Computers and Mathematics with Applications 56 (2008) 1255–1270

Contents lists available at ScienceDirect

Computers and Mathematics with Applications

journal homepage: www.elsevier.com/locate/camwa

Superlinear convergence for PCG using band plus algebra preconditioners for Toeplitz systems[☆]

D. Noutsos[a,*], P. Vassalos[b]

[a] *Department of Mathematics, University of Ioannina, GR45110, Greece*

[b] *Department of Informatics, Athens University of Economics and Business, GR10434, Greece*

ARTICLE INFO

Article history:
Received 20 January 2008
Accepted 6 February 2008

Keywords:
Toeplitz
Preconditioning
Trigonometric algebras
PCG

ABSTRACT

The paper studies fast and efficient solution algorithms for $n \times n$ symmetric ill conditioned Toeplitz systems $T_n(f)x = b$ where the generating function f is known a priori, real valued, nonnegative, and has isolated roots of even order. The preconditioner that we propose is a product of a band Toeplitz matrix and matrices that belong to a certain trigonometric algebra. The basic idea behind the proposed scheme is to combine the advantages of all components of the product that are well known when every component is used as a stand-alone preconditioner. As a result we obtain a flexible preconditioner which can be applied to the system $T_n(f)x = b$ infusing superlinear convergence to the PCG method. The important feature of the proposed technique is that it can be extended to cover the 2D case, i.e. ill-conditioned block Toeplitz matrices with Toeplitz blocks. We perform many numerical experiments, whose results confirm the theoretical analysis and effectiveness of the proposed strategy.

1. Introduction

In this paper, we introduce and analyze a new approach for the solution, by means of the Preconditioned Conjugate Gradient (PCG) method, of ill conditioned linear systems $Tx = b$ where $T = T_n(f)$ is a Toeplitz matrix. A matrix is called Toeplitz matrix if its (i,j) entry depends only on the difference $i - j$ of the subscripts, i.e. $t_{i,j} = t_{i-j}$. The function $f(x)$ whose Fourier coefficients give the diagonals of $T_n(f)$ i.e.

$$T_{j,k} = t_{j-k} = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(x)\mathrm{e}^{-\mathbf{i}(j-k)x}\mathrm{d}x, \quad 1 \le j, k < n,$$

is called the generating function of $T_n(f)$ and in the rest of the paper we will assume that it is a priori known.

Such kind of matrices arise in a wide variety of fields of pure and applied mathematics such as signal theory, image processing, probability theory, harmonic analysis, control theory etc. Therefore, a fast and effective solver is not only welcome but in fact necessary.

Several direct methods for solving Toeplitz systems have been proposed; the most efficient algorithms are called “superfast” and require $O(n \log^2 n)$ operations to compute the solution. The stability properties of these methods are discussed in [6]. Their main disadvantage is that in 2D they cannot exploit efficiently the block Toeplitz structure, and as a consequence they require an order of $nm^2 \log nm$ arithmetic operations, which is very far from optimum.

[☆] This research was co-funded by the European Union - European Social Fund (ESF) & National Sources, in the framework of the program “Pythagoras I” of the “Operational Program for Education and Initial Vocational Training” of the 3rd Community Support Framework of the Hellenic Ministry of Education.

[*] Corresponding author.
E-mail addresses: dnoutsos@uoi.gr (D. Noutsos), pvassal@aueb.gr (P. Vassalos).

0898-1221/$ – see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.camwa.2008.02.046

We focus on the case where the generating function f is real-valued continuous 2π-periodic and defined on $I = [-\pi, \pi]$, thus the associated Toeplitz matrix is Hermitian.

In the case where f is a positive function, the matrix becomes well-conditioned Hermitian and positive definite. In addition, if f is also an even function, the matrix becomes well-conditioned symmetric and positive definite (spd). In this case, preconditioners belonging to some trigonometric matrix algebras have been proposed to achieve superlinear convergence of the PCG method. Circulant preconditioners have been proposed by Strang [24], by Chan [7] and by Chan and Yeung [11] for well conditioned spd systems. τ preconditioners have been proposed for the same systems by Bini and Di Benedeto [2] and by Di Benedeto [13]. To cover the well conditioned Hermitian positive definite case, Hartley prconditioners have been proposed by Bini and Favati [3] and by Jin [16].

It is well known that preconditioners from any trigonometric matrix algebra cannot support superlinear convergence [17,18], when f has roots. Moreover, there are cases where the corresponding matrices are singular, as e.g., in the case where f is a nonnegative function having roots of even order and the preconditioner is a circulant matrix of Strang type. In this specific case, the system becomes an ill conditioned symmetric positive definite one. Problems with such matrices arise in a variety of applications: signal and image processing, tomography, harmonic analysis and partial differential equations.

Band Toeplitz preconditioners are ideal in this case of ill conditioned systems. They succeed in making the condition number of the preconditioned system independent of the dimension n. First, Chan [8] proposed a band Toeplitz preconditioner generated by a trigonometric polynomial g that matches the roots of f. Chan and Tang [10] extended the preconditioner to the ones based on certain approximation of f. Finally, Serra Capizzano [22] proposed a band Toeplitz preconditioner based on trigonometric polynomial g that matches the roots and on the best trigonometric Chebyshev approximation of the remaining positive part $\frac{f}{g}$.

Preconditioners based on τ algebra have studied by Di Benedeto, Fiorentino and Serra Capizzano [14], by Di Benedeto [12] and by Serra Capizzano [23], while ω-circulant preconditioners have been proposed by Potts and Steidl [21] and by Chan and Ching [9].

Finally, a preconditioner of mixed type, being a product of band Toeplitz matrices and inverses of band Toeplitz matrices, based on the best rational approximation of the remaining positive part, has been studied and proposed by the authors in [19].

In this paper, we propose and study a preconditioner defined as a product of the band Toeplitz matrix generated by g and matrices that belong to any trigonometric algebra and correspond to an approximation of the positive part. The idea underlining our scheme is to combine the well known advantages that each of the components of the product presents when it is used as a stand-alone preconditioner. As a result, we obtain a flexible preconditioner that can be applied to the system $T_n(f)x = b$ infusing superlinear convergence to the PCG method. Convergence theory of the proposed preconditioner is developed and an alternating technique is proposed in cases where convergence is not achieved. Finally, we compare our method with the known techniques.

The paper is organized as follows. In Section 2 we introduce the basic idea for the construction of our preconditioners and study their computational cost. In Section 3 we develop the convergence theory in both cases of using band plus τ preconditioners and band plus circulant ones. For both cases, in Section 4 we propose and study an alternating smoothing technique, where the convergence properties studied in Section 3 do not hold. Section 5 is devoted to applications, numerical experiments and concluding remarks.

2. Band plus Algebra preconditioners

Let $f \in \mathcal{C}_{2\pi}$ be a 2π-periodic nonnegative function with roots $x_0, x_1, \dots, x_l$ of multiplicities $2k_1, 2k_2, \dots, 2k_l$ respectively, with $k_1 + k_2 + \cdots + k_l = k$. Then f can be written as a product $g \cdot w$ where

$$g(x) = \prod_{i=1}^{l} (2 - 2\cos(x - x_i))^{k_i} \tag{2.1}$$

and with $w(x) > 0$ for every $x \in [-\pi, \pi]$.

We define as a preconditioner for the system

$$T_n(f)x = b, \tag{2.2}$$

the product of matrices

$$K_n^{\mathcal{A}}(f) = \mathcal{A}_n(\sqrt{w})T_n(g)\mathcal{A}_n(\sqrt{w}) = \mathcal{A}_n(h)T_n(g)\mathcal{A}_n(h) \tag{2.3}$$

with $\mathcal{A}_n \in \{\tau, \mathcal{C}, \mathcal{H}\}$, where $\{\tau, \mathcal{C}, \mathcal{H}\}$ is the set of matrices belonging to τ, Circulant and Hartley algebra, respectively. We have put, for simplicity, $h = \sqrt{w}$.

It is obvious from the construction of K, that it fulfils the fundamental properties that each preconditioner must have, i.e the positive definiteness and symmetry (Hermitian).

Although the idea of using as preconditioners for the system (2.2) a product of band Toeplitz matrices with τ, circulant or Hartley ones is not new (see e.g [9] or [23]), what we propose is more general and flexible in the sense that it can use as $\mathcal{A}_n$ any matrix belonging to $\{\tau, \mathcal{C}, \mathcal{H}\}$, can treat both symmetric and Hermitian systems ([23]), and can be efficiently extended to the 2D case.

2.1. Construction of the preconditioner-Computation cost

For the band Toeplitz matrix $T_n(g)$ things are straightforward. To construct $\mathcal{A}_n(h)$ we use the relation

$$\mathcal{A}_n(h) = Q_n \cdot \text{Diag}\,(h(\mathbf{u}^n)) \cdot Q_n^H,$$

where the entries of the vector $\mathbf{u}^n$ are $\mathbf{u}_i^n = \frac{2\pi(i-1)}{n}$, $i = 1(1)n$ and Q_n is the Fourier matrix F_n for the circulant case or the matrix $\text{Re}(F_n) + \text{Im}(F_n)$ for the Hartley case. For the τ case, we have $\mathbf{u}_i^n = \frac{\pi i}{(n+1)}$, $i = 1(1)n$ and $Q_n = \sqrt{\frac{2}{n+1}}[\sin(j\mathbf{u}_i^n)]_{i,j=1}^n$.

The evaluation of the function h at the points $\mathbf{u}^n$ requires the evaluation of the function w and the computation of real square roots, which can be done by a fast and simple algorithm based on "Newton's Method" and is of $O(n)$ ops. In any case, the above procedure does not incur in the total asymptotic complexity of the method as it is implemented once per every n. The computation $Q \cdot \mathbf{v}$ is performed via Fast Fourier Transforms (or Fast Sine Transforms in the τ case) and requires $O(n \log n)$ ops. Finally, the 'inversion' of $T_n(g)$ can be done in $O(n \log p + p \log^2 p \log \frac{n}{p})$ ops, where p is its bandwidth, using the algorithm proposed in [4] or even better in $O(n)$ using the multigrid technique proposed in [15]. So, the total optimal cost of $O(n \log n)$ is preserved per each iteration of PCG.

3. Convergence Theory

3.1. Convergence of the method: τ case

We start with the case where $\mathcal{A}_n \in \tau$. We will show that the main mass of the eigenvalues of the preconditioned matrix

$$(\tau_n(h)T_n(g)\tau_n(h))^{-1}T_n(f) \tag{3.1}$$

is clustered around unity. Before we give the main results for this case, we report a useful lemma.

