer that any preconditioners of the form $I_n + C_n$ for $I_n + sA_n$ are not superlinear.

One may remark still that matrices $I_n + C_n$ cannot be easily invertible in the case of $V_n \neq U_n^*$ (they might not be QE-based), and hence, we scarcely use them as preconditioners. All the same, in the case where $V_n = U_n^*$ the matrices $I_n + C_n$ remain QE-based so long as the $C_n$ are. For the above $A_n$, consider the splitting $A_n = H_n1 + iH_n2$, where $H_n1$ and $H_n2$ are Hermitian, and approximate them by QE-based (with common $V_n$) matrices $C_n1$ and $C_n2$, respectively. If the $\varepsilon$-rank bounds for $H_n1 - C_n1$ and $H_n2 - C_n2$ are both $O(1)$, then $A_n$ should be $\varepsilon$-close to $C_n = C_n1 + iC_n2$ with the rank bound $O(1)$. Since $C_n$ is also QE-based, it contradicts the above negative result. Now, choose a real value $s$ and consider the following Hermitian 2-level Toeplitz matrices: $I_n + sH_n1$ and $I_n + sH_n2$ (to guarantee that $I_n + sH_n1$ and $I_n + sH_n2$ are both invertible, we may choose $s \in (0, 1)$). We have proved that at least one of the two does not admit a superlinear preconditioner among QE-based matrices of Hermitian pattern. Still, we cannot say definitely which of the two.
In this respect, our negative result for the Hermitian case does not look very satisfactory. In pursuit of using Theorem 2.1 in a direct way, we need to produce a symbol giving Hermitian multilevel Toeplitz matrices with sufficiently many zero columns, which is barely possible.

Another criticism of Theorem 2.1 is that it ignores some important matrix algebra preconditioners. In particular, the sine-transform matrices \( S_n = \sqrt{\frac{2}{n}} \left[ \sin \frac{\pi ij}{n+1} \right]_{ij=1}^n \) are not quasi-equimodular, and hence they cannot be used as \( V_n \). To cover such cases, we now introduce a new concept of partially equimodular matrices.

**Definition.** Matrices \( V_n \) are called partially equimodular if there exist two positive constants \( c \) and \( d \), independent of \( n \), so that, in any column of \( V_n \) for any \( n \), the number of entries which are not less in modulus than \( c/\sqrt{n} \) is greater than or equal to \( dn \). Matrices \( C_n = U_n D_n V_n \), where \( U_n \) are unitary, \( D_n \) are diagonal, and \( V_n \) are unitary partially equimodular, will be referred to as PE-based matrices.

In this paper we modify Theorem 2.1 so that matrices \( C_n \) are allowed to be PE-based. Since any PE-based matrices are also QE-based, we thus weaken premise (b). At the same time, we have to strengthen premise (a), in such a way that it is still equally easy to fulfill for producing the same negative results.

The paper is organized as follows. In Section 3, we discuss the concept of partially equimodular matrices and their relation to trigonometric matrix algebras. We show that all known algebras consist of PE-based matrices. In Section 4 we propose a new technique to study the \( \varepsilon \)-closeness of matrices. Then, in Section 5 we show how this technique can be applied to the multilevel Toeplitz matrices and in Section 6 we discuss the (more difficult!) Hermitian case.

### 3. Unitary partially equimodular matrices and matrix algebras

We begin with a very convenient indication for unitary matrices to be partially equimodular.

**Theorem 3.1.** Let \( V_n \) be unitary \( n \times n \) matrices, and assume that the maximal in modulus entry of \( V_n \) does not exceed \( M/\sqrt{n} \), where \( M \) is a constant independent of \( n \). Then the matrices \( V_n \) are partially equimodular.

