# Approximate Inverse-Free Preconditioners for Toeplitz Matrices

You-Wei Wen* Wai-Ki Ching † Michael K. Ng ‡

**Abstract**

In this paper, we propose approximate inverse-free preconditioners for solving Toeplitz systems. The preconditioners are constructed based on the famous Gohberg-Sememcul formula. We show that if a Toepltiz matrix is generated by a positive bounded function and its entries enjoys the off-diagonal decay property, then the eigenvalues of the preconditioned matrix are clustered around one. Experimental results show that the proposed preconditioners are superior to other existing preconditioners in the literature.

**Keywords:** Toeplitz matrices, Gohberg-Semencul formula, approximate inverse-free, preconditioned conjugate gradient method, preconditioners.

## 1 Introduction

In this paper, we are interested in solving the Toeplitz system

$$\boldsymbol{T}_n\boldsymbol{x} = \boldsymbol{b} \tag{1}$$

by the Preconditioned Conjugate Gradient (PCG) method. A matrix is called a Toeplitz matrix if it has constant diagonals, i.e., it takes the following form:

$$\boldsymbol{T}_n = \begin{pmatrix} t_0 & t_{-1} & t_{-2} & \cdots & t_{2-n} & t_{2-n} \\ t_1 & t_0 & t_{-1} & \cdots & \cdots & t_{1-n} \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ t_{n-2} & \cdots & \cdots & \ddots & \ddots & t_{-1} \\ t_{n-1} & t_{n-2} & \cdots & \cdots & t_1 & t_0 \end{pmatrix}. \tag{2}$$

*Faculty of Science, South China Agricultural University, Wushan, Guangzhou, P.R.China. E-mail: wenyouwei@gmail.com.

†Advanced Modeling and Applied Computing Laboratory, Department of Mathematics, The University of Hong Kong, Pokfulam Road, Hong Kong. E-mail: wching@hkusua.hku.hk Research supported in part by RGC Grants, HKU CRCG Grants, Hung Hing Ying Physical Sciences Research Fund and HKU Strategic Research Theme Fund on Computational Physics and Numerical Methods.

‡Department of Mathematics, Hong Kong Baptist University, Kowloon Tong, Hong Kong. E-mail: mng@math.hkbu.edu.hk. Research supported in part by RGC 7045/04P and 7045/05P, and HKBU FRGs.

Very often, a sequence of $n \times n$ Toeplitz matrices is generated by a function $f(\theta) \in C_{2\pi}$, called the generating function, in Wiener class (this means it has a Fourier series expansion with absolutely summable Fourier coefficients) and is denoted by $\boldsymbol{T}_n = \mathcal{T}_n[f]$. When the generate function $f(\theta)$ of a sequence of Toeplitz matrices $\boldsymbol{T}_n$ is known, its $(j,k)$-th entry is given by the $(j-k)$th Fourier coefficient of $f(\theta)$, i.e.,

$$t_k = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(\theta)e^{-ik\theta}d\theta, \quad k = 0, \pm 1, \pm 2, \ldots.$$

Toeplitz and block-Toeplitz matrices arise from many applications in applied sciences and engineering sciences, see for example, Ching [11] and Chan and Ng [6] and the references therein. There are two main types of methods for solving Toeplitz systems. The first type is direct methods. The direct methods are based on the idea of solving Toeplitz systems recursively, see for instance, Levinson (1946) [22], Durbin (1960) [14] and Trench (1964) [31]. The operational cost of these direct methods is $O(n^2)$. Superfast algorithms of complexity $O(n \log^2 n)$ for solving Toeplitz systems have been proposed by different groups of researchers, see for instance, Bitmead and Anderson (1980) [3], Brent, Gustavson and Yun (1980) [4], Morf (1980) [24], de Hoog (1987) [13], Ammar and Gragg (1988) [1] and Huckle (1998) [19]. The second type of method is iterative methods. Conjugate Gradient (CG) method is a popular method for solving Toeplitz systems. An important property of a Toeplitz matrix is that it can be embedded into $2n \times 2n$ circulant matrix. Thus the operational cost for a Toeplitz matrix-vector multiplication is $\mathrm{O}(n \log n)$ by using the Fast Fourier Transforms (FFT). In addition, when a suitable preconditioner is chosen, the convergence rate of the method can be speeded up. In the past two decades, there has been an intensive research with regard to preconditioning techniques for Toeplitz systems, see for instance [5, 8, 26, 27, 28, 29]. Here we propose inverse-free preconditioners for solving Toeplitz systems. The preconditioners are constructed based on Gohberg-Sememcul formula.

The Gohberg-Semencul formula states that the inverse of a Toeplitz matrix $T_n$ can be written a sum of multiplications of lower-triangular and upper-triangular Toeplitz matrices. In this paper, we will first present an interesting property that when such lower-triangular and upper-triangular Toeplitz matrices pre-multiplies and post-multiplies the original Toeplitz matrix $\boldsymbol{T}_n$, the resulting matrix is an identity matrix up to a scalar. Based on this useful identity, we will then construct approximate inverse-free preconditioners for Toeplitz matrices. We will show that if the Toeplitz matrix is generated by a positive bounded function and its entries enjoys the off-diagonal decay property, then the eigenvalues of the preconditioned matrix are clustered around one. Experimental results in Section 4 show that such approximate inverse-free preconditioners are superior to other popular preconditioners in the literature.

The rest of this paper is organized as follows. In §2, we give some preliminary results on Toeplitz matrice and the Gohberg-Semencul formula. Then we construct approximate inverse-free preconditioners and present our numerical algorithm. In §3, we give a convergence analysis on the preconditioning method. In §4, we present numerical examples to demonstrate that the proposed preconditioners are both efficient and effective. Finally concluding remarks are given in §5.

# 2 The Gohberg-Sememcul Formula and Construction of Preconditioner

In this section, we first present the Gohberg-Sememcul formula. We then construct preconditioners for Toeplitz systems.