Lemma 3.1. *Let $w \in \mathcal{C}_{2\pi}$ be a positive and even function. Then, for any positive ϵ, there exist N and $M > 0$ such that for every $n > N$, at most M eigenvalues of the matrix $T_n(w) - \tau_n(w)$ have absolute value greater than ϵ.*

Proof. See [23], Theorem 2.1. □

Theorem 3.2. *Let $T_n(f)$ be the Toeplitz matrix produced by a nonnegative function f in $\mathcal{C}_{2\pi}$ which can be written as $f = g \cdot w$, where g the trigonometric polynomial of order k as it given by (2.1) and $w = h^2$ is a strictly positive even function belonging to $\mathcal{C}_{2\pi}$. Then, for every $\epsilon > 0$ there exist N and $\hat{M} > 0$ such that for every $n > N$, at most $\hat{M}$ eigenvalues of the preconditioned matrix (3.1) lie outside the interval $(1-\epsilon, 1+\epsilon)$.*

Proof. We begin with the observation that the matrix $T_n(f)$ can be written (see [5]) as $T_n(g)T_n(w)+L_1$, where L_1 is a low rank matrix. Taking into account the specific form of L_1, which contains only nonzero columns at the first and last k columns, we obtain that $\text{rank}(L_1) = \text{rank}(L_1^T) = 2k$ and $\text{rank}(L_1 + L_1^T) = 4k$. From the close relationship between τ matrices and band Toeplitz matrices, we have that

$$\begin{aligned} T_n(f) &= \frac{1}{2}(T_n(g)T_n(w) + L_1) + \frac{1}{2}(T_n(w)T_n(g) + L_1^T) \\ &= \frac{1}{2}((\tau_n(g) + L_2)T_n(w) + L_1) + \frac{1}{2}(T_n(w)(\tau_n(g) + L_2) + L_1^T) \\ &= \frac{1}{2}\tau_n(g)T_n(w) + \frac{1}{2}\tau_n(g)T_n(w) + L_3, \end{aligned}$$

where L_2 and L_3 are low rank symmetric matrices. More specifically, as L_2 has nonzero elements only at the upper left and lower right corner, the factor $L_2T_n(\omega) + T_n(\omega)L_2$ has nonzero entries only in the $k-1$ first and last rows and columns, i.e it is a border matrix. So, the rank of the matrix L_3 is at most $4k$. To study the spectrum of the preconditioned matrix $K_n^\tau(f)^{-1}T_n(f)$ with $K_n^\tau(f)^{-1}$ as in (2.3), we consider the symmetric form of it $\hat{T}_n = T_n(g)^{-\frac{1}{2}}\tau_n(h)^{-1}T_n(f)\tau_n(h)^{-1}T_n(g)^{-\frac{1}{2}}$, which is similar to the first one. So

$$\begin{aligned} \hat{T}_n &= T_n(g)^{-\frac{1}{2}}\tau_n(h)^{-1}T_n(f)\tau_n(h)^{-1}T_n(g)^{-\frac{1}{2}} \\ &= \frac{1}{2}T_n(g)^{-\frac{1}{2}}\tau_n(h)^{-1}\left(\tau_n(g)T_n(w) + T_n(w)\tau_n(g) + L_3\right)\tau_n(h)^{-1}T_n(g)^{-\frac{1}{2}} \\ &= \frac{1}{2}T_n(g)^{-\frac{1}{2}}\tau_n(g)\tau_n(h)^{-1}T_n(w)\tau_n(h)^{-1}T_n(g)^{-\frac{1}{2}} + \frac{1}{2}T_n(g)^{-\frac{1}{2}}\tau_n(h)^{-1}T_n(w)\tau_n(h)^{-1}\tau_n(g)T_n(g)^{-\frac{1}{2}} + L_4 \\ &= \frac{1}{2}T_n(g)^{-\frac{1}{2}}(T_n(g) - L_2)\tau_n(h)^{-1}T_n(w)\tau_n(h)^{-1}T_n(g)^{-\frac{1}{2}} + \frac{1}{2}T_n(g)^{-\frac{1}{2}}\tau_n(h)^{-1}T_n(w)\tau_n(h)^{-1}(T_n(g) - L_2)T_n(g)^{-\frac{1}{2}} + L_4 \\ &= \frac{1}{2}T_n(g)^{\frac{1}{2}}\tau_n(h)^{-1}T_n(w)\tau_n(h)^{-1}T_n(g)^{-\frac{1}{2}} + \frac{1}{2}T_n(g)^{-\frac{1}{2}}\tau_n(h)^{-1}T_n(w)\tau_n(h)^{-1}T_n(g)^{\frac{1}{2}} + L_5, \end{aligned}$$

where L_3, L_4 and L_5 are dense symmetric matrices of low rank, with $\operatorname{rank}(L_4) = \operatorname{rank}(L_3)$ and therefore the rank of L_5 is at most $8k - 4$ (the rank of L_4 plus twice the rank of L_2).

From Lemma 3.1 we obtain that for the choice of $\epsilon_h > 0$, there exist a low rank (of constant rank) matrix L_6 and a matrix E of small norm ($\|E\|_2 \le \epsilon_h$), such that

$$\tau_n(h)^{-1}T_n(w)\tau_n(h)^{-1} = I + E + L_6, \tag{3.2}$$

where I is the n-dimensional identity matrix. Hence

$$\begin{aligned}\hat{T}_n &= \frac{1}{2}T_n(g)^{\frac{1}{2}}(I + E + L_6)T_n(g)^{-\frac{1}{2}} + \frac{1}{2}T_n(g)^{-\frac{1}{2}}(I + E + L_6)T_n(g)^{\frac{1}{2}} + L_5\\ &= I + \frac{1}{2}T_n(g)^{\frac{1}{2}}ET_n(g)^{-\frac{1}{2}} + \frac{1}{2}T_n(g)^{-\frac{1}{2}}ET_n(g)^{\frac{1}{2}} + L,\end{aligned}$$

where L is a symmetric low rank matrix with its rank being no greater than the sum of the rank of L_5 and the double of the one L_6.

The proof of the main issue that $\hat{T}_n$ has a clustering at one, is reduced to the proof that for every $\epsilon > 0$, there exists $\epsilon_h > 0$, with $\|E\|_2 \le \epsilon_h$, such that all the eigenvalues of the matrix

$$\hat{A}_n = \frac{1}{2}T_n(g)^{\frac{1}{2}}ET_n(g)^{-\frac{1}{2}} + \frac{1}{2}T_n(g)^{-\frac{1}{2}}ET_n(g)^{\frac{1}{2}}$$

belong in the interval $(-\epsilon, \epsilon)$. Equivalently, since $\hat{A}_n$ is symmetric, we have to prove that both matrices $\epsilon I + \hat{A}_n$ and $\epsilon I - \hat{A}_n$ are positive definite matrices.

First, we prove that $\epsilon I + \hat{A}_n$ is positive definite. This is equivalent to proving that

$$T_n(g)^{\frac{1}{2}}(\epsilon I + \hat{A}_n)T_n(g)^{\frac{1}{2}} = \epsilon T_n(g) + \frac{1}{2}T_n(g)E + \frac{1}{2}ET_n(g)$$

is a positive definite matrix. For this, we consider a normalized vector $x \in \mathbb{R}^n$, ($\|x\|_2 = 1$) and take the Rayleigh quotient

$$\mathrm{r} = \epsilon x^T T_n(g)x + \frac{1}{2}x^T T_n(g)Ex + \frac{1}{2}x^T ET_n(g)x = \epsilon x^T T_n(g)x + x^T T_n(g)Ex.$$

The norm of the vector $y = Ex$ is given by

$$\hat{\epsilon} = \|y\|_2 = \|Ex\|_2 \le \|E\|_2\|x\|_2 \le \epsilon_h.$$

Let z be the normalized vector of y, so $y = \hat{\epsilon}z$; then the Rayleigh quotient takes the form

$$r = \epsilon x^T T_n(g)x + \hat{\epsilon}x^T T_n(g)z. \tag{3.3}$$

The second term of (3.3) takes the minimum value for z being the normalized vector of $-T_n(g)x$. So,

$$\begin{aligned}r &\ge \epsilon x^T T_n(g)x - \hat{\epsilon}\frac{x^T T_n(g)^2 x}{\|T_n(g)x\|_2} \ge \epsilon\|T_n(g)^{\frac{1}{2}}x\|_2 - \epsilon_h\frac{\|T_n(g)x\|_2^2}{\|T_n(g)x\|_2}\\ &= \epsilon\|T_n(g)^{\frac{1}{2}}x\|_2 - \epsilon_h\|T_n(g)x\|_2 \ge \epsilon\|T_n(g)^{\frac{1}{2}}x\|_2 - \epsilon_h\|T_n(g)^{\frac{1}{2}}\|_2\|T_n(g)^{\frac{1}{2}}x\|_2\\ &= \left(\epsilon - \epsilon_h\|T_n(g)^{\frac{1}{2}}\|_2\right)\|T_n(g)^{\frac{1}{2}}x\|_2.\end{aligned}$$

Since the operator $T(g)$ is bounded, we can choose the value of ϵ_h to be such that

$$\epsilon > \epsilon_h\|T_n(g)^{\frac{1}{2}}\|_2, \tag{3.4}$$

so that the Rayleigh quotient r will be positive since $\|x\|_2 = 1$. This holds true for every choice of x, so the matrix $\epsilon I + \hat{A}_n$ is a positive definite matrix.

To prove that the second matrix $\epsilon I - \hat{A}_n$ is positive definite we follow exactly the same argumentation and we end up with

$$r = \epsilon x^T T_n(g)x - \hat{\epsilon}x^T T_n(g)z.$$

in the place of (3.3). Then, the second term takes its maximum value for z being the normalized vector of $T_n(g)x$. After that, the proof follows the same step and the same conclusion is deduced. □

We will prove now the important feature that our preconditioner fulfils and leads to superlinear convergence of PCG. The clustering of the eigenvalues around 1 has been proven in Theorem 3.2. So, we have to prove that the outliers are uniformly far away from zero and from infinity. For this we will study Rayleigh quotients of the preconditioned matrix:

$$\lambda_{\min}(K_n^{\tau^{-1}}T_n(f)) = \inf_{x\in\mathbb{R}^n}\frac{x^T K_n^\tau(f)^{-\frac{1}{2}}T_n(f)K_n^\tau(f)^{-\frac{1}{2}}x}{x^T x} = \inf_{x\in\mathbb{R}^n}\frac{x^T T_n(f)x}{x^T K_n^\tau(f)x} \tag{3.5}$$

and

$$\lambda_{\max}(K_n^{\tau\,-1}T_n(f)) = \sup_{x\in\mathbb{R}^n}\frac{x^TK_n^{\tau}(f)^{-\frac{1}{2}}T_n(f)K_n^{\tau}(f)^{-\frac{1}{2}}x}{x^Tx} = \sup_{x\in\mathbb{R}^n}\frac{x^TT_n(f)x}{x^TK_n^{\tau}(f)x}.$$

Thus, we have to study the range of the Rayleigh quotient

$$\frac{x^TT_n(f)x}{x^TK_n^{\tau}(f)x} = \frac{x^TT_n(f)x}{x^T\tau_n(h)T_n(g)\tau_n(h)x} = \frac{x^TT_n(f)x}{x^TT_n(g)x}\cdot\frac{x^TT_n(g)x}{x^T\tau_n(h)T_n(g)\tau_n(h)x}.$$

It is well known that the range of the first Rayleigh quotient is contained in the range of the function $w=\frac{f}{g}$ which is positive and far from zero and infinity. Therefore, we have to prove that

$$\begin{aligned}&\liminf_{n\to\infty}\inf_{x\in\mathbb{R}^n}\frac{x^TT_n(g)x}{x^T\tau_n(h)T_n(g)\tau_n(h)x} > 0,\\ &\limsup_{n\to\infty}\sup_{x\in\mathbb{R}^n}\frac{x^TT_n(g)x}{x^T\tau_n(h)T_n(g)\tau_n(h)x} < \infty.\end{aligned} \tag{3.6}$$

We will prove only the first inequality of (3.6). The proof of the second one is similar. This is obtained from the observations that

$$\limsup_{n\to\infty}\sup_{x\in\mathbb{R}^n}\frac{x^TT_n(g)x}{x^T\tau_n(h)T_n(g)\tau_n(h)x} = \infty \Leftrightarrow \liminf_{n\to\infty}\inf_{x\in\mathbb{R}^n}\frac{x^T\tau_n(h)T_n(g)\tau_n(h)x}{x^TT_n(g)x} = 0$$

and

$$\liminf_{n\to\infty}\inf_{x\in\mathbb{R}^n}\frac{x^T\tau_n(h)T_n(g)\tau_n(h)x}{x^TT_n(g)x} = \liminf_{n\to\infty}\inf_{x\in\mathbb{R}^n}\frac{x^TT_n(g)x}{x^T\tau_n(h^{-1})T_n(g)\tau_n(h^{-1})x}.$$

So, the proof of the second inequality of (3.6) is equivalent to the proof of the first one with the function h^{-1} in the place of h.