**Proof.** With \( 0 < c < M \), consider an arbitrary column of \( V_n \) and denote by \( \zeta_n \) the number of its entries which are not less in modulus than \( c/\sqrt{n} \). Since the Euclidean length of the column is equal to 1, we obtain

\[
1 \leq (n - \zeta_n) \frac{c^2}{n} + \zeta_n \frac{M^2}{n} \quad \Rightarrow \quad \zeta_n \geq \frac{1 - c^2}{M^2 - c^2} n. \quad \square
\]

Consider a sequence of grids \( W_n = \{x_{in}\} \) with \( n \) nodes \( x_{1n} < \ldots < x_{nn} \) defined on a given basic interval \( I \), and suppose that for any \( n \) there is a set \( \Phi_n = \{\phi_{in}\} \) of \( n \) functions orthogonal in the following sense (see [19]):

\[
[\phi_{in}, \phi_{jn}]_n = \sum_{i=1}^{n} (\phi_{in} \bar{\phi}_{jn}) (x_{in}) = \delta_{ij}, \quad 1 \leq i, j \leq n.
\]

It follows that the generalized Vandermonde matrices \([\Phi_n(x_{in})]^n_{ij=1}\) are unitary. By \( \{V_n^* D_n V_n\} \), we denote a sequence of matrix algebras, each, for a fixed
$V_n^* D_n V_n$, where $V_n$ is fixed and $D_n$ is an arbitrary diagonal matrix.

When the functions of $\Phi_n$ are trigonometric polynomials, we come up with the well-known matrix algebras associated with the classical fast transforms $\{\text{circulants}\}$. Canonical examples of such matrix algebras are the circulant, the $\tau$, and the Hartley $\mathbb{H}$. The corresponding functions $\phi_{jn}$ and nodes $x_{in}$ are the following:

\[
\phi_{jn} = \frac{1}{\sqrt{n}} \exp\{i(j-1)x\}, \quad x_{in} = \frac{2\pi(i-1)}{n} \in [-\pi, \pi];
\]
\[
\phi_{jn} = \frac{1}{\sqrt{n}} \sin(jx), \quad x_{in} = \frac{\pi i}{n+1} \in [0, \pi];
\]
\[
\phi_{jn} = \frac{1}{\sqrt{n}} \left(\sin((j-1)x) + \cos((j-1)x)\right), \quad x_{in} = \frac{2\pi(i-1)}{n} \in [-\pi, \pi].
\]

In addition, 7 other examples of unitary transforms $V_n$ related to cosine/sine functions are presented in [13]. One more example is the matrix algebra of $\varepsilon$–circulants $\mathbb{E}$ with $|\varepsilon| = 1$; if $\varepsilon = \exp\{12\pi \psi\}$, then the corresponding functions and nodes are of the form

\[
\phi_{jn} = \frac{1}{\sqrt{n}} \exp\{i(j-1+\psi)x\}, \quad x_{in} = \frac{2\pi(i-1)}{n} \in [-\pi, \pi].
\]

In the above cases, it is easy to check that unitary matrices $V_n$ satisfy the hypothesis of Theorem 3.1 and hence, are partially equimodular. The same holds even in a more general context.

**Theorem 3.2.** Assume that the nodes $x_{in}$ are quasi-uniform on $I$ in the sense that $\sum_{i=1}^{n} |I|/n - (x_{i+1} - x_{in})| = o(1)$, and let $\phi_{jn}(x) = \theta_{n}\phi_{j}(x)$, where the $\theta_n$ are constants, and the $\phi_{j}(x)$ are continuous functions uniformly bounded in $j$ with a finite number of zeroes on $I$. Assume that the matrices $V_n = [\phi_{jn}(x_{in})]$ are unitary. Then, they satisfy the hypothesis of Theorem 3.1.

**Proof.** Let $\phi = \phi_1$. It is sufficient to prove that $c_1 n \leq \sum_{i=1}^{n} |\phi(x_{in})|^2 \leq c_2 n$ for some positive $c_1$ and $c_2$. The second inequality follows from the boundedness of $\phi$. The first one stems from the demand that the grids are quasi-uniform and $\phi$ has only a finite number of zeroes. With these assumptions, there is $\delta > 0$ such that the number of the indices $i$ for which $|\phi(x_{in})| > \delta$ is bounded from below by $c(\delta)n$. Hence, we can take $c_1 = c_2 = c(\delta)$. Now, since $V_n$ is unitary, we deduce that

\[1 = \sum_{i=1}^{n} |(V_n)_{i,1}|^2 = \theta_n^2 \sum_{i=1}^{n} |\phi(x_{in})|^2.
\]

Therefore the relation $1/\sqrt{c_1 n} \leq \theta_n \leq 1/\sqrt{c_2 n}$ is proved, and consequently, due to the uniform boundedness of the $f_j$, the inequalities $\max |(V_n)_{i,j}| \leq M/\sqrt{n}$ hold true with $M$ an absolute constant.