## 2.1 The Gohberg-Sememcul Formula

Let us first introduce some notations for our discussion. Here we assume that $z_i (i = 1, 2 \ldots, n)$ are either scalars or block matrices and $\boldsymbol{J}_n$ is the $n \times n$ anti-identity matrix. We let $\boldsymbol{z} = (z_1, z_2, \ldots, z_n)^t$ and denote $\boldsymbol{L}(\boldsymbol{z})$ as the lower-triangular Toeplitz matrix with its first column entries being given by $\boldsymbol{z}$, and $\widetilde{\boldsymbol{L}}(\boldsymbol{z})$ denote the lower-triangular Toeplitz matrix with its first column being given by $(0, z_1, \ldots, z_{n-1})^t$, i.e.,

$$\boldsymbol{L}(\boldsymbol{z}) = \begin{pmatrix} z_1 & & & 0 \\ z_2 & z_1 & & \\ \vdots & \ddots & \ddots & \\ z_n & \cdots & z_2 & z_1 \end{pmatrix} \quad \text{and} \quad \widetilde{\boldsymbol{L}}(\boldsymbol{z}) = \begin{pmatrix} 0 & & & 0 \\ z_1 & 0 & & \\ \vdots & \ddots & \ddots & \\ z_{n-1} & \cdots & z_1 & 0 \end{pmatrix}. \tag{3}$$

Similarly, we denote $\boldsymbol{U}(\boldsymbol{z})$ as the upper-triangular Toeplitz matrix with its first row being given by $(z_n, z_{n-1}, \ldots, z_1)^t$, and $\widetilde{\boldsymbol{U}}(\boldsymbol{z})$ denote the upper-triangular Toeplitz matrix with its first column being given by $(0, z_n, \ldots, z_2)^t$, i.e.,

$$\boldsymbol{U}(\boldsymbol{z}) = \begin{pmatrix} z_n & z_{n-1} & \cdots & z_1 \\ & \ddots & \ddots & \vdots \\ & & z_n & z_{n-1} \\ 0 & & & z_n \end{pmatrix} \quad \text{and} \quad \widetilde{\boldsymbol{U}}(\boldsymbol{z}) = \begin{pmatrix} 0 & z_n & \cdots & z_2 \\ & \ddots & \ddots & \vdots \\ & & 0 & z_n \\ 0 & & & 0 \end{pmatrix}. \tag{4}$$

Let

$$\boldsymbol{x}_n = (x_1, x_2, \ldots, x_n)^t \quad \text{and} \quad \boldsymbol{y}_n = (y_1, y_2, \ldots, y_n)^t$$

be respectively the solutions of the following linear systems

$$\begin{cases} \boldsymbol{T}_n \boldsymbol{x}_n & = & \boldsymbol{e}_n^{(1)} \\ \boldsymbol{T}_n \boldsymbol{y}_n & = & \boldsymbol{e}_n^{(n)}. \end{cases} \tag{5}$$

Here $\boldsymbol{e}_n^{(1)} = (1, 0, \ldots, 0)^t$ and $\boldsymbol{e}_n^{(n)} = (0, \ldots, 0, 1)^t$ are the unit vectors, the subscript denotes the length of the vector and the superscript denotes the position of the nonzero element. It is well-known that the inverse of a Toeplitz matrix can be written as a sum of multiplications of lower-triangular and upper-triangular Toeplitz matrices [15]. If the linear system (5) is solvable, then the inverse of $\boldsymbol{T}_n$ can be constructed by using the famous Gohberg-Sememcul formula [15].

For any positive definite Toeplitz matrix $\boldsymbol{T}_n$, if $\boldsymbol{x}_n$ and $\boldsymbol{y}_n$ are the solutions of (5), we must have $x_1 \neq 0$, see for instance [18] and the inverse of $\boldsymbol{T}_n$ can be represented in terms of $\boldsymbol{x}_n$ and $\boldsymbol{y}_n$ by using Gohberg-Sememcul formula [15] as follows:

$$\boldsymbol{T}_n^{-1} = \frac{1}{x_1}[\boldsymbol{L}(\boldsymbol{x}_n)\boldsymbol{U}(\boldsymbol{y}_n) - \widetilde{\boldsymbol{L}}(\boldsymbol{y}_n)\widetilde{\boldsymbol{U}}(\boldsymbol{x}_n)]. \tag{6}$$

When $x_1$ is zero, Ben-Artzi and Shalom [2], Labahn and Shalom [21], Ng, Rost and Wen [25] and Heinig [17] have studied the representation. Furthermore, if the nonsingular matrix $\boldsymbol{T}_n$ is well-conditioned, Gohberg-Sememcul formula have been shown to be numerically forward stable, see for instance [16, 32].

### 2.2 An Identity

In this subsection, we present an interesting property that when the triangular Toeplitz matrices (3) and (4) pre-multiplies and post-multiplies the original Toeplitz matrix $\boldsymbol{T}_n$, the resulting matrix is an identity matrix up to a scalar.

The following lemma gives the relationship between the solutions of (5) from the size of $n$ to the size of $2n$.

**Lemma 1** *Let $\boldsymbol{T}_n = (t_{i-j})_{ij}$ be a nonsingular Toeplitz matrix and we denote*

$$\boldsymbol{c}_n = (t_0, t_1, \ldots, t_{n-1})^t \quad \text{and} \quad \boldsymbol{r}_n = (t_0, t_{-1}, \ldots, t_{1-n})^t.$$

*Suppose that $\boldsymbol{x}_n$ and $\boldsymbol{y}_n$ satisfy*

$$\boldsymbol{T}_n\boldsymbol{x}_n = \boldsymbol{e}_n^{(1)} \quad \text{and} \quad \boldsymbol{T}_n\boldsymbol{y}_n = \boldsymbol{e}_n^{(n)},$$

*respectively and $x_1 \neq 0$ then the following equations are solvable:*

$$\boldsymbol{L}(\boldsymbol{x}_n)\boldsymbol{u}_n = -\widetilde{\boldsymbol{U}}(\boldsymbol{c}_n)\boldsymbol{x}_n \quad \text{and} \quad \boldsymbol{U}(\boldsymbol{y}_n)\boldsymbol{v}_n = -\widetilde{\boldsymbol{L}}(\boldsymbol{J}_n\boldsymbol{r}_n)\boldsymbol{y}_n. \tag{7}$$

*Here*

$$\boldsymbol{u}_n = (u_1, u_2, \ldots, u_n)^t \quad \text{and} \quad \boldsymbol{v}_n = (v_1, v_2, \ldots, v_n)^t.$$

*Moreover, if $\boldsymbol{P}_{2n}$ is defined as*

$$\boldsymbol{P}_{2n} = \begin{pmatrix} \boldsymbol{T}_n & \boldsymbol{A}_n \\ \boldsymbol{B}_n & \boldsymbol{T}_n \end{pmatrix} \tag{8}$$

*where*

$$\boldsymbol{A}_n = \widetilde{\boldsymbol{L}}(\boldsymbol{J}_n\boldsymbol{r}_n) + \boldsymbol{U}(\boldsymbol{v}_n) \quad \text{and} \quad \boldsymbol{B}_n = \boldsymbol{L}(\boldsymbol{u}_n) + \widetilde{\boldsymbol{U}}(\boldsymbol{c}_n), \tag{9}$$

*then*

$$\boldsymbol{P}_{2n}\begin{pmatrix} \boldsymbol{x}_n \\ 0 \end{pmatrix} = \boldsymbol{e}_{2n}^{(1)} \quad \text{and} \quad \boldsymbol{P}_{2n}\begin{pmatrix} 0 \\ \boldsymbol{y}_n \end{pmatrix} = \boldsymbol{e}_{2n}^{(2n)}.$$