By inverting the ratio of the first inequality of (3.6) it is equivalent to proving that

$$\limsup_{n\to\infty}\sup_{x\in\mathbb{R}^n}\frac{x^T\tau_n(h)T_n(g)\tau_n(h)x}{x^TT_n(g)x} < \infty, \tag{3.7}$$

so, we have to study the ratio

$$r_x = \frac{x^T\tau_n(h)T_n(g)\tau_n(h)x}{x^TT_n(g)x}. \tag{3.8}$$

It is well known that the band Toeplitz matrix $T_n(g)$ is written as a τ plus a Hankel matrix

$$T_n(g) = \tau_n(g) + H_n(g), \tag{3.9}$$

where $H_n(g)$ is the Hankel matrix of rank $2(k-1)$ of the form

$$H_n(g) = E_n(g) + E_n(g)^R, \tag{3.10}$$

with

$$E_n(g) = \text{Hankel}(g_2, g_3, \ldots, g_k, 0, \ldots, 0) \tag{3.11}$$

and $E_n(g)^R$ is obtained from the matrix $E_n(g)$ by taking all its rows and columns in reverse order. The entries g_i are the Fourier coefficients of the trigonometric polynomial g ($g(x) = g_0 + 2g_1\cos(x) + 2g_2\cos(2x) + \cdots + 2g_k\cos(kx)$). In the special case where the root is 0 of multiplicity $2k$, we have that $g_i = \binom{2k}{k-i}$. It is obvious that for $k=1$, $H_n(g)=0$, which means that $T_n(g)$ (the Laplace matrix) is a τ matrix and the problem is solved. In the case where $k=2$, we have that $H_n(g)$ is a semi-positive definite matrix of rank 2 with just ones in the positions $(1,1)$ and (n,n) and zeros elsewhere. In the case $k>2$, the matrix $H_n(g)$ becomes indefinite. We denote by Δ the $(k-1)\times(k-1)$ matrix formed by the first $k-1$ rows and columns of $E_n(g)$:

$$\Delta = \begin{pmatrix} g_2 & g_3 & \cdots & g_k \\ g_3 & & ⋰ & 0 \\ \vdots & ⋰ & & \vdots \\ g_k & 0 & \cdots & 0 \end{pmatrix} \tag{3.12}$$

and by Δ^R, the matrix obtained from Δ by taking all its rows and columns in reverse order. For an n-dimensional vector x we denote by $\overline{x}^{(m)}$ and by $\underline{x}^{(m)}$ the m-dimensional vectors formed from the first and last m entries of x, respectively.

Recalling ratio (3.8), we get

$$\begin{aligned} r_x &= \frac{x^T\tau_n(h)T_n(g)\tau_n(h)x}{x^TT_n(g)x} = \frac{x^T\tau_n(h)\tau_n(g)\tau_n(h)x + x^T\tau_n(h)H_n(g)\tau_n(h)x}{x^T\tau_n(g)x + x^TH_n(g)x} \\ &= \frac{x^T\tau_n(h^2g)x + x^T\tau_n(h)H_n(g)\tau_n(h)x}{x^T\tau_n(g)x + \overline{x}^{(k-1)^T}\Delta\overline{x}^{(k-1)} + \underline{x}^{(k-1)^T}\Delta^R\underline{x}^{(k-1)}}. \end{aligned} \tag{3.13}$$

Lemma 3.3. *Let x be a normalized n-dimensional vector ($\|x\|_2 = 1$) and the sequence of the vectors $\overline{x}^{(k-1)}$ be bounded, i.e. $0 < c \le \|\overline{x}^{(k-1)}\|_2 \le 1$ for all n or the sequence of the vectors $\underline{x}^{(k-1)}$ is bounded i.e. $0 < c \le \|\underline{x}^{(k-1)}\|_2 \le 1$ for all n, with c being constant independent of n; then the ratio r_x is bounded.*

Proof. The assumption $0 < c \le \|\overline{x}^{(k-1)}\|_2 \le 1$ or $0 < c \le \|\underline{x}^{(k-1)}\|_2 \le 1$ means that $\|\overline{x}^{(k-1)}\|_2 = O(1)\bigcap\Omega(1)$ or $\|\underline{x}^{(k-1)}\|_2 = O(1)\bigcap\Omega(1)$, respectively. Without loss of generality, we suppose that $\|\overline{x}^{(k-1)}\|_2 = O(1)\bigcap\Omega(1)$. The proof for the case where $\|\underline{x}^{(k-1)}\|_2 = O(1)\bigcap\Omega(1)$ being the same. It is easily proved that there is a constant integer m independent of n such that $\|\overline{x}^{(m)}\|_2 = O(1)\bigcap\Omega(1)$ and $\|y^{(k)}\|_2 = o(1)$, where $y^{(k)}$ is the k-dimensional vector of the entries of x followed by the vector $\overline{x}^{(m)}$. This is true since otherwise, there would be an infinitely large integer m, depending on n, such that every block of size k of the vector $\overline{x}^{(m)}$ would have constant norm independent of n. The latter is a contradiction, since then $\|\overline{x}^{(m)}\|_2 \to \infty$. Since both the numerator and the denominator of the ratio in (3.8) are bounded from above, to prove that this ratio is bounded is equivalent to proving that the denominator $x^TT_n(g)x$ is bounded from below far from zero for x of unit Euclidean norm. For this, we write the matrix $T_n(g)$ and the vector x in the following block form:

$$T_n(g) = \left(\begin{array}{c|c} T_m(g) & G \mid 0 \\ \hline \begin{array}{c} G^T \\ \hline 0 \end{array} & T_{n-m}(g) \end{array}\right), \qquad x = \left(\begin{array}{c} \overline{x}^{(m)} \\ \hline y^{(k)} \\ \hline z \end{array}\right),$$

where G is an $m \times k$ Toeplitz matrix with nonzero entries only in the k diagonals in the left bottom corner. We take now the denominator:

$$\begin{aligned} x^TT_n(g)x &= (\overline{x}^{(m)^T}|y^{(k)^T}|z^T)\left(\begin{array}{c|c} T_m(g) & G \mid 0 \\ \hline \begin{array}{c} G^T \\ \hline 0 \end{array} & T_{n-m}(g) \end{array}\right)\left(\begin{array}{c} \overline{x}^{(m)} \\ \hline y^{(k)} \\ \hline z \end{array}\right) \\ &= \overline{x}^{(m)^T}T_m(g)\overline{x}^{(m)} + 2\overline{x}^{(m)^T}Gy^{(k)} + (y^{(k)^T}|z^T)T_{n-m}(g)\left(\begin{array}{c} y^{(k)} \\ \hline z \end{array}\right). \end{aligned} \tag{3.14}$$

Since $T_m(g)$ and $T_{n-m}(g)$ are positive definite matrices, the first and the third terms in the sum of (3.14) are both positive numbers. The minimum value of the first term depends only on m, which is constant, and is of order $\frac{1}{m^{2k}}$ independently of n, and far from zero. The third term depends on n and may take small values near zero. The second term is the only one which may take negative values, but

$$|2\overline{x}^{(m)^T}Gy^{(k)}| = 2\|\overline{x}^{(m)^T}Gy^{(k)}\|_2 \le 2\|\overline{x}^{(m)}\|_2\|G\|_2\|y^{(k)}\|_2 = o(1),$$

since $\|y^{(k)}\|_2 = o(1)$, and the other norms are constants. As a consequence, the first term is absolutely greater in order of magnitude than the second one, which characterizes the bounded behavior of all the sum, and our assertion has been proven. □

It remains to study the quantity r_x for vector sequences x such that

$$\|\overline{x}^{(m)}\|_2 = o(1) \quad \text{and} \quad \|\underline{x}^{(m)}\|_2 = o(1) \tag{3.15}$$

for each constant m independent of n. First, we write the vector x as a convex combination of the eigenvectors v_is of τ algebra, with entries $(v_i)_j = \sqrt{\frac{2}{n+1}}\sin(\frac{\pi ij}{n+1})$:

$$x = \sum_{i=1}^n c_iv_i, \qquad \sum_{i=1}^n |c_i|^2 = 1. \tag{3.16}$$

We denote by **D** the denominator and by **N** the numerator of the ratio r_x of (3.13). So the denominator is given by

$$\begin{aligned} \mathbf{D} &= \sum_{i=1}^n c_iv_i^T\tau_n(g)\sum_{i=1}^n c_iv_i + \sum_{i=1}^n c_iv_i^TH_n(g)\sum_{i=1}^n c_iv_i \\ &= \sum_{i=1}^n c_i^2g_i + \sum_{i=1}^n c_iv_i^TH_n(g)\sum_{i=1}^n c_iv_i \\ &= \sum_{i=1}^n c_i^2g_i + \sum_{i=1}^n c_i\overline{v}_i^T\Delta\sum_{i=1}^n c_i\overline{v}_i + \sum_{i=1}^n c_i\underline{v}_i^T\Delta^R\sum_{i=1}^n c_i\underline{v}_i, \end{aligned} \tag{3.17}$$

while the numerator is given by

$$\begin{aligned}\mathbf{N} &= \sum_{i=1}^{n} c_i v_i^T \tau_n(h^2 g) \sum_{i=1}^{n} c_i v_i + \sum_{i=1}^{n} c_i v_i^T \tau_n(h) H_n(g) \tau_n(h) \sum_{i=1}^{n} c_i v_i \\ &= \sum_{i=1}^{n} c_i^2 h_i^2 g_i + \sum_{i=1}^{n} c_i h_i v_i^T H_n(g) \sum_{i=1}^{n} c_i h_i v_i \\ &= \sum_{i=1}^{n} c_i^2 h_i^2 g_i + \sum_{i=1}^{n} c_i h_i \bar{v}_i^T \varDelta \sum_{i=1}^{n} c_i h_i \bar{v}_i + \sum_{i=1}^{n} c_i h_i \underline{v}_i^T \varDelta^R \sum_{i=1}^{n} c_i h_i \underline{v}_i, \end{aligned} \tag{3.18}$$

where $h_i = h(\frac{\pi i}{n+1}) > h_{\min} > 0$ and $g_i = g(\frac{\pi i}{n+1}) = (2 - 2\cos(\frac{\pi i}{n+1}))^k = (2\sin(\frac{\pi i}{2(n+1)}))^{2k}$. For simplicity, we have put $\bar{v}_i$ and $\underline{v}_i$ instead of $\bar{v}_i^{(k-1)}$ and $\underline{v}_i^{(k-1)}$, respectively. The first sum in both numerator and denominator is positive and we call it the τ-term, since it corresponds to the Rayleigh quotient of a τ matrix. We call the other two terms, corresponding to the low rank correction matrices $\varDelta$ and $\varDelta^R$, correction terms. The correction terms may take negative values. It is obvious that the τ-terms of the numerator and the denominator coincide with each other in order of magnitude for all the choices of the vector x, since

$$\sum_{i=1}^{n} c_i^2 h_i^2 g_i = \hat{h}^2 \sum_{i=1}^{n} c_i^2 g_i, \quad 0 < h_{\min} \le \hat{h} \le h_{\max} < \infty.$$