Using Theorem 3.1, we now conclude that the matrices $V_n$ in Theorem 3.1 are partially equimodular (PE).

For the $p$-level case, the matrix algebras $\{V_n^* D_n V_n\}$, where $n = (n_1, \ldots, n_p)$, are constructed through the Kronecker products of $p$ sequences of unilevel matrix algebras $\{V_{nk} D_{nk} V_{nk}\}$ so that

\[V_n = \bigotimes_{1 \leq k \leq p} V_{nk}, \quad D_n = \bigotimes_{1 \leq k \leq p} D_{nk}.\]
Formally, the $V_{n_k}$ may correspond to $\rho$ different sequences of quasi-uniform grids $W_{n_k}^{(k)}$ on $I$ and functions $\Phi_{n_k}^{(k)}$, though we usually assume that there is no dependence on the upper index.

4. RANK BOUNDS FOR $\varepsilon$-CLOSENESS

When choosing $C_n$ to be $\varepsilon$-close to a given $A_n$, we are interested in knowing how the rank bounds may behave. For this we propose a new version of Theorem 2.1 in order to relax the assumption that the $C_n$ are QE-based. We can now consider arbitrary PE-based $C_n$ (rather than QE-based).

**Theorem 4.1.** Assume that $A_n$ and $C_n$ are $\varepsilon$-close by rank with the rank bound $r(n, \varepsilon)$, and, for any $\varepsilon > 0$, let $A_n$ have $\rho(n, \varepsilon)$ columns of the Euclidean length not greater than $\varepsilon$. Let $\|A_n\|_2$ be bounded uniformly in $n$. Assume also that

(a) the singular values $\sigma_1(A_n) \geq \cdots \geq \sigma_n(A_n)$ behave so that for any $0 < d < 1$, there is a positive $q(d)$ providing that

$$S(d, A_n) = \sum_{d_n \leq j \leq n} \sigma^2_j(A_n) \geq q(d)n$$

for all sufficiently large $n$, and

(b) $C_n = U_nD_nV_n$ are PE-based matrices.

Then $r(n, \varepsilon) \neq o(\rho(n, \varepsilon))$.

**Proof.** We may assume that $\|C_n\|_2$ are bounded uniformly in $n$. If this is not the case, we pass to other matrices $\tilde{C}_n$ satisfying the same premises with $r(n, \varepsilon)$ being possibly replaced by $2r(n, \varepsilon)$. Indeed, the number of singular values for $C_n$ that are greater than $\|A_n\|_2 + \varepsilon$ cannot exceed $r(n, \varepsilon)$, and we may thus get to $\tilde{C}_n$ by cutting off the largest diagonal entries of $D_n$ in the equation $C_n = U_nD_nV_n$.

To get a contradiction, assume that $r = r(n, \varepsilon) = o(\rho(n, \varepsilon))$ and show, based on this, that for any $\varepsilon > 0$ there must be a column in $C_n$ whose $2$-norm is less than or equal to $\varepsilon$ for any $n$ large enough.

Let $E_{n, \varepsilon}$ contain $\rho = \rho(n, \varepsilon)$ columns of $I_n$ that correspond to $\varepsilon$-small columns of $A_n$. By contradiction again, let us suppose that for some $\delta > 0$, independent of $n$, every column of $C_nE_{n, \varepsilon}$ is greater than $\delta$ in the $2$-norm for infinitely many $n$.

Therefore,

$$\delta \rho < \|C_nE_{n, \varepsilon}\|_F = \|A_n + (C_n - A_n)E_{n, \varepsilon}\|_F$$

$$\leq \|A_nE_{n, \varepsilon}\|_F + \|(C_n - A_n)E_{n, \varepsilon}\|_F$$

$$\leq \varepsilon \rho + \|(C_n - A_n)E_{n, \varepsilon}\|_F$$

$$\leq \varepsilon \rho + \sigma_1(C_n - A_n)\varepsilon + \varepsilon(\rho - r)$$

$$= O(r + \varepsilon \rho),$$

which is at odds with $r = o(\rho)$.