**Proof:** Since $x_1 = y_n \neq 0$, we know that $\boldsymbol{L}(\boldsymbol{x}_n)$ and $\boldsymbol{U}(\boldsymbol{y}_n)$ are nonsingular. Now because

$$\boldsymbol{B}_n\boldsymbol{x}_n = \boldsymbol{L}(\boldsymbol{u}_n)\boldsymbol{x}_n + \widetilde{\boldsymbol{U}}(\boldsymbol{c}_n)\boldsymbol{x}_n = \boldsymbol{L}(\boldsymbol{u}_n)\boldsymbol{x}_n - \boldsymbol{L}(\boldsymbol{x}_n)\boldsymbol{u}_n$$

and

$$\boldsymbol{L}(\boldsymbol{u}_n)\boldsymbol{x}_n = \boldsymbol{L}(\boldsymbol{x}_n)\boldsymbol{u}_n$$

we have

$$\begin{pmatrix} \boldsymbol{T}_n & \boldsymbol{A}_n \\ \boldsymbol{B}_n & \boldsymbol{T}_n \end{pmatrix}\begin{pmatrix} \boldsymbol{x}_n \\ 0 \end{pmatrix} = \begin{pmatrix} \boldsymbol{T}_n\boldsymbol{x}_n \\ \boldsymbol{B}_n\boldsymbol{x}_n \end{pmatrix} = \boldsymbol{e}_{2n}^{(1)}.$$

Similarly, we note that

$$\boldsymbol{U}(\boldsymbol{y}_n)\boldsymbol{v}_n = \boldsymbol{U}(\boldsymbol{v}_n)\boldsymbol{y}_n,$$

thus we have

$$\boldsymbol{A}_n\boldsymbol{y}_n = \widetilde{\boldsymbol{L}}(\boldsymbol{J}_n\boldsymbol{r}_n)\boldsymbol{y}_n + \boldsymbol{U}(\boldsymbol{v}_n)\boldsymbol{y}_n = \widetilde{\boldsymbol{L}}(\boldsymbol{J}_n\boldsymbol{r}_n)\boldsymbol{y}_n + \boldsymbol{U}(\boldsymbol{y}_n)\boldsymbol{v}_n = 0.$$

Therefore, we obtain

$$\begin{pmatrix} \boldsymbol{T}_n & \boldsymbol{A}_n \\ \boldsymbol{B}_n & \boldsymbol{T}_n \end{pmatrix}\begin{pmatrix} 0 \\ \boldsymbol{y}_n \end{pmatrix} = \begin{pmatrix} \boldsymbol{A}_n\boldsymbol{y}_n \\ \boldsymbol{T}_n\boldsymbol{y}_n \end{pmatrix} = \boldsymbol{e}_{2n}^{(2n)}.$$

Thus the lemma is proved. □

**Theorem 1** *Let $\boldsymbol{T}_n$ be an $n\times n$ positive definite Toeplitz matrix, $\boldsymbol{x}_n$ and $\boldsymbol{y}_n$ be the solutions of (5), $\boldsymbol{L}(\boldsymbol{x}_n), \boldsymbol{U}(\boldsymbol{y}_n), \widetilde{\boldsymbol{L}}(\boldsymbol{y}_n)$ and $\widetilde{\boldsymbol{U}}(\boldsymbol{x}_n)$ be defined as the formula in (3) and (4). Then we have*

$$\boldsymbol{U}(\boldsymbol{y}_n)\boldsymbol{T}_n\boldsymbol{L}(\boldsymbol{x}_n) - \widetilde{\boldsymbol{L}}(\boldsymbol{y}_n)\boldsymbol{T}_n\widetilde{\boldsymbol{U}}(\boldsymbol{x}_n) = x_1\boldsymbol{I}_n \tag{10}$$

*where $\boldsymbol{I}_n$ is an $n \times n$ identity matrix.*

**Proof:** Since $\boldsymbol{x}_n$ and $\boldsymbol{y}_n$ are the solutions of (5), $\boldsymbol{T}_n$ is invertible and $x_1 \neq 0$ and $y_n \neq 0$, see for instance [18]. Let $\boldsymbol{P}_{2n}$ be defined by (8), and we define

$$\widehat{\boldsymbol{x}}_{2n} = (\boldsymbol{x}_n^t, 0)^t \quad \text{and} \quad \widehat{\boldsymbol{y}}_{2n} = (0, \boldsymbol{y}_n^t)^t.$$

Here $\boldsymbol{P}_{2n}^{-1}$ can be represented by using the Goherg-Sememcul formula as follows:

$$\begin{aligned}
\boldsymbol{P}_{2n}^{-1} &= \frac{1}{x_1}[\boldsymbol{L}(\widehat{\boldsymbol{x}}_{2n})\boldsymbol{U}(\widehat{\boldsymbol{y}}_{2n}) - \widetilde{\boldsymbol{L}}(\widehat{\boldsymbol{y}}_{2n})\widetilde{\boldsymbol{U}}(\widehat{\boldsymbol{x}}_{2n})] \\
&= \frac{1}{x_1}\left[\begin{pmatrix} \boldsymbol{L}(\boldsymbol{x}_n) & 0 \\ \widetilde{\boldsymbol{U}}(\boldsymbol{x}_n) & \boldsymbol{L}(\boldsymbol{x}_n) \end{pmatrix}\begin{pmatrix} \boldsymbol{U}(\boldsymbol{y}_n) & \widetilde{\boldsymbol{L}}(\boldsymbol{y}_n) \\ 0 & \boldsymbol{U}(\boldsymbol{y}_n) \end{pmatrix} - \begin{pmatrix} 0 & 0 \\ \widetilde{\boldsymbol{L}}(\boldsymbol{y}_n) & 0 \end{pmatrix}\begin{pmatrix} 0 & \widetilde{\boldsymbol{U}}(\boldsymbol{x}_n) \\ 0 & 0 \end{pmatrix}\right] \\
&= \frac{1}{x_1}\begin{pmatrix} \boldsymbol{L}(\boldsymbol{x}_n)\boldsymbol{U}(\boldsymbol{y}_n) & \boldsymbol{L}(\boldsymbol{x}_n)\widetilde{\boldsymbol{L}}(\boldsymbol{y}_n) \\ \widetilde{\boldsymbol{U}}(\boldsymbol{x}_n)\boldsymbol{U}(\boldsymbol{y}_n) & \widetilde{\boldsymbol{U}}(\boldsymbol{x}_n)\widetilde{\boldsymbol{L}}(\boldsymbol{y}_n) + \boldsymbol{L}(\boldsymbol{x}_n)\boldsymbol{U}(\boldsymbol{y}_n) - \widetilde{\boldsymbol{L}}(\boldsymbol{y}_n)\widetilde{\boldsymbol{U}}(\boldsymbol{x}_n) \end{pmatrix}.
\end{aligned}$$