So, if the τ-terms are greater, in order of magnitude, than the associated correction terms, then r_x is bounded. The only case where r_x tends to infinity is that where the correction terms in the numerator exceed, in order of magnitude, either the associated τ-term and/or that of the denominator. We will try to find such cases by comparing the τ-terms with the correction terms. Since the correction term corresponding to $\varDelta^R$ behaves exactly as the one corresponding to $\varDelta$, for simplicity we will compare only the τ-terms with the correction terms corresponding to $\varDelta$. In other words, we consider that $|\bar{x}^T \varDelta \bar{x}|$ is greater than or equal to $|\underline{x}^T \varDelta^R \underline{x}|$, in order of magnitude. Given $\{N_n\}$ with $N_n = \{1, 2, \dots, n\}$, we define the sequence of subsets $\{S_n\}$ such that

$$\begin{aligned} &(1)\ S_n \subset N_n \forall n \\ &(2)\ \forall i_n \text{ sequence to which } i_k \in S_k \text{ we have } \lim_{n\to\infty} \frac{i_n}{n} = 0\ (i_n = o(n)). \end{aligned} \tag{3.19}$$

Accordingly the complementary sequence of subsets, $\{Q_n\}$ is defined as

$$Q_n = N_n \setminus S_n. \tag{3.20}$$

It is obvious that the border of the above subsets S_n and Q_n is not clear, but this does not present any problem in the analysis that follows. However, we have to be careful to take only sequences belonging to $o(n)$ when dealing with $\{S_n\}$. We write the vector x as the sum $x = x_S + x_Q$ where

$$x_S = \sum_{i\in S_n} c_i v_i, \qquad x_Q = \sum_{i\in Q_n} c_i v_i. \tag{3.21}$$

We denote also by $\bar{x}_S = \sum_{i\in S_n} c_i \bar{v}_i$, $\underline{x}_S = \sum_{i\in S_n} c_i \underline{v}_i$, $\bar{x}_Q = \sum_{i\in Q_n} c_i \bar{v}_i$ and $\underline{x}_Q = \sum_{i\in Q_n} c_i \underline{v}_i$. In other words we separate the eigenvectors into those that correspond to "small" eigenvalues ($o(1)$) and those that correspond to "large" ones ($O(1) \bigcap \varOmega(1)$).

We consider the sequences

$$\{q_n\}_n = \left\{ \sum_{i\in Q_n} c_i^2 \right\}_n \quad \text{and} \quad \{s_n\}_n = \left\{ \sum_{i\in S_n} c_i^2 \right\}_n . \tag{3.22}$$

Lemma 3.4. *Let x be such that $\|\bar{x}^{(k-1)}\|_2 = o(1)$ and $\|\underline{x}^{(k-1)}\|_2 = o(1)$ and the sequence $\{q_n\}_n$ of (3.22) is bounded, i.e. $0 < c \le q_n \le 1$; then the ratio r_x is bounded.*

Proof. In this case, we have

$$x^T \tau_n(g) x = x_S^T \tau_n(g) x_S + x_Q^T \tau_n(g) x_Q = \sum_{i\in S_n} c_i^2 g_i + \sum_{i\in Q_n} c_i^2 g_i \sim c > 0,$$

since the eigenvalues of the second sum are bounded from below. On the other hand, we have

$$|\bar{x}^{(k-1)^T} \varDelta \bar{x}^{(k-1)}| \le \|\varDelta\|_2 \|\bar{x}^{(k-1)}\|_2^2 = o(1),$$

since $\|\bar{x}^{(k-1)}\|_2 = o(1)$. We get the same conclusion for the term $|\underline{x}^{(k-1)^T} \varDelta \underline{x}^{(k-1)}|$. So, the τ-term is the dominant term which is bounded from below. Since the numerator is bounded from above, r_x is bounded. □

Lemma 3.5. *Let x be such that $\|\bar{x}^{(k-1)}\|_2 = o(1)$ and $\|\underline{x}^{(k-1)}\|_2 = o(1)$ and for the sequences $\{s_n\}_n$ and $\{q_n\}_n$ of (3.22) it holds that $\lim_{n\to\infty} s_n = 1$, $\lim_{n\to\infty} q_n = 0$ with $\|\bar{x}_S\|_2 = o\left((q_n)^{\frac{1}{2}}\right)$; then the ratio r_x is bounded.*

Proof. We suppose that the sequence $\{q_n\}_n$ tends to zero monotonically, since otherwise it can be split into monotonic subsequences.

The τ-term gives:

$$x^T \tau_n(g) x = \sum_{i\in S_n} c_i^2 g_i + \sum_{i\in Q_n} c_i^2 g_i, \tag{3.23}$$

while the correction term gives:

$$\bar{x}^T \Delta \bar{x} = (\bar{x}_S + \bar{x}_Q)^T \Delta (\bar{x}_S + \bar{x}_Q) = \bar{x}_S^T \Delta \bar{x}_S + 2\bar{x}_S^T \Delta \bar{x}_Q + \bar{x}_Q^T \Delta \bar{x}_Q. \tag{3.24}$$

For the vector $\bar{x}_Q$ we have

$$\|\bar{x}_Q\|_2 = \left\| \sum_{i\in Q_n} c_i \bar{v}_i \right\|_2 \le \sum_{i\in Q_n} |c_i| \|\bar{v}_i\|_2 \le \left(\sum_{i\in Q_n} c_i^2\right)^{\frac{1}{2}} \left(\sum_{i\in Q_n} \|\bar{v}_i\|_2^2\right)^{\frac{1}{2}} \sim (q_n)^{\frac{1}{2}},$$

since $\|\bar{v}_i\|_2^2 \sim \frac{1}{n}$, for all $i \in N_n$ and the cardinality of Q_n is $n - o(n) \sim n$. So, $\|\bar{x}_Q\|_2 = O\left((q_n)^{\frac{1}{2}}\right)$. Let $\|\bar{x}_Q\|_2 = o\left((q_n)^{\frac{1}{2}}\right)$, then $|\bar{x}_Q^T \Delta \bar{x}_Q| \le \|\Delta\|_2 \|\bar{x}_Q\|_2^2 = o(q_n)$, which means that the second sum of (3.23) exceeds the last one of (3.24), so,

$$x_Q^T T_n(g) x_Q = \sum_{i\in Q_n} c_i^2 g_i + \bar{x}_Q^T \Delta \bar{x}_Q + \underline{x}_Q^T \Delta^R \underline{x}_Q \sim q_n. \tag{3.25}$$

In the case where $\|\bar{x}_Q\|_2 \sim (q_n)^{\frac{1}{2}}$, we consider the quantity $x_Q^T T_n(g) x_Q$ and normalize the vector x_Q to the vector $\hat{x}_Q$ by multiplying by a number of order $(q_n)^{-\frac{1}{2}}$, such that $\|\hat{x}_Q\|_2 = 1$. If we consider the vector $\hat{x}_Q$ in the place of x, which means that there are no vectors of indices belonging to S_n in the convex combination, we get that $\sum_{i\in Q_n} c_i^2 = 1$ for the new coefficients c_is. Since $\|\bar{x}_Q\|_2 \sim (q_n)^{\frac{1}{2}}$, we obtain that $\|\bar{\hat{x}}_Q\|_2 \sim c > 0$. From Lemma 3.3, by replacing $\hat{x}_Q$ in the place of x, we obtain that $\hat{x}_Q^T T_n(g) \hat{x}_Q$ is bounded from below. If we come back to the quantity $x_Q^T T_n(g) x_Q$ by dividing the vector $\hat{x}_Q$ by the same number, we obtain the validity of (3.25). For the estimation of the associated term $x_Q^T \tau_n(h)^T T_n(g) \tau(h) x_Q$ of the numerator, we follow exactly the same steps in the proof by considering the vector $\tau_n(h)x$ in the place of x. So, we obtain

$$x_Q^T \tau_n(h)^T T_n(g) \tau_n(h) x_Q \sim x_Q^T T_n(g) x_Q \sim q_n. \tag{3.26}$$

Under the last assumption, $\|\bar{x}_S\|_2 = o\left((q_n)^{\frac{1}{2}}\right)$, the remaining terms of (3.24) $\bar{x}_S^T \Delta \bar{x}_S$ and $2\bar{x}_S^T \Delta \bar{x}_Q$ are both absolutely smaller than q_n in order of magnitude. Exactly the same happens with the corresponding terms of the numerator. So, the order of the denominator of r_x is just the order of $\sum_{i\in S_n} c_i^2 g_i$ if it exceeds q_n or q_n otherwise, while the one of the numerator is just the order of $\sum_{i\in S_n} c_i^2 h_i^2 g_i$ if it exceeds q_n or q_n otherwise. In any case, the numerator and the denominator coincide with each other, meaning that r_x is bounded. □

A useful definition is given here.

Definition 3.6. A positive and even function $h \in \mathcal{C}_{2\pi}$ is said to be an (m, ρ)-smooth function if it is an m times differentiable function in an open region of the point $\rho \in (-\pi, \pi)$ with $h^{(j)}(\rho) = 0, j = 1(1)m - 1$ and $h^{(m)}(\rho)$ being bounded.