Thus, from $r(n, \varepsilon) = o(\rho(n, \varepsilon))$ we infer that, for any $\varepsilon > 0$, for all sufficiently large $n$ there exists a column $e_{n, \varepsilon}$ of $I_n$ such that $\|C_ne_{n, \varepsilon}\|_2 \leq \varepsilon$. Moreover, from $r(n, \varepsilon) = o(n)$, it follows that $A_n$ and $C_n$ have the same clusters \cite{25, 27}, and therefore for any positive $\delta$ we have

$$S(d, A_n) = S(d, C_n) + o(n).$$
Finally, taking into account that the $V_n$ are partially equimodular with the constants $c$ and $d$, we deduce that
\[
\varepsilon \geq \frac{\|C_n e_{\varepsilon n}\|_2}{\|D_n V_n e_{\varepsilon n}\|_2} = \frac{\sqrt{\frac{c^2}{n} S(d, C_n)}}{\sqrt{\frac{c^2}{n} (S(d, A_n) + o(n))}} \geq \frac{\sqrt{\frac{c^2}{n} (q(d)n + o(n))}}{\sqrt{\frac{c^2}{n} (q(d)n + o(n))}} \geq c \sqrt{q(d)} + o(1),
\]
which is impossible due to the arbitrariness of $\varepsilon$. \hfill \Box

**Theorem 4.2.** Under the hypotheses of Theorem 4.1, assume that
\[
\lim_{n \to \infty} \rho(n, \varepsilon) = \infty \quad \forall \varepsilon > 0.
\]
Then the singular values of $A_n - C_n$ cannot have a proper cluster at zero.

**Proof.** If $r(n, \varepsilon)$ is bounded uniformly in $n$, then $r(n, \varepsilon) = o(\rho(n, \varepsilon))$, which contradicts the conclusion of Theorem 4.1. \hfill \Box

**Theorem 4.3.** Under the hypotheses of Theorem 4.2, consider any matrices $P_n$ such that $P_n + A_n$ and $P_n + C_n$ are nonsingular and $P_n + A_n$ is uniformly bounded in $n$ in the spectral norm. Then the matrices $P_n + C_n$ cannot be superlinear preconditioners for $P_n + A_n$.

**Proof.** By direct application of Proposition 2.3, if the singular values of $A_n - C_n$ do not have a proper cluster at zero, then the preconditioner $P_n + C_n$ is not superlinear for $P_n + A_n$. It remains to apply Theorem 4.2. \hfill \Box

In Theorem 4.1 the assumption that $A_n$ has many small columns indicates the direction in which we may seek negative results. This assumption is not easy to reconcile with some structural requirements on $A_n$. To this end, it might be useful to present a modification of Theorem 4.1 that allows us to consider $A_n$ with no small columns. A price to pay for this is that we require a bit more than the PE property of $V_n$, though still less than the QE property.

Let us say that the matrices $A_n$ are $\varepsilon$-zeroed on $\rho(n, \varepsilon)$ columns of the matrices $Z_n$ if $A_n Z_n$ has $\rho(n, \varepsilon)$ columns of Euclidean length not greater than $\varepsilon$ for all sufficiently large $n$. Also, the matrices $Z_n$ will be referred to as uniformly sparse if the number of nonzeroes in every column of $Z_n$ is bounded above uniformly in $n$.

**Theorem 4.4.** Assume that $A_n$ and $C_n$ are $\varepsilon$-close by rank with the rank bound $r(n, \varepsilon)$, and the $C_n$ are $\varepsilon$-zeroed on $\rho(n, \varepsilon)$ columns of uniformly sparse unitary matrices $Z_n$. Let $\|A_n\|_2$ be bounded uniformly in $n$. Assume also that
(a) the singular values $\sigma_1(A_n) \geq \cdots \geq \sigma_n(A_n)$ behave so that for any $0 < d < 1$ there is a positive $q(d)$ providing that
\[
S(d, A_n) = \sum_{d n \leq j \leq n} \sigma_j^2(A_n) \geq q(d)n
\]
for all sufficiently large $n$, and
(b) $C_n = U_n D_n V_n$ are PE-based matrices and, in addition, the maximal modulus entry of $V_n$ does not exceed $M/\sqrt{n}$, where $M$ does not depend on $n$.