We then partition $\boldsymbol{P}_{2n}^{-1}$ into a 2-by-2 block matrix as follows:

$$\boldsymbol{P}_{2n}^{-1} \equiv \begin{pmatrix} \boldsymbol{P}_{11} & \boldsymbol{P}_{12} \\ \boldsymbol{P}_{21} & \boldsymbol{P}_{22} \end{pmatrix} \tag{11}$$

and denote the Schur complement of $\boldsymbol{P}_{2n}^{-1}$ as

$$\boldsymbol{K} = \boldsymbol{P}_{22} - \boldsymbol{P}_{21}\boldsymbol{P}_{11}^{-1}\boldsymbol{P}_{12}.$$

Then we have

$$\boldsymbol{P}_{2n} = \begin{pmatrix} \boldsymbol{P}_{11}^{-1} + \boldsymbol{P}_{11}^{-1}\boldsymbol{P}_{12}\boldsymbol{K}^{-1}\boldsymbol{P}_{21}\boldsymbol{P}_{11}^{-1} & -\boldsymbol{P}_{11}^{-1}\boldsymbol{P}_{12}\boldsymbol{K}^{-1} \\ -\boldsymbol{K}^{-1}\boldsymbol{P}_{21}\boldsymbol{P}_{11}^{-1} & \boldsymbol{K}^{-1} \end{pmatrix}.$$

Comparing with $\boldsymbol{P}_{2n}$ with Equation (8), we have

$$\boldsymbol{P}_{11}^{-1} + \boldsymbol{P}_{11}^{-1}\boldsymbol{P}_{12}\boldsymbol{K}^{-1}\boldsymbol{P}_{21}\boldsymbol{P}_{11}^{-1} = \boldsymbol{T}_n \tag{12}$$

and

$$\boldsymbol{K}^{-1} = \boldsymbol{T}_n.$$

Substituting

$$\boldsymbol{P}_{11} = \frac{1}{x_1}\boldsymbol{L}(\boldsymbol{x}_n)\boldsymbol{U}(\boldsymbol{y}_n), \quad \boldsymbol{P}_{12} = \frac{1}{x_1}\boldsymbol{L}(\boldsymbol{x}_n)\widetilde{\boldsymbol{L}}(\boldsymbol{y}_n), \quad \text{and} \quad \boldsymbol{P}_{21} = \frac{1}{x_1}\widetilde{\boldsymbol{U}}(\boldsymbol{x}_n)\boldsymbol{U}(\boldsymbol{y}_n)$$

into Equation (12), we obtain

$$x_1\boldsymbol{U}(\boldsymbol{y}_n)^{-1}\boldsymbol{L}(\boldsymbol{x}_n)^{-1} + \boldsymbol{U}(\boldsymbol{y}_n)^{-1}\widetilde{\boldsymbol{L}}(\boldsymbol{y}_n)\boldsymbol{T}_n\widetilde{\boldsymbol{U}}(\boldsymbol{x}_n)\boldsymbol{L}(\boldsymbol{x}_n)^{-1} = \boldsymbol{T}_n.$$

Therefore the result follows. □

We note that if $\boldsymbol{T}_n$ is a symmetric Toeplitz matrix, using the fact that $\boldsymbol{y}_n = \boldsymbol{J}_n\boldsymbol{x}_n$, then it is straightforward to prove the following theorem.

**Theorem 2** *Let $\boldsymbol{T}_n$ be an $n \times n$ symmetric Toeplitz matrix, $\boldsymbol{x}_n$ be the solution of (5),*

$\boldsymbol{L}(\boldsymbol{x}_n)$ *and* $\widetilde{\boldsymbol{U}}(\boldsymbol{x}_n)$ *are defined in (3). Then we have*

$$\boldsymbol{L}(\boldsymbol{x}_n)^t \boldsymbol{T}_n \boldsymbol{L}(\boldsymbol{x}_n) - \widetilde{\boldsymbol{U}}(\boldsymbol{x}_n)^t \boldsymbol{T}_n \widetilde{\boldsymbol{U}}(\boldsymbol{x}_n) = x_1 \boldsymbol{I}_n. \tag{13}$$

### 2.3 Approximate Inverse-Free Preconditioner

In this subsection, we construct an approximate inverse-free preconditioners for Toeplitz matrices. We focus on Toeplitz matrices $\boldsymbol{T}_{2n} = (t_{i-j})_{2n\times 2n}$ such that $t_k = t^*_{-k}$. If $t_k$ are scalars, then $\boldsymbol{T}_{2n}$ is a symmetric matrix. However, if $t_k$ are block matrices, we do not assume that $\boldsymbol{T}_{2n} = \boldsymbol{T}^t_{2n}$.

For simplicity, we only consider the scalar matrices. We note that there is a natural partitioning of a Toeplitz matrix into $2 \times 2$ blocks as follows:

$$\boldsymbol{T}_{2n} = \begin{pmatrix} \boldsymbol{T}_n & \boldsymbol{S}_n \\ \boldsymbol{S}^t_n & \boldsymbol{T}_n \end{pmatrix} \tag{14}$$

where $\boldsymbol{T}_n$ is the principal submatrix of $\boldsymbol{T}_{2n}$ and $\boldsymbol{S}_n$ is also an $n \times n$ Toeplitz matrix. We propose to use $\boldsymbol{P}_{2n}$ defined in (8) as a preconditioner for $\boldsymbol{T}_{2n}$. We remark that $\boldsymbol{P}_{2n}$ is also a symmetric matrix.

**Lemma 2** *Let* $\boldsymbol{T}_n$ *be a symmetric Toeplitz matrix, and* $\boldsymbol{P}_{2n}$ *be defined in Lemma 1, then* $\boldsymbol{P}_{2n}$ *is also a symmetric Toeplitz matrix.*

**Proof:** We note that $\boldsymbol{x}_n$ and $\boldsymbol{y}_n$ are the first column and the last column of $\boldsymbol{T}_n$ respectively and we have $\boldsymbol{y}_n = \boldsymbol{J}_n \boldsymbol{x}_n$. Since

$$\widetilde{\boldsymbol{L}}(\boldsymbol{J}_n \boldsymbol{r}_n) = \boldsymbol{J}_n \widetilde{\boldsymbol{U}}(\boldsymbol{c}_n) \boldsymbol{J}_n \quad \text{and} \quad \boldsymbol{L}(\boldsymbol{x}_n) = \boldsymbol{J}_n \boldsymbol{U}(\boldsymbol{y}_n) \boldsymbol{J}_n,$$

from (7) we have

$$\boldsymbol{J}_n \boldsymbol{U}(\boldsymbol{y}_n) \boldsymbol{J}_n \boldsymbol{u}_n = -\widetilde{\boldsymbol{U}}(\boldsymbol{c}_n) \boldsymbol{x}_n \quad \text{and} \quad \boldsymbol{U}(\boldsymbol{y}_n) \boldsymbol{v}_n = -\boldsymbol{J}_n \widetilde{\boldsymbol{U}}(\boldsymbol{r}_n) \boldsymbol{J}_n \boldsymbol{y}_n.$$