Lemma 3.7. *Let x be such that $\|\bar{x}^{(k-1)}\|_2 = o(1)$ and $\|\underline{x}^{(k-1)}\|_2 = o(1)$ and for the sequences $\{s_n\}_n$ and $\{q_n\}_n$ of (3.22) it holds that $\lim_{n\to\infty} s_n = 1$, $\lim_{n\to\infty} q_n = 0$ with $\|\bar{x}_S\|_2 = \Omega\left((q_n)^{\frac{1}{2}}\right)$. Let also that h is a $(k-1, 0)$-smooth function. Then, the ratio r_x is bounded.*

Proof. The proof follows exactly the same steps of Lemma 3.5 to obtain the same results until (3.26). In the sequel, we use the assumption that the function h is a $(k-1, 0)$-smooth function. By taking the Taylor expansion of h_is about the point zero, we find

$$h_i = h\left(\frac{i\pi}{n+1}\right) = h_0 + \frac{\left(\frac{i\pi}{n+1}\right)^{k-1}}{(k-1)!} h^{(k-1)}(\xi_i), \quad \xi_i \in \left(0, \frac{i\pi}{n+1}\right). \tag{3.27}$$

Thus, the vector corresponding to $\bar{x}_S$ in the numerator is given by

$$\sum_{i\in S_n} h_i c_i \bar{v}_i = \sum_{i\in S_n} \left(h_0 + \frac{\left(\frac{i\pi}{n+1}\right)^{k-1}}{(k-1)!} h^{(k-1)}(\xi_i) \right) c_i \bar{v}_i = h_0 \bar{x}_S + \sum_{i\in S_n} \left(\frac{i\pi}{n+1}\right)^{k-1} \eta_i c_i \bar{v}_i,$$

where $\eta_i = \frac{h^{(k-1)}(\xi_i)}{(k-1)!}$, $i \in S_n$, bounded. The correction term of the numerator corresponding to Δ, is $\mathbf{Z} = \sum_{i=1}^n h_i c_i \bar{v}_i^T \Delta \sum_{i=1}^n h_i c_i \bar{v}_i$ which takes the form

$$\mathbf{Z} = \sum_{i\in S_n} h_i c_i \bar{v}_i^T \Delta \sum_{i\in S_n} h_i c_i \bar{v}_i + 2\sum_{i\in S_n} h_i c_i \bar{v}_i^T \Delta \sum_{i\in Q_n} h_i c_i \bar{v}_i + \sum_{i\in Q_n} h_i c_i \bar{v}_i^T \Delta \sum_{i\in Q_n} h_i c_i \bar{v}_i = \mathbf{Z}_1 + 2\mathbf{Z}_2 + \mathbf{Z}_3. \tag{3.28}$$

We have proven that the third term $\mathbf{Z}_3$ coincides with q_n. The first term gives

$$\begin{aligned}\mathbf{Z}_1 &= \left(h_0 \bar{x}_S^T + \sum_{i\in S_n} \left(\frac{i\pi}{n+1}\right)^{k-1} \eta_i c_i \bar{v}_i^T\right) \Delta \left(h_0 \bar{x}_S^T + \sum_{i\in S_n} \left(\frac{i\pi}{n+1}\right)^{k-1} \eta_i c_i \bar{v}_i\right)\\ &= h_0^2 \bar{x}_S^T \Delta \bar{x}_S + 2h_0 \bar{x}_S^T \Delta \sum_{i\in S_n} \left(\frac{i\pi}{n+1}\right)^{k-1} \eta_i c_i \bar{v}_i + \sum_{i\in S_n} \left(\frac{i\pi}{n+1}\right)^{k-1} \eta_i c_i \bar{v}_i^T \Delta \sum_{i\in S_n} \left(\frac{i\pi}{n+1}\right)^{k-1} \eta_i c_i \bar{v}_i,\end{aligned} \tag{3.29}$$

while the second one gives

$$\begin{aligned}\mathbf{Z}_2 &= \left(h_0 \bar{x}_S^T + \sum_{i\in S_n} \left(\frac{i\pi}{n+1}\right)^{k-1} \eta_i c_i \bar{v}_i^T\right) \Delta \sum_{i\in Q_n} h_i c_i \bar{v}_i\\ &= h_0 \bar{x}_S^T \Delta \sum_{i\in Q_n} h_i c_i \bar{v}_i + \sum_{i\in S_n} \left(\frac{i\pi}{n+1}\right)^{k-1} \eta_i c_i \bar{v}_i^T \Delta \sum_{i\in Q_n} h_i c_i \bar{v}_i.\end{aligned} \tag{3.30}$$

First we will estimate the quantity $\mathbf{q} = \| \sum_{i\in S_n} \left(\frac{i\pi}{n+1}\right)^{k-1} \eta_i c_i \bar{v}_i \|_2$. From $i \in S_n$ and the fact that $\bar{v}_i = \left(\sqrt{\frac{2}{n+1}} \sin\left(\frac{ij\pi}{n+1}\right)\right)_{j=1}^{k-1}$, we get that $\|\bar{v}_i\|_2 \sim \frac{i}{n^{\frac{3}{2}}}$. So,

$$\begin{aligned}\mathbf{q} &\le \sum_{i\in S_n} |\eta_i||c_i| \left(\frac{i\pi}{n+1}\right)^{k-1} \|\bar{v}_i\|_2 \sim \frac{\eta}{\sqrt{n}} \sum_{i\in S_n} |c_i| \left(\frac{i}{n}\right)^k\\ &\le \frac{\eta}{\sqrt{n}} \left(\sum_{i\in S_n} 1\right)^{\frac{1}{2}} \left(\sum_{i\in S_n} c_i^2 \left(\frac{i}{n}\right)^{2k}\right)^{\frac{1}{2}} \sim \sqrt{\frac{\#S_n}{n}} \left(\sum_{i\in S_n} c_i^2 g_i\right)^{\frac{1}{2}},\end{aligned} \tag{3.31}$$

where $\eta \in (\min_i |\eta_i|, \max_i |\eta_i|)$ and $\#S_n$ means the cardinality of the set S_n. Since $\frac{\#S_n}{n} = o(1)$, we get that the quantity $(\sum_{i\in S_n} c_i^2 g_i)^{\frac{1}{2}}$, which is just the square root of the τ-term, exceeds $\| \sum_{i\in S_n} \left(\frac{i\pi}{n+1}\right)^{k-1} \eta_i c_i \bar{v}_i \|_2$ in order of magnitude. Coming back to the terms $\mathbf{Z}_1$ and $\mathbf{Z}_2$ of the numerator, we deduce that the order of the first term of $\mathbf{Z}_1$ in (3.29) is

$$|h_0^2 \bar{x}_S^T \Delta \bar{x}_S| \le h_0^2 \|\bar{x}_S\|_2^2 \|\Delta\|_2 = \Omega(q_n),$$

which coincides with $\bar{x}_S^T \Delta \bar{x}_S$ of the denominator in (3.24). On the other hand, we can prove that $|\bar{x}_S^T \Delta \bar{x}_S| \sim \|\bar{x}_S\|_2^2$ by taking into account the proof of Lemma 2.6 of [18]. In that work, it was proved that

$$\bar{v}_i^T \Delta \bar{v}_j = \frac{2\sin^2(\theta)}{n+1} z_{ij}(\theta), \quad \theta = \frac{\pi}{n+1}, \quad i, j \in S_n$$

where

$$\lim_{\theta\to 0} z_{ij}(\theta) = ij \binom{2k-4}{k-2}.$$

Finally, we obtain that

$$\bar{x}_S^T \Delta \bar{x}_S = \frac{2\sin^2(\theta)}{n+1} \sum_{i\in S_n} \sum_{j\in S_n} c_i c_j z_{ij}(\theta) = \frac{2\sin^2(\theta)}{n+1} z(\theta),$$

where

$$\lim_{\theta\to 0} z(\theta) = \binom{2k-4}{k-2} \sum_{i\in S_n} \sum_{j\in S_n} i c_i j c_j = \binom{2k-4}{k-2} \left(\sum_{i\in S_n} i c_i\right)^2 \ge 0.$$

By applying the same considerations to the quantity $\|\bar{x}_S\|_2^2$, after a simple analysis, we have

$$\|\bar{x}_S\|_2^2 = \frac{2\sin^2(\theta)}{n+1} y(\theta),$$

where

$$\lim_{\theta\to 0} y(\theta) = \frac{(k-1)k(2k-1)}{6} \left(\sum_{i\in S_n} i c_i\right)^2 \ge 0.$$

From the relations above, we conclude that the quantities $\bar{x}_S^T \Delta \bar{x}_S$ and $\|\bar{x}_S\|_2^2$ have the same order of magnitude.

The order of the second term of $\mathbf{Z}_1$ in (3.29) is

$$\begin{aligned}\left|2h_0\bar{x}_S^T\Delta\sum_{i\in S_n}\left(\frac{i\pi}{n+1}\right)^{k-1}\eta_i c_i \bar{v}_i\right| &\leq 2h_0\|\bar{x}_S\|_2\|\Delta\|_2\left\|\sum_{i\in S_n}\left(\frac{i\pi}{n+1}\right)^{k-1}\eta_i c_i\bar{v}_i\right\|_2\\ &= \|\bar{x}_S\|_2\times o\left(\left(\sum_{i\in S_n}c_i^2 g_i\right)^{\frac{1}{2}}\right).\end{aligned}$$

This term is less than the first one, in order of magnitude, if $\sum_{i\in S_n} c_i^2 g_i = O\left(\|\bar{x}_S\|_2^2\right)$, while it is less than the corresponding τ-term, in order of magnitude if $\sum_{i\in S_n} c_i^2 g_i = \Omega\left(\|\bar{x}_S\|_2^2\right)$. In any case, it does not play a role in the order of magnitude of the numerator. We arrive at the same conclusion regarding the order of the third term of $\mathbf{Z}_1$ in (3.29) which is $o\left(\sum_{i\in S_n} c_i^2 g_i\right)$.

For the terms of $\mathbf{Z}_2$ in (3.30) we first estimate the term $\left\|\sum_{i\in Q_n} h_i c_i \bar{v}_i\right\|_2$:

$$\left\|\sum_{i\in Q_n} h_i c_i\bar{v}_i\right\|_2 \leq \sum_{i\in Q_n} h_i|c_i|\|\bar{v}_i\|_2 \leq \left(\sum_{i\in Q_n}c_i^2\right)^{\frac{1}{2}}\left(\sum_{i\in Q_n}h_i^2\|\bar{v}_i\|_2^2\right)^{\frac{1}{2}} \sim (q_n)^{\frac{1}{2}}.$$

Therefore, the order of the first term of $\mathbf{Z}_2$ in (3.30) is given by

$$\left|h_0\bar{x}_S^T\Delta\sum_{i\in Q_n}h_i c_i\bar{v}_i\right| \leq h_0\|\bar{x}_S\|_2\|\Delta\|_2\left\|\sum_{i\in Q_n}h_i c_i\bar{v}_i\right\|_2 = \|\bar{x}_S\|_2\times O\left((q_n)^{\frac{1}{2}}\right),$$

which is less, in order of magnitude, than $\bar{x}_S^T\Delta\bar{x}_S$ in the denominator of (3.24). The order of the second term of $\mathbf{Z}_2$ in (3.30) is given by

$$\begin{aligned}\left|\sum_{i\in S_n}\left(\frac{i\pi}{n+1}\right)^{k-1}\eta_i c_i\bar{v}_i\Delta\sum_{i\in Q_n}h_i c_i\bar{v}_i\right| &\leq \left\|\sum_{i\in S_n}\left(\frac{i\pi}{n+1}\right)^{k-1}\eta_i c_i\bar{v}_i\right\|_2\|\Delta\|_2\left\|\sum_{i\in Q_n}h_i c_i\bar{v}_i\right\|_2\\ &= o\left(\left(\sum_{i\in S_n}c_i^2 g_i\right)^{\frac{1}{2}}\right)\times O\left((q_n)^{\frac{1}{2}}\right),\end{aligned}$$

which is less, in order of magnitude, than the same term $\bar{x}_S^T\Delta\bar{x}_S$, if $\sum_{i\in S_n} c_i^2 g_i = O\left(\|\bar{x}_S\|_2^2\right)$, while it is less than the corresponding τ-term, in order of magnitude, if $\sum_{i\in S_n} c_i^2 g_i = \Omega\left(\|\bar{x}_S\|_2^2\right)$, since $\|\bar{x}_S\|_2 = \Omega\left((q_n)^{\frac{1}{2}}\right)$. □

Theorem 3.8. *Let $f \in \mathcal{C}_{2\pi}$ be an even function with roots $x_1, x_2, \ldots, x_l$ with multiplicities $2k_1, 2k_2, \ldots, 2k_l$, respectively, g be the trigonometric polynomial of order $k = \sum_{j=1}^{l} k_j$ given by (2.1), that rises the roots and w be the remaining positive part of f ($f = g\cdot w$). If the function $h = \sqrt{w}$ is a $(k_j - 1, x_j)$-smooth function for all $j = 1(1)l$, then the spectrum of the preconditioned matrix $K_n^\tau(f)^{-1}T_n(f)$ is bounded from above as well as from below:*

$$c < \lambda_{\min}(K_n^\tau(f)^{-1}T_n(f)) < \lambda_{\max}(K_n^\tau(f)^{-1}T_n(f)) < C,$$

where c and C are constants independent of the size n.