Then $r(n, \varepsilon) \neq o(\rho(n, \varepsilon))$. 

Proof. Denote by $k$ the maximal number of nonzeros in any column of $Z_n$. Then, since the $Z_n$ are uniformly sparse, $k < +\infty$. Allowing for the entries of $V_n$ being not greater in modulus than $M/\sqrt{n}$, we conclude now that the entries of $Z_n V_n$ in modulus do not exceed $k M/\sqrt{n}$. Since $Z_n V_n$ are unitary as a product of unitary matrices, they are partially equimodular by Theorem 3.1. Now, $\rho(n, \varepsilon)$ columns of $B_n \equiv Z_n^* A_n Z_n$ are $\varepsilon$-small, and it remains to apply Theorem 4.1 to the matrices $B_n$ and arbitrary PE-based $C_n$. □

5. What is the best for multilevel Toeplitz matrices?

We start with a remark on the assumption (a) used in Theorem 4.1. In the case of multilevel Toeplitz matrices generated by $f(z)$, it reduces to an assumption on $f(z)$. Note that the inequality $S(d, A_n) \geq q(d) N(n)$, for any $d \in (0, 1)$, holds if $f(z)$ does not vanish in a subset of $[-\pi, \pi]^p$ of positive measure; in other words, $f(z)$ is at most sparsely vanishing [24, 7]. It is remarkable that a slightly weaker property of $f$ accounts for the assumption (a) in the previous Theorem 2.1. This means that the stronger assumption we have introduced does not seem to substantially affect the construction of “bad examples”.

Consider a $p$-variable symbol of the form

$$f(x_1, \ldots, x_p) = \frac{1}{2} \exp\{ik_1 x_1 + \cdots + ik_p x_p\}, \quad k_j \geq 1, \ j = 1, \ldots, p,$$

and the corresponding $p$-level Toeplitz matrices $A_n = A_n(f)$, $n = (n_1, \ldots, n_p)$. Since $|f| = 1/2$, assumption (a) of Theorem 4.1 is fulfilled. Moreover, the number $\rho(n)$ of zero columns of $A_n$ is easily estimated as follows:

$$\rho(n) \geq c_f N(n) \sum_{k=1}^{p} \frac{1}{n_k},$$

where $c_f > 0$ is independent of $n$. Thus, we come up with the following negative results.

**Theorem 5.1.** For $I_n + A_n$, any sublinear preconditioner of the form $I_n + C_n$, where $C_n$ is a $p$-level PE-based matrix, provides the singular value cluster for which, for some $c(\varepsilon) > 0$ and infinitely many $n$,

$$\gamma_n(\varepsilon) \geq c(\varepsilon) \rho(n),$$

where $\rho(n)$ is defined by (1).

**Corollary.** There exist Hermitian $p$-level Toeplitz matrices for which any sublinear PE-based preconditioner with $U_n = V_n^*$ provides the singular value cluster with $\gamma_n(\varepsilon)$ outliers subject to (2).

These results witness that some well-known sublinear preconditioners are “optimal” for the whole class of multilevel Toeplitz matrices.

Let $f(x_1, \ldots, x_p) > 0$ belong to the Wiener class. Then, by a direct extension of R. Chan’s arguments for the unilevel case, it is shown in [28] that multilevel circulant preconditioners of G. Strang’s and T. Chan’s type provide the general clusters with the number of outliers

$$\gamma_n(\varepsilon) \leq k_f N(n) \sum_{k=1}^{p} \frac{1}{n_k}$$

License or copyright restrictions may apply to redistribution; see http://www.ams.org/journal-terms-of-use
with \( k_f > 0 \) independent of \( n \). In [20] it was found that the same property holds true for T. Chan’s circulants for any positive continuous symbol and, what is more, for all known trigonometric matrix algebra preconditioners. Therefore, we can say that this preconditioning technique, unless very unsatisfactory for large \( p \), is the best we may count on when PE-based preconditioners \( \{V_n^* D_n V_n\} \) are considered.