Since

$$\boldsymbol{J}_n \boldsymbol{J}_n = \boldsymbol{I}_n \quad \text{and} \quad \boldsymbol{U}(\boldsymbol{c}_n) = \boldsymbol{U}(\boldsymbol{r}_n)$$

we have

$$\begin{aligned} \boldsymbol{J}_n \boldsymbol{U}(\boldsymbol{y}_n) \boldsymbol{v}_n &= -\boldsymbol{J}_n \boldsymbol{J}_n \widetilde{\boldsymbol{U}}(\boldsymbol{r}_n) \boldsymbol{J}_n \boldsymbol{y}_n \\ &= -\widetilde{\boldsymbol{U}}(\boldsymbol{c}_n) \boldsymbol{J}_n \boldsymbol{y}_n \\ &= -\widetilde{\boldsymbol{U}}(\boldsymbol{c}_n) \boldsymbol{x}_n \\ &= \boldsymbol{J}_n \boldsymbol{U}(\boldsymbol{y}_n) \boldsymbol{J}_n \boldsymbol{u}_n. \end{aligned}$$

Hence we obtain $\boldsymbol{u}_n = \boldsymbol{J}_n \boldsymbol{v}_n$. From (9) we conclude that $\boldsymbol{A}_n = \boldsymbol{B}^t_n$ and the result follows.

$\square$

We recall that the inverse of $\boldsymbol{P}_{2n}$ is given by

$$\boldsymbol{P}_{2n}^{-1} = \frac{1}{\hat{x}_1}[\boldsymbol{L}(\widehat{\boldsymbol{x}}_{2n})\boldsymbol{L}(\widehat{\boldsymbol{x}}_{2n})^t - \widetilde{\boldsymbol{L}}(\widehat{\boldsymbol{x}}_{2n})\widetilde{\boldsymbol{L}}(\widehat{\boldsymbol{x}}_{2n})^t] \tag{15}$$

where $\widehat{\boldsymbol{x}}_{2n} = (\boldsymbol{x}_n^t, 0)^t$.

We remark that it is not necessary to construct $\boldsymbol{P}_{2n}$ explicitly. Once

$$\boldsymbol{T}_n\boldsymbol{x}_n = \boldsymbol{e}_n^{(1)} \tag{16}$$

is solved, the inverse of $\boldsymbol{P}_{2n}$ can be represented by using (15). We will show that $\boldsymbol{P}_{2n}$ is a good preconditioner for $\boldsymbol{T}_{2n}$. However, the inverses of $\boldsymbol{P}_{2n}$ involves the solution of Equation (16). The computational cost can be expensive. Therefore in the remains of this subsection, we present a recursive method to construct the preconditioner $\boldsymbol{P}_{2n}$ efficiently.

### 2.4 The Recursive Scheme and Computational Cost

In fact, when the solution $\boldsymbol{x}_n$ is obtained, $\boldsymbol{P}_{2n}^{-1}$ can be represented by the formula in (15). Equation (16) can be solved efficiently by using the Preconditioned Conjugate Gradient (PCG) method with the preconditioner $\boldsymbol{P}_n$. The solution of $\boldsymbol{x}_{n/2}$ involved in the preconditioner $\boldsymbol{P}_n$ can be recursively generated by solving Equation (16) until the size of the linear system is sufficiently small. The procedure of the recursive computation of $\boldsymbol{P}_{2n}$ is described as follows.

---

Procedure Input($T_k$,$k$) Output($\boldsymbol{x}_k$)

---

If $k \le N$, then
  solve the linear system
  $$\boldsymbol{T}_k\boldsymbol{x}_k = \boldsymbol{e}_k^{(1)}$$
  exactly by direct methods;
else
  compute $\boldsymbol{x}_{k/2}$ by using the procedure with the input matrix $\boldsymbol{T}_{k/2}$ and the integer $k/2$;
  construct $\boldsymbol{P}_k^{-1}$ by using the output $\boldsymbol{x}_{k/2}$ via the formula in (11);
  solve the linear system $\boldsymbol{T}_k\boldsymbol{x}_k = \boldsymbol{e}_k^{(1)}$ by using the preconditioned conjugate gradient method with the preconditioner $\boldsymbol{P}_k$;
end.

---

We remark that the above procedure is suitable for scalar case. For the block case, one needs to solve one more equation

$$\boldsymbol{T}_k\boldsymbol{y}_k = \boldsymbol{e}_k^{(k)}.$$

Since the procedure is similar to that of the scalar case, we therefore omitted here.

The main computational cost of the method comes from the matrix-vector multiplications of the forms $\boldsymbol{T}_{2n}\boldsymbol{z}$ and $\boldsymbol{P}_{2n}^{-1}\boldsymbol{z}$ in each PCG iteration, where $\boldsymbol{z}$ is an $n \times 1$ vector. For the scalar case, the overall operations is of $\mathrm{O}(n \log n)$. The total cost of the recursive procedure is also $\mathrm{O}(n \log n)$. For the block case, if the block size is $m \times m$ and the number of block is $n$, then the total cost is $\mathrm{O}(m^2 n \log n + m^3 n)$. For more details on computational costs required for both the scalar and the block cases, we refer readers to [26].

## 3 Convergence Analysis

In this section, we give an convergence analysis of PCG method with our proposed preconditioner. Here the strong norm are used to study the asymptotic behavior. We first give the definition of a strong norm. Let $\boldsymbol{A} = \{a_{k,j}\}$ be a $n \times n$ matrix, the strong norm $\|\boldsymbol{A}\|$ is defined by

$$\|\boldsymbol{A}\|_2 = \max_{\boldsymbol{x} \neq 0} \frac{\|\boldsymbol{A}\boldsymbol{x}\|_2}{\|\boldsymbol{x}\|_2}.$$

In many theoretical and practical problems, the entries of Toeplitz matrices enjoy an exponential or polynomial decay in their off-diagonals, see for instance [30] and the references therein. We then focus on Toeplitz systems with certain decay properties. We need the the following definitions.

**Definition 3** *[30] Let $\boldsymbol{A} = [a_{i,j}]_{i,j \in \mathcal{I}}$ be a matrix, where the index set is $\mathcal{I} = \mathbb{Z}$, $\mathbb{N}$, or $\{1, 2, \ldots, N\}$.*

1. *$\boldsymbol{A}$ belongs to the space $\mathcal{E}_\gamma$ if*

$$|a_{i,j}| \leq c e^{-\gamma|i-j|} \quad \textit{for} \;\; \gamma > 0$$

*and some constant $c > 0$.*

2. *$\boldsymbol{A}$ belongs to the space $\mathcal{Q}_s$ if*

$$|a_{i,j}| \leq c(1 + |i - j|)^{-s} \quad \textit{for} \;\; s > 1,$$

*and some constant $c > 0$.*

For a sequence $\{t_j\}$, if its entries enjoy a certain decay properties, we have the following lemma.