Proof. For the case of one zero at 0, Lemmata 3.3–3.5 and 3.7 cover all possible choices of the vector $x \in \mathbb{R}^n$ to obtain that the Rayleigh quotient r_x is bounded. The case of one zero at a point different from 0 is simple, since it can be transformed to zero by a shift transformation. The generalization to more roots is straightforward. The main difference concerns the definition of the sets S_n and Q_n of (3.19). Under the assumption of l roots $x_1, x_2, \ldots, x_l$, we give the new definition of the above sets as

$$\begin{aligned}&(1)\ S_n \subset N_n \forall n\\ &(2)\ \forall i_n \text{ sequence to which } i_k \in S_k \text{ we have } \lim_{n\to\infty}\frac{i_n}{n} - x_j = 0 \qquad (3.32)\\ &\quad (i_n - nx_j = o(n)),\quad j = 1, 2, \ldots, l.\end{aligned}$$

and

$$Q_n = N_n \setminus S_n. \qquad (3.33)$$

After that definition, Lemmata 3.3–3.5 and 3.7 work well to yield our result that r_x is bounded, which completes the proof of the Theorem. □

As a subsequent result we have that the minimum eigenvalue of $K_n^\tau(f)^{-1}T_n(f)$ is bounded far away from zero. Hence, from the theorem of Axelsson and Lindskog [1], it follows immediately that the PCG method will have superlinear convergence.

We have to remark here that if the smoothing condition of the function h does not hold, the Rayleigh quotient r_x may not be bounded and consequently the PCG method may not have superlinear convergence. The worst case, where we get the maximum value of r_x, occurs when choosing $x = x_S$. In that case the denominator coincides with $\frac{1}{n^{2k}}$, and so for the numerator to be of the same order the $(k-1,0)$-smoothness of the function h is necessary. Otherwise, if h is a $(k-2,0)$-smooth function, which is the best possible choice, we deduce that the numerator coincides with $\frac{1}{n^{2k-1}}$. As a consequence, r_x tends to infinity with a rate coinciding with n.

3.2. Convergence of the method: Circulant case

For circulant matrices, in order to show the clustering of the eigenvalues of the preconditioned matrix sequence

$$(C_n(h)T_n(g)C_n(h))^{-1}T_n(f) \tag{3.34}$$

around unity, we first remark that although a band Toeplitz matrix and a circulant one do not commute, they very nearly have the commutativity property, since

$$\operatorname{rank}(T_n(g)\cdot C - C\cdot T_n(g)) \le 2k,$$

where k is the bandwidth of the band matrix and which is obviously independent of the dimension n of the problem. We will show that the main mass of the eigenvalues of the preconditioned matrix (3.34) is clustered around unity. Before giving the main results for this case, we report a useful lemma.

Lemma 3.9. *Let $w \in \mathcal{C}_{2\pi}$ be a positive and even function. Then, for any positive ϵ, there exist N and $M > 0$ such that for every $n > N$, at most M eigenvalues of the matrix $C_n^{-1}T_n(w)$ have absolute value greater than ϵ.*

Proof. See [23], Theorem 2.1 (The proof for the circulant case is just the same as the one for τ case). □

Theorem 3.10. *Let $T_n(f)$ be the Toeplitz matrix produced by a nonnegative function f in $\mathcal{C}_{2\pi}$, which can be written as $f = g\cdot w$, where g is the even trigonometric polynomial as is defined in (2.1) and $w = h^2$ is a strictly positive even function belonging to $\mathcal{C}_{2\pi}$. Then for every $\epsilon > 0$, there exist N and $\hat{M} > 0$ such that for every $n > N$, at most $\hat{M}$ eigenvalues of the preconditioned matrix (3.34) lie outside the interval $(1-\epsilon, 1+\epsilon)$.*

Proof. We follow exactly the same steps and the same considerations as in the proof of Theorem 3.2 for the τ case, with the only difference being that the matrices $C_n(g)$ and $C_n(h)$ replace $\tau_n(g)$ and $\tau_n(h)$, respectively. First, we obtain that

$$\hat{T}_n = \frac{1}{2}T_n(g)^{\frac{1}{2}}C_n(h)^{-1}T_n(w)C_n(h)^{-1}T_n(g)^{-\frac{1}{2}} + \frac{1}{2}T_n(g)^{-\frac{1}{2}}C_n(h)^{-1}T_n(w)C_n(h)^{-1}T_n(g)^{\frac{1}{2}} + L_5, \tag{3.35}$$

with L_5 being symmetric and a low rank matrix (of constant rank). It is noted that we have used the same notation $\hat{T}_n$ for the associated symmetric form of the preconditioned matrix.

From Lemma 3.9, we obtain that for the choice of $\epsilon_h > 0$ there exist a low rank (of constant rank) matrix L_6 and a matrix E of small norm ($\|E\|_2 \le \epsilon_h$), such that

$$C_n(h)^{-1}T_n(w)C_n(h)^{-1} = I + E + L_6. \tag{3.36}$$

Consequently, we obtain the relation

$$\hat{T}_n = I + \frac{1}{2}T_n(g)^{\frac{1}{2}}ET_n(g)^{-\frac{1}{2}} + \frac{1}{2}T_n(g)^{-\frac{1}{2}}ET_n(g)^{\frac{1}{2}} + L,$$

which is nothing but relation (3.3) for the τ case.

After the latter manipulations, the proof follows step by step the one given in Theorem 3.2; the same result is obtained. □

As in the case of τ matrices, we will prove the important feature that our preconditioner satisfies and leads to superlinear convergence of the PCG.

The clustering of the eigenvalues around 1 has been proven in Theorem 3.10. We have to prove now that there does not exist any eigenvalue, belonging to the outliers, that tends to zero or to infinity. For this, we will study the Rayleigh quotients of the preconditioned matrix, as in the τ case. It is easily proved that the previous analysis, from relation (3.5) to relation (3.7), for the τ case, holds also for the circulant case by simply replacing $\tau_n(h)$ by $C_n(h)$.

Therefore, we have to prove that

$$\limsup_{n\to\infty}\sup_{x\in\mathbb{R}^n}\frac{x^TC_n(h)T_n(g)C_n(h)x}{x^TT_n(g)x} < \infty. \tag{3.37}$$

For this, we have to study the ratio

$$r_x = \frac{x^TC_n(h)T_n(g)C_n(h)x}{x^TT_n(g)x}. \tag{3.38}$$

It is well known that the band Toeplitz matrix $T_n(g)$ is written as a circulant minus a low rank Toeplitz matrix

$$T_n(g) = C_n(g) - \tilde{T}_n(g), \tag{3.39}$$

where $\tilde{T}_n(g)$ is a Toeplitz matrix of rank $2k$ of the form

$$\tilde{T}_n(g) = \tilde{J}_n(g) + \tilde{J}_n(g)^T = \text{Toep}(0, \dots, 0, g_k, g_{k-1}, \dots, g_1), \tag{3.40}$$

where the entries g_i are the Fourier coefficients of the trigonometric polynomial g ($g(x) = g_0 + 2g_1 \cos(x) + 2g_2 \cos(2x) + \cdots + 2g_k \cos(kx)$). It is obvious that $\tilde{T}_n(g)$ is an indefinite matrix, while C_n is a semi positive definite one. We define by $\varDelta$ the $k \times k$ matrix formed by the first k rows and the last k columns of $\tilde{J}_n(g)$:

$$\varDelta = \begin{pmatrix} g_k & \cdots & g_2 & g_1 \\ 0 & \ddots & & g_2 \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & 0 & g_k \end{pmatrix}. \tag{3.41}$$

We use the same notations $\overline{x}^{(m)}$ and $\underline{x}^{(m)}$ for the first and the last m-dimensional blocks of the vector x, respectively.

Recalling ratio (3.38), we find

$$\begin{aligned} r_x &= \frac{x^T C_n(h) T_n(g) C_n(h) x}{x^T T_n(g) x} = \frac{x^T C_n(h) C_n(g) C_n(h) x - x^T C_n(h) \tilde{T}_n(g) C_n(h) x}{x^T C_n(g) x - x^T \tilde{T}_n(g) x} \\ &= \frac{x^T C_n(h^2 g) x - x^T C_n(h) \tilde{T}_n(g) C_n(h) x}{x^T C_n(g) x - \overline{x}^{(k)^T} \varDelta \underline{x}^{(k)} - \underline{x}^{(k)^T} \varDelta^T \overline{x}^{(k)}} = \frac{x^T C_n(h^2 g) x - x^T C_n(h) \tilde{T}_n(g) C_n(h) x}{x^T C_n(g) x - 2\overline{x}^{(k)^T} \varDelta \underline{x}^{(k)}}. \end{aligned} \tag{3.42}$$

We state here, without proof, a sequence of lemmata analogous to Lemmata 3.3–3.5 and 3.7 and finally, a theorem analogous to Theorem 3.8. The proofs are given in an similar way as the proofs in the τ case, although specific difficulties appear in some points. These proofs can be found in the Technical Report [20].