Note that the trigonometric matrix algebras other than circulants require a different technique [20]. Given a matrix \( A_n \), we choose a preconditioner \( C_n = U_n D_n U_n^* \) of Hermitian pattern so that it minimizes \( \|A_n - C_n\|_F \) over all diagonal matrices \( D_n \). It is clear that the minimum is attained at \( D_n = \operatorname{diag} U_n^* A_n U_n \). Assume that we have \( p \) sequences of grids \( W_{nk} = \{x_{1nk}\} \) and functions \( \Phi_{nk} = \{\phi_{jnk}\} \). Consider the vector functions \( \psi_{nk}(x_k) = [\phi_{1nk}(x_k), \ldots, \phi_{nk_nk}(x_k)]^* \) and, with the notation \( x = (x_1, \ldots, x_p) \), set

\[
\sigma_n(f; x) = \psi^* A_n \psi, \quad \psi = \bigotimes_{1 \leq k \leq p} \psi_{nk}(x_k).
\]

Obviously, \( D_n \) consists of the values of \( \sigma_n(f; x_{i_1 n_1}, \ldots, x_{i_p n_p}) \), where \( 1 \leq i_k \leq n_k \), \( k = 1, \ldots, p \). The crucial observation is now that, in the case of \( p \)-level Toeplitz matrices \( A_n = A_n(f) \), the \( \varepsilon \)-closeness of \( C_n \) and \( A_n \) can be naturally related to the closeness of functions \( f(x) \) and \( \sigma_n(f; x) \). Since a continuous periodic function is uniformly approximated by trigonometric polynomials, it is sufficient (for continuous symbols) to study the case when \( f \) is a trigonometric polynomial.

Instead of the case study of different algebras, when following the above lines it is possible to propose a unifying approach that covers at once all cases of interest. It is based on a matrix interpretation of Korovkin’s results in approximation theory (see [20]), allowing us to study the latter problem only for a finite set of very simple polynomials.

6. Negative results for Hermitian multilevel matrices

Consider PE-based matrices of Hermitian pattern \( (U_n = V_n^*) \) and symbols of the form

\[
f(x_1, \ldots, x_p) = \frac{1}{2} \exp \{i k_1 x_1 + \cdots + i k_p x_p \},
\]

\( k_j \geq 1 \), for \( j = 1, \ldots, p \). Let \( \operatorname{re}(f) \) and \( \operatorname{im}(f) \) denote the real and the imaginary part of \( f \). It is obvious that for any real value \( s \) at least one of the two Hermitian Toeplitz matrices \( A_n(s + \operatorname{re}(f)) \) and \( A_n(s + \operatorname{im}(f)) \) does not admit a superlinear PE-based preconditioner. Therefore, for any fixed positive integer \( k \), for any \( j = 1, \ldots, p \), and any real value \( s \), one of the two matrices

\[
A_n(s + \cos(kx_j)) \quad \text{or} \quad A_n(s + \sin(kx_j))
\]

cannot be superlinearly preconditioned. We cannot say definitely which of the two cannot, but it is reasonable to think that both cannot. In fact, it is easy to see that \( A_n(s + \cos(kx_j)) \) is similar to \( A_n(s + \cos(k(x_j + v))) \) through a unitary diagonal transformation. Choosing \( v = -\frac{k}{2\pi} \), we have

\[
A_n(s + \cos(k(x_j + v))) = A_n(s + \sin(kx_j)),
\]

and, therefore, by supposing that the considered PE-based matrices \( V_n^* D_n V_n \) possess no structural symmetries, we do not see any reason why \( A_n(s + \sin(kx_j)) \) should

License or copyright restrictions may apply to redistribution; see http://www.ams.org/journal-terms-of-use
be better than \( A_n(s + \cos(kx_j)) \) or vice versa. It remains to give a rigorous proof of the preceding qualitative argument. This will be the subject of future research.

References


Dipartimento di Chimica, Fisica e Matematica, Università dell’Insubria - Sede di Como, Via Valleggio 11, 22100 Como, Italy
E-mail address: stefano.serrac@uninsubria.it; serra@monge.dm.unipi.it

Institute of Numerical Mathematics, Russian Academy of Sciences, Gubkina 8, Moscow 117333, Russia
E-mail address: tee@inm.ras.ru