**Lemma 3** *Let $t_j (j = 0, 1, 2, \ldots)$ be a sequence satisfying*

$$|t_j| \leq c e^{-\gamma|j|} \tag{17}$$

*for some $c > 0$ and $\gamma > 0$, or*

$$|t_j| \leq c(|j| + 1)^{-s} \tag{18}$$

*for some $c > 0$ and $s > 1$. Then the sequence $\{t_j\}$ is absolutely summable, i.e., $\sum_{k=0}^{\infty} |t_j|$ is bounded, and for any given $\epsilon > 0$, there exists a constant $K > 0$ independent of $n$, such that for all $n > K$,*

$$\sum_{k=K}^{n} |t_j| < \epsilon. \tag{19}$$

**Proof:** First we consider the case that

$$|t_j| \leq ce^{-\gamma|j|}.$$

We have

$$\sum_{k=p}^{n} |t_j| \leq \sum_{k=p}^{n} ce^{-\gamma j} = \frac{1}{1-e^{-\gamma}} ce^{-\gamma p}(1 - e^{-\gamma(n-p+1)}) < \frac{ce^{-\gamma p}}{1-e^{-\gamma}}$$

Therefore, when $p = 0$, then $\{t_j\}$ is absolutely summable. Let

$$K > -\gamma^{-1} \ln(c^{-1}(1-e^{-\gamma})\epsilon),$$

we obtain (19).

When $|t_j| \leq c(|j|+1)^{-s}$, we have

$$\sum_{k=p}^{n} |t_j| \leq \sum_{k=p}^{n} c(|j|+1)^{-s} \leq \sum_{k=p}^{\infty} c(|j|+1)^{-s} \leq c \int_{p-1}^{\infty} (x+1)^{-s} dx \leq \frac{cp^{1-s}}{s-1}.$$

Therefore $\{t_j\}$ is absolutely summable. Let

$$K > \left(\frac{c}{(s-1)\epsilon}\right)^{1/(s-1)}$$

then (19) holds.

□

With the above definitions, the following theorem shows that the off-diagonal decay property is preserved under the inverse operator.

**Theorem 4** *[20] Let $\boldsymbol{A} : l^2(\mathcal{I}) \to l^2(\mathcal{I})$ be an invertible matrix, where $\mathcal{I} = \mathbb{Z}, \mathbb{N}$ or $\{1, 2, \ldots, N\}$.*

1. *If $\boldsymbol{A} \in \mathcal{E}_\gamma$, then $\boldsymbol{A}^{-1} \in \mathcal{E}_{\gamma_1}$ for some $\gamma_1 \in (0, \gamma)$.*
2. *If $\boldsymbol{A} \in \mathcal{Q}_s$, then $\boldsymbol{A}^{-1} \in \mathcal{Q}_s$.*

The following theorem shows that the two sequences $\boldsymbol{T}_{2n}$ and $\boldsymbol{P}_{2n}$ can be very close to each other as $n$ goes to infinity.

**Theorem 5** *Let $\boldsymbol{T}_n$ be an $n \times n$ positive definite symmetric Toeplitz matrix with its diagonal entries satisfying*

$$|t_j| \leq ce^{-\gamma|j|} \tag{20}$$

*for some $c > 0$ and $\gamma > 0$, or*

$$|t_j| \leq c(|j| + 1)^{-s} \tag{21}$$

*for some $c > 0$ and $s > 1$. Then for any given $\epsilon_0 > 0$, there exists a constant $K > 0$ independent of $n$, such that for all $n > K$,*

$$\|\boldsymbol{T}_{2n} - \boldsymbol{P}_{2n}\|_2 < \epsilon_0. \tag{22}$$

*Here $\boldsymbol{P}_{2n}$ is defined in (8). Moreover, for any given $\epsilon_1 > 0$, there exists a constant $K_1 > 0$ such that for all $n > K_1$,*

$$\|\boldsymbol{P}_{2n}^{-1} - \boldsymbol{T}_{2n}^{-1}\|_2 < \epsilon_1. \tag{23}$$

**Proof:** Here we will only consider the case

$$|t_j| \leq ce^{-\gamma|j|},$$

the second case can be proved similarly.

Denote

$$\boldsymbol{E} = \boldsymbol{T}_{2n} - \boldsymbol{P}_{2n} \quad \text{and} \quad \boldsymbol{t}_2 = (t_n, t_{n+1}, \ldots, t_{2n-1})^t,$$

then we have

$$\begin{aligned} \boldsymbol{E} = \boldsymbol{T}_{2n} - \boldsymbol{P}_{2n} &= \begin{pmatrix} 0 & \boldsymbol{S}_n - \boldsymbol{A}_n \\ \boldsymbol{S}_n^t - \boldsymbol{A}_n^t & 0 \end{pmatrix} \\ &= \begin{pmatrix} 0 & \boldsymbol{L}(\boldsymbol{t}_2)^t - \boldsymbol{L}(\boldsymbol{u}_n)^t \\ \boldsymbol{L}(\boldsymbol{t}_2) - \boldsymbol{L}(\boldsymbol{u}_n) & 0 \end{pmatrix} \end{aligned}$$

where $\boldsymbol{A}_n$ and $\boldsymbol{S}_n$ are defined in (8) and (14) respectively. Therefore, we have

$$\|\boldsymbol{E}\|_1 = \|\boldsymbol{L}(\boldsymbol{t}_2) - \boldsymbol{L}(\boldsymbol{u}_n)\|_1 = \|\boldsymbol{t}_2 - \boldsymbol{u}_n\|_1 \leq \|\boldsymbol{t}_2\|_1 + \|\boldsymbol{u}_n\|_1. \tag{24}$$

Since $\boldsymbol{T}_n \in \mathcal{E}_\gamma$, we have

$$\|\boldsymbol{t}_2\|_1 = \sum_{k=n}^{2n-1} |t_k| \leq \sum_{k=n}^{2n-1} ce^{-\gamma|k|} \leq \frac{ce^{-\gamma n}}{1 - e^{-\gamma}}$$

Therefore, for any given $\epsilon > 0$, if

$$n > -\gamma^{-1} \ln(c^{-1}(1 - e^{-\gamma})\epsilon),$$

we have

$$\|\boldsymbol{t}_2\|_1 < \epsilon.$$

Next, we will estimate the bound of $\|\boldsymbol{u}_n\|_1$. We note that

$$\boldsymbol{u}_n = -\boldsymbol{L}(\boldsymbol{x}_n)^{-1}\widetilde{\boldsymbol{U}}(\boldsymbol{c}_n)\boldsymbol{x}_n,$$

we estimate the bound of $\|\boldsymbol{L}(\boldsymbol{x}_n)^{-1}\|_1$ and $\widetilde{\boldsymbol{U}}(\boldsymbol{c}_n)\boldsymbol{x}_n$.