Lemma 3.11. *Let x be a normalized n-dimensional vector ($\|x\|_2 = 1$) and suppose the sequence of the vectors $\overline{x}^{(k)}$ is bounded, i.e. $0 < c \le \|\overline{x}^{(k)}\|_2 \le 1$ for all n or the sequence of the vectors $\underline{x}^{(k)}$ is bounded i.e. $0 < c \le \|\underline{x}^{(k)}\|_2 \le 1$ for all n, with c being constant independent of n; then the ratio r_x in (3.42) is bounded.*

Lemma 3.12. *Let x be such that $\|\overline{x}^{(k)}\|_2 = o(1)$ and $\|\underline{x}^{(k)}\|_2 = o(1)$ and the sequence $\{q_n\}_n$ is bounded, i.e. $0 < c \le q_n \le 1$; then the ratio r_x is bounded.*

Lemma 3.13. *Let x be such that $\|\overline{x}^{(k)}\|_2 = o(1)$ and $\|\underline{x}^{(k)}\|_2 = o(1)$ and for the sequences $\{s_n\}_n$ and $\{q_n\}_n$ it holds that $\lim_{n\to\infty} s_n = 1$, $\lim_{n\to\infty} q_n = 0$ with $\|\overline{x}_S\|_2 = o\left((q_n)^{\frac{1}{2}}\right)$ and $\|\underline{x}_S\|_2 = o\left((q_n)^{\frac{1}{2}}\right)$; then the ratio r_x is bounded.*

Lemma 3.14. *Let x be such that $\|\overline{x}^{(k)}\|_2 = o(1)$ and $\|\underline{x}^{(k)}\|_2 = o(1)$ and for the sequences $\{s_n\}_n$ and $\{q_n\}_n$ it holds that $\lim_{n\to\infty} s_n = 1$, $\lim_{n\to\infty} q_n = 0$ with $\|\overline{x}_S\|_2 = \varOmega\left((q_n)^{\frac{1}{2}}\right)$ or $\|\underline{x}_S\|_2 = \varOmega\left((q_n)^{\frac{1}{2}}\right)$. Suppose also that h is a $(k, 0)$-smooth function. Then, the ratio r_x is bounded.*

Theorem 3.15. *Let $f \in \mathcal{C}_{2\pi}$ be an even function with roots $x_0, x_1, \dots, x_l$ with multiplicities $2k_1, 2k_2, \dots, 2k_l$, respectively, and let g be the trigonometric polynomial of order $k = \sum_{j=1}^{l} k_j$ given by (2.1), that rises the roots and w the remaining positive part of f ($f = g \cdot w$). If the function $h = \sqrt{w}$ is a (k_j, x_j)-smooth function for all x_js, $j = 1(1)l$, then the spectrum of the preconditioned matrix $K_n^C(f)^{-1} T_n(f)$ is bounded from above as well as from below:*

$$c < \lambda_{\min}(K_n^C(f)^{-1} T_n(f)) < \lambda_{\max}(K_n^C(f)^{-1} T_n(f)) < C, \tag{3.43}$$

where c and C are constants independent of the size n.

As a subsequent result, we have that the minimum eigenvalue of $K_n^C(f)^{-1} T_n(f)$ is bounded far away from zero. Hence, from the theorem of Axelsson and Lindskog [1], it follows immediate that the PCG method will have superlinear convergence.

We have to remark here that one order of smoothness more is required for the circulant case than the one required for the τ case. If the smoothing condition of the function h does not hold, the Rayleigh quotient r_x may not be bounded, and consequently the PCG method may not have superlinear convergence.

Remark 3.1. Following a theory closely related to that just developed, band plus Hartley preconditioners could be applied for the solution of ill-conditioned Hermitian Toeplitz systems. In this paper, we do not study this case. We simply remark that a similar analysis could be applied to obtain analogous results for the superlinearity of the convergence. Since Hartley matrices are closely related to circulant matrices, we believe that $(k, 0)$-smoothing, for the function h, is needed.

4. Smoothing technique

Our analysis brings up the following question: Is the condition of smoothing valid for most of the applications? The answer to this question is not positive. There are problems where the positive part h is smooth enough, but in most of them we are not guaranteed the required smoothness. In some of the problems the function h is not differentiable at 0, nor continuous. In the following two subsections, we propose a smoothing technique which approximates h with a $(k-1, 0)$-smooth function for the τ case and with a $(k, 0)$-smooth function for the Circulant case, respectively, in order to get superlinear convergence.

4.1. Smoothing technique: τ case

Let us assume that the factor h of the generating function f is not a $(k-1, 0)$-smooth function. We define the function $\hat{h}$ as follows

$$\hat{h}(x) = \begin{cases} P_k[h](x) & \text{if } x \in (-\epsilon, \epsilon) \\ h(x) & \text{if } x \in [-\pi, -\epsilon] \cup [\epsilon, \pi], \end{cases} \tag{4.1}$$

where ϵ is a small positive constant and $P_k[h]$ is an even and $(k-1, 0)$-smooth function which interpolates h at the points $-\epsilon, 0, \epsilon$. It is obvious that we can choose as $P_k[h]$ the function

$$P_k[h](x) = \frac{h(\epsilon) - h_0}{\epsilon^k} |x|^k + h_0, \tag{4.2}$$

which is a k degree interpolation polynomial on the interval $(0, \epsilon)$, or the function

$$P_k[h](x) = \frac{h(\epsilon) - h_0}{(2 - 2\cos(\epsilon))^{\frac{k}{2}}} (2 - 2\cos(x))^{\frac{k}{2}} + h_0, \tag{4.3}$$

which, for even k, is a k degree interpolation trigonometric polynomial on the interval $(-\epsilon, \epsilon)$. For small ϵ, the function $P_k[h]$ is a very good approximation of h on the interval $(-\epsilon, \epsilon)$. For this reason we propose as preconditioner the matrix

$$K_n^{\tau}(\hat{f}) = \tau_n(\hat{h}) T_n(g) \tau_n(\hat{h}). \tag{4.4}$$

The smoothing identity of the function $\hat{f} = g \cdot \hat{h}^2$ is valid and Theorem 3.8 guarantees superlinear convergence of the PCG method with preconditioned matrix sequence $K_n^{\tau}(\hat{f})^{-1} T_n(f)$. We state here the generalization of Theorem 3.8.

Theorem 4.1. *Let $f \in \mathcal{C}_{2\pi}$ be an even function with roots $x_0, x_1, \ldots, x_l$ with multiplicities $2k_1, 2k_2, \ldots, 2k_l$, respectively, g be the trigonometric polynomial of order $k = \sum_{j=1}^{l} k_j$ given by (2.1), that raises the roots, w the remaining positive part of f ($f = g \cdot w$) and $h = \sqrt{w}$. We define the function $\hat{h}$ as follows:*

$$\hat{h}(x) = \begin{cases} P_{k_j}[h](x) & \text{if } x \in (x_j - \epsilon_j, x_j + \epsilon_j), \quad j = 1, 2, \ldots, l \text{ and} \\ & h \text{ is not a } (k_j - 1, x_j)\text{-smooth function} \\ h(x) & \text{elsewhere}, \end{cases} \tag{4.5}$$

where $\epsilon_j, j = 1, 2, \ldots, l$ are small positive constants and

$$P_{k_j}[h](x) = \frac{(x - x_j + \epsilon_j)h(x_j + \epsilon_j) - (x - x_j - \epsilon_j)h(x_j - \epsilon_j) - 2\epsilon_j h(x_j)}{2\epsilon_j^{k+1}} |x - x_j|^k + h(x_j) \quad \text{or}$$

$$P_{k_j}[h](x) = \frac{(2 - 2\cos(x - x_j + \epsilon_j))h(x_j + \epsilon_j) + (2 - 2\cos(x - x_j - \epsilon_j))h(x_j - \epsilon_j) - (2 - 2\cos(2\epsilon_j))h(x_j)}{(2 - 2\cos(2\epsilon_j))(2 - 2\cos(\epsilon_j))^{\frac{k}{2}}} \times (2 - 2\cos(x - x_j))^{\frac{k}{2}} + h(x_j).$$

Then, the spectrum of the preconditioned matrix $K_n^{\tau}(\hat{f})^{-1} T_n(f)$ ($\hat{f} = g \cdot \hat{h}^2$) is bounded from above as well as from below:

$$c < \lambda_{\min}(K_n^{\tau}(\hat{f})^{-1} T_n(f)) < \lambda_{\max}(K_n^{\tau}(\hat{f})^{-1} T_n(f)) < C,$$

where c and C are constants independent of the size n.

$P_{k_j}[h]$ have been taken to be interpolation functions of h at the points $x_j - \epsilon_j, x_j, x_j + \epsilon_j$.

4.2. Smoothing technique: Circulant case

The same smoothing technique could be applied if h is not a $(k, 0)$-smooth function. We state here the generalization of Theorem 3.15.

Table 5.1
Number of iterations for $f_1(x) = x^4$

n	R	S^{*3}	$M^{1,2}$	W	τ	$\mathcal{C}$
32	15	11	6	7	5	6
64	20	11	8	8	5	6
128	24	12	10	8	6	6
256	27	12	11	9	7	7
512	29	13	11	9	7	7
1024	30	13	12	9	7	7

Theorem 4.2. *Let $f \in \mathcal{C}_{2\pi}$ be an even function with roots $x_0, x_1, \ldots, x_l$ with multiplicities $2k_1, 2k_2, \ldots, 2k_l$, respectively, g the trigonometric polynomial of order $k = \sum_{j=1}^{l} k_j$ given by (2.1), that raises the roots, w the remaining positive part of f ($f = g \cdot w$) and $h = \sqrt{w}$. We define the function $\hat{h}$ as follows:*

$$\hat{h}(x) = \begin{cases} P_{k_j}[h](x) & \text{if } x \in (x_j - \epsilon_j, x_j + \epsilon_j), \quad j = 1, 2, \ldots, l \text{ and} \\ & h \text{ is not a } (k_j, x_j)\text{-smooth function} \\ h(x) & \text{elsewhere}, \end{cases} \tag{4.6}$$

where $\epsilon_j, j = 1, 2, \ldots, l$ are small positive constants and

$$P_{k_j}[h](x) = \frac{(x - x_j + \epsilon_j)h(x_j + \epsilon_j) - (x - x_j - \epsilon_j)h(x_j - \epsilon_j) - 2\epsilon_j h(x_j)}{2\epsilon_j^{k+2}} |x - x_j|^{k+1} + h(x_j) \quad \text{or}$$

$$\begin{aligned} P_{k_j}[h](x) \;=\; & \frac{(2 - 2\cos(x - x_j + \epsilon_j))h(x_j + \epsilon_j) + (2 - 2\cos(x - x_j + \epsilon_j))h(x_j - \epsilon_j) - (2 - 2\cos(2\epsilon_j))h(x_j)}{(2 - 2\cos(2\epsilon_j))(2 - 2\cos(\epsilon_j))^{\frac{k+1}{2}}} \\ & \times (2 - 2\cos(x - x_j))^{\frac{k+1}{2}} + h(x_j). \end{aligned}$$

Then, the spectrum of the preconditioned matrix $K_n^C(\hat{f})^{-1}T_n(f)$ ($\hat{f} = g \cdot \hat{h}^2$) is bounded from above as well as from below:

$$c < \lambda_{\min}(K_n^C(\hat{f})^{-1}T_n(f)) < \lambda_{\max}(K_n^C(\hat{f})^{-1}T_n(f)) < C, \tag{4.7}$$

where c and C are constants independent of the size n.

Remark 4.1. The same smoothing technique could be applied for the band plus Hartley preconditioners, when the function h is not a $(k, 0)$-smooth function.

5. Numerical experiments

In this section, we report some numerical examples to show the efficiency of the proposed preconditioners and to confirm the validity of the presented theory. The experiments were carried out using Matlab. In all the examples, the right-hand side of the system was $(11 \cdots 1)^T$ in order to compare our method with methods proposed by other researchers. We have run also our examples with the right-hand side being random vectors and we have obtained results with the same behavior. The zero vector was as our initial guess for the PCG method and as stopping criterion was taken the validity of the inequality $\frac{\|r^{(k)}\|_2}{\|r^{(0)}\|_2} \leq 10^{-7}$, where $r^{(k)}$ is the residual vector in the kth iteration.