We first estimate of the bound of $\|\boldsymbol{L}(\boldsymbol{x}_n)^{-1}\|_1$. ¿From Theorem 4, we deduce that

$$|x_j| \leq c_1 e^{-\gamma_1 |j|}$$

for some constant $c_1$ and $\gamma_1 \in (0, \gamma)$. We know $\boldsymbol{L}(\boldsymbol{x}_n) \in \mathcal{E}_{\gamma_1}$, from which we obtain $\boldsymbol{L}(\boldsymbol{x}_n)^{-1} \in \mathcal{E}_{\gamma_2}$ for some constant $\gamma_2 \in (0, \gamma_1)$.

We assume that $a_{i,j}$ are the entries of $\boldsymbol{L}(\boldsymbol{x}_n)^{-1}$. Since $\boldsymbol{L}(\boldsymbol{x}_n)$ is a lower-triangle matrix, $\boldsymbol{L}(\boldsymbol{x}_n)^{-1}$ is a lower-triangle matrix too. We consider the sum of the $j$-th column of $\boldsymbol{L}(\boldsymbol{x}_n)^{-1}$ as follows:

$$\sum_{k=1}^{n} a_{k,j} = \sum_{k=j}^{n} a_{k,j} \leq \sum_{k=j}^{n} c_2 e^{-\gamma_2(k-j)} < \frac{c_2}{1 - e^{-\gamma_2}}.$$

Denote

$$M = \frac{c_2}{1 - e^{-\gamma_2}},$$

we obtain

$$\|\boldsymbol{L}(\boldsymbol{x}_n)^{-1}\|_1 < M. \tag{25}$$

We then estimate the bound of $\widetilde{\boldsymbol{U}}(\boldsymbol{c}_n)\boldsymbol{x}_n$. Since $\boldsymbol{T}_n$ enjoys an off-diagonal decay property, from Theorem 4 and Lemma 3, there exist two constants $M_1, M_2$ independent of $n$, such that

$$\sum_{k=0}^{2n-1} |t_k| < M_1 \quad \text{and} \quad \sum_{k=1}^{n} |x_n| < M_2.$$

Thus, for any given $\varepsilon$, there exists a constant $N_1$, such that for all $n > N_1$,

$$\sum_{k=N_1}^{2n-1} |t_k| < \varepsilon.$$

Moreover, there exists a constant $N_2$, such that for all $n > N_2$,

$$\sum_{k=N_2}^{n} |x_n| < \varepsilon.$$

We denote

$$\widetilde{\boldsymbol{x}}_n = (0, \ldots, 0, x_{n-N_1+2}, x_{n-N_1+3}, \ldots, x_n)^t$$

and

$$\boldsymbol{E}_1 = \begin{pmatrix} 0 & \widehat{\boldsymbol{E}}_1 \\ 0 & 0 \end{pmatrix}$$

with

$$\widehat{\boldsymbol{E}}_1 = \begin{pmatrix} t_{N_1} & t_{N_1-1} & \cdots & t_2 \\ & t_{N_1} & \ddots & \vdots \\ & & \ddots & t_{N_1-1} \\ 0 & & & t_{N_1} \end{pmatrix}.$$

For the above analysis, if $n > N_1 + N_2$, we have

$$\|\widetilde{\boldsymbol{U}}(\boldsymbol{c}_n) - \boldsymbol{E}_1\|_1 < \varepsilon, \quad \|\boldsymbol{E}_1\|_1 < M_1, \quad \|\widetilde{\boldsymbol{x}}_n\|_1 < \varepsilon \quad \text{and} \quad \|\boldsymbol{x}_n\|_1 < M_2.$$

Thus, using $\boldsymbol{E}_1\boldsymbol{x}_n = \boldsymbol{E}_1\widetilde{\boldsymbol{x}}_n$, we have

$$\begin{aligned} \|\widetilde{\boldsymbol{U}}(\boldsymbol{c}_n)\boldsymbol{x}_n\|_1 &= \|(\widetilde{\boldsymbol{U}}(\boldsymbol{c}_n) - \boldsymbol{E}_1)\boldsymbol{x}_n + \boldsymbol{E}_1\widetilde{\boldsymbol{x}}_n\|_1 \\ &\leq \|(\widetilde{\boldsymbol{U}}(\boldsymbol{c}_n) - \boldsymbol{E}_1)\boldsymbol{x}_n + \boldsymbol{E}_1\widetilde{\boldsymbol{x}}_n\|_1 \\ &\leq \|(\widetilde{\boldsymbol{U}}(\boldsymbol{c}_n) - \boldsymbol{E}_1)\|_1\|\boldsymbol{x}_n\|_1 + \|\boldsymbol{E}_1\|_1\|\widetilde{\boldsymbol{x}}_n\|_1 \\ &< (M_1 + M_2)\varepsilon. \end{aligned} \tag{26}$$

By using (25) and (26), we obtain

$$\begin{aligned} \|\boldsymbol{u}_n\|_1 &= \| - \boldsymbol{L}(\boldsymbol{x}_n)^{-1}\widetilde{\boldsymbol{U}}(\boldsymbol{c}_n)\boldsymbol{x}_n\|_1 \\ &\leq \|\boldsymbol{L}(\boldsymbol{x}_n)^{-1}\|_1\|\widetilde{\boldsymbol{U}}(\boldsymbol{c}_n)\boldsymbol{x}_n\|_1 \\ &< M(M_1 + M_2)\varepsilon. \end{aligned}$$

Thus, by using (24), we obtain

$$\|\boldsymbol{E}\|_1 < \|\boldsymbol{t}_2\|_2 + \|\boldsymbol{u}_n\|_2 < \varepsilon + M(M_1 + M_2)\varepsilon.$$

Similarly, we also have

$$\|\boldsymbol{E}\|_\infty < \varepsilon + M(M_1 + M_2)\varepsilon.$$

By letting

$$\varepsilon_0 = \frac{1}{M(M_1 + M_2) + x_1}\epsilon,$$

(22) holds.

We note that both $\boldsymbol{T}_{2n}^{-1}$ and $\boldsymbol{P}_{2n}^{-1}$ are bounded, then (23) holds. □

**Theorem 6** *Let $\boldsymbol{T}_n$ be a Toeplitz matrix generated by a positive definite function $f(\theta)$, and $\boldsymbol{P}_{2n}$ is defined in (8). Then for any given $\epsilon > 0$, there exists a constant $K > 0$ independently of $n$, such that for all $n > K$, all eigenvalues of $\boldsymbol{P}_{2n}^{-1}\boldsymbol{T}_{2n}$ lie inside the* *interval* $(1-\epsilon, 1+\epsilon)$.