Example 5.1. We consider the function $f_1(x) = x^4$ as generating function. The associated function $h = \frac{x^2}{2-2\cos(x)}$ is a $(2, 0)$-smooth function and so, a smoothing technique is not needed for both band plus τ and band plus circulant preconditioners. In Table 5.1 the number of iterations needed to achieve the predefined accuracy are illustrated. We compare the performance of our preconditioners with a variety of other well known and optimal preconditioners: R is the pioneering one proposed by Chan [8]. S^{*3} is the proposal of Serra Capizzano in [22] using best Chebyshev approximation (3 is the degree of the polynomial). $M^{(1,2)}$ is the preconditioner proposed by Noutsos and Vassalos in [19], which is based on best rational approximation with 1, 2 being the degrees of the numerator and denominator, respectively. W is the ω circulant preconditioner proposed by Potts and Steidl in [21]. Finally, by τ and $\mathcal{C}$, we denote the band plus τ and band plus circulant preconditioners, respectively. The efficiency of our preconditioners is clearly shown.

Example 5.2. Let

$$f_2(x) = \begin{cases} x^2(|x| + 1) & |x| \leq \dfrac{\pi}{2} \\ \left(\dfrac{\pi}{2} + 2\right)x^2 & x \in [-\pi, \pi] \setminus \left[-\dfrac{\pi}{2}, \dfrac{\pi}{2}\right] \end{cases}$$

Table 5.2
$f_2(x)$

n	$\lambda_{\max}\tau$	$\lambda_{\min}\tau$	τ	$\lambda_{\max}\mathcal{C}$	$\lambda_{\min}\mathcal{C}$	$\mathcal{C}$	B
32	1.7612	0.9003	6	4.2123	0.7960	9	8
64	1.7694	0.8925	7	4.2465	0.8027	10	24
128	1.7736	0.8869	7	4.2648	0.8070	10	27
256	1.7758	0.8825	7	4.2742	0.8098	11	29
512	1.7771	0.8791	7	4.2791	0.8116	12	30
1024	1.7778	0.8764	7	4.2815	0.8127	12	31

Table 5.3
$f_3(x)\tau$ without smoothing

n	$\lambda_{\max}\tau$	$\lambda_{\min}\tau$	τ	B
32	5.5929	0.843	8	17
64	6.049	0.835	10	34
128	6.3624	0.8291	11	45
256	6.5669	0.8249	11	54
512	6.6955	0.8221	11	61
1024	6.7744	0.8205	12	67

Table 5.4
$f_3(x)$ circulant and smoothing circulant in $[-.5, .5]$

n	$\lambda_{\max}\mathcal{C}$	$\lambda_{\min}\mathcal{C}$	$\mathcal{C}$	$\lambda_{\max}\hat{\mathcal{C}}$	$\lambda_{\min}\hat{\mathcal{C}}$	$\hat{\mathcal{C}}$	B
32	49.417	0.2286	13	32.369	0.34827	13	17
64	83.835	0.1386	15	34.260	0.34001	14	34
128	146.42	0.0789	18	35.552	0.3328	15	45
256	263.63	0.0428	23	36.218	0.3292	17	54
512	488.33	0.0224	26	36.556	0.3273	18	61
1024	926.19	0.0115	29	36.725	0.3265	18	67

be the generating function. Even though the corresponding function $h = \sqrt{\frac{f_2(x)}{2-2\cos(x)}}$ is not differentiable at the point $\frac{\pi}{2}$, it is an $(1, 0)$-smooth function. Hence, our preconditioners ensure superlinear convergence without any smoothing technique. In Table 5.2, we give the minimum and the maximum eigenvalues of the preconditioned matrix and the iterations of the PCG method needed for both τ and circulant cases. In the last column, denoted by B, we give for comparison the iterations needed if we use the band Toeplitz preconditioner generated by the trigonometric polynomial which raises the roots.

Example 5.3. For the generated function

$$f_3(x) = \begin{cases} x^4(|x|+1) & |x| \le \dfrac{\pi}{2} \\ \left(\dfrac{\pi}{2}+2\right)x^4 & x \in [-\pi, \pi] \setminus \left[-\dfrac{\pi}{2}, \dfrac{\pi}{2}\right] \end{cases}$$

we have that $k = 2$. It is easily checked that the corresponding function $h(x) = \frac{\sqrt{f_3(x)}}{2-2\cos(x)}$, is a $(1, 0)$-smooth function. Consequently, the τ plus band preconditioner works well without any smoothing technique, while the circulant plus band one needs a further smoothing step. In Table 5.3 we give the corresponding results, as in Table 5.2 for the τ case without smoothing, while in Table 5.4 we give the results for the circulant case without and with the smoothing technique. The band plus circulant preconditioner is denoted by $\hat{\mathcal{C}}$. It is easily seen that the smoothing technique is required for the circulant case to achieve superlinearity.

Example 5.4. Finally, we consider the function

$$f_4(x) = \begin{cases} x^6(|x|+1) & |x| \le \dfrac{\pi}{2} \\ \left(\dfrac{\pi}{2}+2\right)x^6 & x \in [-\pi, \pi] \setminus \left[-\dfrac{\pi}{2}, \dfrac{\pi}{2}\right] \end{cases}$$

as our generating function. In this example, we have $k = 3$ and moreover the corresponding function $h(x) = \sqrt{\frac{f_4(x)}{(2-2\cos(x))^3}}$ is also a $(1, 0)$-smooth function. Thus, the smoothing technique is necessary for both cases to achieve superlinearity. In Table 5.5 we give the iterations of the PCG method needed for both τ and circulant cases with and without using our smoothing technique. The meaning of the asterisks is that the iterations required are over 100. The presented numerical results fully confirm the theory developed in the previous sections.

Table 5.5
$f_4(x)\tau$ and τ with smoothing in $[-.5, .5]$

n	τ	$\hat{\tau}$	$\mathcal{C}$	$\hat{\mathcal{C}}$	B
32	14	10	17	13	20
64	20	11	25	16	48
128	33	13	43	19	*
256	53	14	79	21	*
512	*	15	*	22	*
1024	*	15	*	23	*

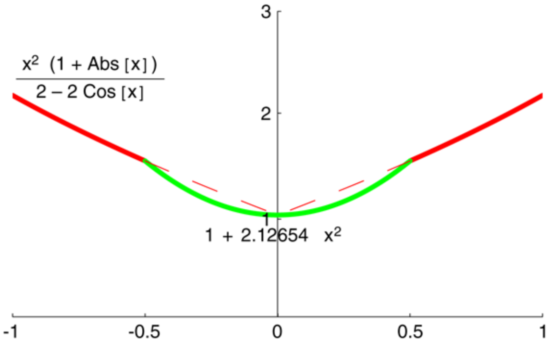

Fig. 5.1. Smoothing of $h(x) = \frac{x^2(1+|x|)}{2-2\cos(x)}$, by interpolation.

In Fig. 5.1, the smoothing technique is shown graphically for the function $h(x) = \frac{x^2(1+|x|)}{2-2\cos(x)}$. We have to remark that h is not a differentiable function at zero.

Acknowledgement

The authors would like to thank professor S. Serra Capizzano for the fruitful discussion during his visit in Greece.

References

[1] O. Axelsson, G. Lindskog, On the rate of convergence of the preconditioned conjugate gradient method, Numer. Math. 48 (1986) 499–523.
[2] D. Bini, F. Di Benedetto, A new preconditioner for the parallel solution of positive definite Toeplitz systems, in: Proc. 2nd ACM SPAA, Crete, Greece, 1990, pp. 220–223.
[3] D. Bini, P. Favati, On a matrix algebra related to the discrete Hartley transform, SIAM J. Matrix Anal. Appl. 14 (1993) 500–507.
[4] D. Bini, B. Meini, Effective methods for solving banded Toeplitz systems, SIAM J. Matrix Anal. Appl. 20 (1999) 700–719.
[5] A. Bottcher, B. Silbermann, Introduction to Large Truncated Toeplitz Matrices, 1st ed., Springer-Verlag, 2001.
[6] J. Bunch, Stability of methods for solving Toeplitz systems of equations, SIAM J. Sci. Statist. Comput. 6 (1985) 349–364.
[7] R.H Chan, Circulant preconditioners for Hermitian Toeplitz matrices, SIAM J. Matrix Anal. Appl. 10 (1989) 542–550.
[8] R.H Chan, Toeplitz preconditioners for Toeplitz systems with nonnegative generating functions, IMA J. Numer. Anal. 11 (1991) 333–345.
[9] R.H Chan, W.K Ching, Toeplitz-circulant preconditioners for Toeplitz systems and their applications on queueing networks with batch arrivals, SIAM J. Sci. Comput. 17 (1996) 762–772.
[10] R.H Chan, P. Tang, Fast band-Toeplitz preconditioners for Hermitian Toeplitz systems, SIAM J. Sci. Comput. 15 (1994) 164–171.
[11] R.H Chan, M. Yeung, Circulant preconditioners from kernels, SIAM J. Numer. Anal. 29 (1992).
[12] F. Di Benedetto, Analysis of preconditioning techniques for ill-conditioned Toeplitz matrices, SIAM J. Sci. Comput. 16 (1995) 682–697.
[13] F. Di Benedetto, Preconditioning of block Toeplitz matrices by sine transform, SIAM J. Sci. Comput. 18 (1997) 499–515.
[14] F. Di Benedetto, G. Fiorentino, S. Serra Capizzano, C.G. preconditioning of Toeplitz matrices, Comput. Math. Appl. 25 (1993) 35–45.
[15] G. Fiorentino, S. Serra Capizzano, Multigrid methods for symmetric positive definite block Toeplitz matrices with nonnegative generating functions, SIAM J. Sci. Comput. 17 (1996) 1068–1081.
[16] X.Q. Jin, Hartley preconditioners for Toeplitz systems generated by positive continuous functions, BIT 34 (1994) 367–371.
[17] D. Noutsos, S. Serra Capizzano, P. Vassalos, Spectral equivalence and matrix algebra preconditioners for multilevel Toeplitz systems: A negative result, Structured Matrices in Mathematics, Computer Science, and Engineering, Comtemporary Math. 323 (2003) 313–322.
[18] D. Noutsos, S. Serra Capizzano, P. Vassalos, Matrix algebra preconditioners for multilevel Toeplitz systems do not insure optimal convergence rate, Theoret. Computer Sci. 315 (2004) 557–579.
[19] D. Noutsos, P. Vassalos, New band Toeplitz preconditioners for ill-conditioned symmetric positive definite Toeplitz systems, SIAM J. Matrix Anal. Appl. 23 (2002) 728–743.
[20] D. Noutsos, P. Vassalos, Superlinear convergence for PCG using band plus algebra preconditioners for Toeplitz systems, Technical Report, Department of Mathematics, University of Ioannina, GR-45110 Ioannina, Greece, 2007, pp. 91–120.
[21] D. Potts, G. Steidl, Preconditioners for ill-conditioned Toeplitz matrices, BIT 39 (1999) 513–533.
[22] S. Serra Capizzano, Optimal, quasi-optimal and superlinear band-Toeplitz preconditioners for asymptotically ill-conditioned positive definite Toeplitz systems, Math. Comput. 66 (1997) 651–665.
[23] S. Serra Capizzano, Superlinear PCG methods for symmetric Toeplitz systems, Math. Comput. 68 (1999) 793–803.
[24] G. Strang, A proposal for Toeplitz matrix calculations, Stud. Appl. Math. 74 (1986) 171–176.