Now we consider a more interesting case that the generating function $f(\theta)$ has zero(s). Suppose that the generating function $f(\theta) \in C_{2\pi}$ be non-negative real function and have finitely many zeros. Let $\theta_k (k = 1, 2, \ldots, m)$ be all roots of $f(\theta)$ in $[-\pi, \pi)$ with order $2a_j$. One can write

$$f(\theta) = h(\theta)|w(\theta)|^2, \quad -\pi \leq \theta < \pi \tag{27}$$

where

$$w(\theta) = \prod_{k=1}^{m} (1 - e^{\mathrm{i}(\theta - \theta_k)})^{a_k}$$

and $h(\theta) > 0$. It is easy to see that $\mathcal{T}_n[w]$ is a lower triangular Toeplitz matrix with bandwidth

$$a = \sum_{k=1}^{m} a_k.$$

By straightforward calculation, we obtain

$$\mathcal{T}_n[f] = \mathcal{T}_n[w]^* \mathcal{T}_n[h] \mathcal{T}_n[w] + \boldsymbol{G}_n \tag{28}$$

where $\boldsymbol{G}_n$ has only non-zeros entries in its last $a$ columns and its last $a$ rows and therefore its rank is less than $2a$, see also [7, 23].

**Theorem 7** *Let $\boldsymbol{T}_n$ be an $n \times n$ positive definite Toeplitz matrix generated by $f(\theta)$ in (27) with its diagonal entries satisfying (20) or (21) for some $c > 0$. Then for any given $\epsilon_0 > 0$, there exists a constant $K > 0$ independent of $n$, such that for all $n > K$, at most $2a$ eigenvalues of*

$$\mathcal{T}_n[f] - \mathcal{T}_n[w]^* \mathcal{P}_n[h] \mathcal{T}_n[w]$$

*have absolute value exceeding $\epsilon_0$.*

**Proof:** Notice that $h(\theta) > 0$, from Theorem 6 we know that for any given $\epsilon > 0$, there exists a constant $K > 0$ independent of $n$, such that for all $n > K$, we have

$$\|\mathcal{T}_n[h] - \mathcal{P}_n[h]\|_2 < \epsilon.$$

Let $\epsilon_0 = \|\mathcal{T}_{2n}[w]\|_2^2 \epsilon$, we have

$$\begin{aligned} & \mathcal{T}_n[f] - \mathcal{T}_n[w]^* \mathcal{P}_n[h] \mathcal{T}_n[w] \\ = \quad & \mathcal{T}_n[w]^* \left( \mathcal{T}_n[h] - \mathcal{P}_n[h] \right) \mathcal{T}_n[w] + \boldsymbol{G}_n \end{aligned}$$

where $\mathrm{rank}(\boldsymbol{G}_{2n}) \leq 2a$ and

$$\|\mathcal{T}_n[w]^* \left( \mathcal{T}_n[h] - \mathcal{P}_n[h] \right) \mathcal{T}_n[w]\|_2 < \epsilon_0.$$

This means that

$$\mathcal{T}_{2n}[f] - \mathcal{T}_{2n}[w]^*\mathcal{P}_{2n}[h]\mathcal{T}_{2n}[w]$$

is the sum of a matrix with small 2-norm and a matrix with rank $a$. Hence the result is proved. $\square$

**Theorem 8** *Let $\boldsymbol{T}_n$ be an $n \times n$ positive definite Toeplitz matrix generated by a complex function $f(\theta)$ in Wiener class with roots $\theta_k (k = 1, 2, \ldots, m)$ in $[-\pi, \pi)$ with order $a_j$. Suppose that the function*

$$w_1(\theta) = \prod_{k=1}^{m}(1 - e^{\mathrm{i}(\theta-\theta_k)})^{a_k}$$

*and $h_1(\theta) > 0$ satisfies:*

$$|f(\theta)|^2 = h_1(\theta)|w_1(\theta)|^2, \quad -\pi \le \theta < \pi. \tag{29}$$

*Then for any given $\epsilon_0 > 0$, there exists a constant $K > 0$ and $M = M(\epsilon_0)$ independent of $n$, such that for all $n > K$, we have*

$$\mathcal{T}_n[f]^*\mathcal{T}_n[f] = \mathcal{T}_n[w_1]\left(\mathcal{P}_n[h_1] + \boldsymbol{B}_n\right)\mathcal{T}_n[w_1]^* + \boldsymbol{F}_n(M)$$

*where $\|\boldsymbol{B}_n\|_2 < \epsilon_0$ and rank$(\boldsymbol{F}_n(M)) \le M$.*

**Proof:** Denote $h_2(\theta) = f(\theta)/w_1(\theta)$, then we have $h_1(\theta) = |h_2(\theta)|^2$. By straightforward calculation, we have

$$\mathcal{T}_n[f] = \mathcal{T}_n[h_2]\mathcal{T}_n[w_1] + \widehat{\boldsymbol{G}}_n$$

where $\widehat{\boldsymbol{G}}_n$ has only non-zeros entries in its last $a$ columns and therefore its rank is less than $a$, see for instance [23]. Therefore

$$\mathcal{T}_n[f]^*\mathcal{T}_n[f] = \mathcal{T}_n[w_1]^*\mathcal{T}_n[h_2]^*\mathcal{T}_n[h_2]\mathcal{T}_n[w_1] + \widetilde{\boldsymbol{G}}_n. \tag{30}$$

Here

$$\widetilde{\boldsymbol{G}}_n = \mathcal{T}_n[w_1]^*\mathcal{T}_n[h_2]^*\widehat{\boldsymbol{G}}_n + \widehat{\boldsymbol{G}}_n^*\mathcal{T}_n[h_2]\mathcal{T}_n[w_1] + \widehat{\boldsymbol{G}}_n^*\widehat{\boldsymbol{G}}_n$$

is a matrix whose rank is at most $s$.

Chan et al. [7] proved that for any given $\epsilon > 0$, there exists a constant $K > 0$ and $M = M(\epsilon)$ independent of $n$, such that for all $n > K$, we have

$$\mathcal{T}_n[h_2]^*\mathcal{T}_n[h_2] = \mathcal{T}_n[|h_2|^2] + \widetilde{\boldsymbol{B}}_n + \widetilde{\boldsymbol{F}}_n(M) \tag{31}$$

where $\|\widetilde{\boldsymbol{B}}_n\|_2 < \epsilon_0$ and rank$(\widetilde{\boldsymbol{F}}_n(M)) \le M$.

We note that

$$h_1(\theta) = |h_2(\theta)|^2 > 0.$$

¿From Theorem 6, we know for any given $\epsilon > 0$, there exists a constant $K_1 > 0$ independent of $n$, such that for all $n > K_1$, we have

$$\|\mathcal{T}_n[|h_2|^2] - \mathcal{P}_n[h_1]\|_2 = \|\mathcal{T}_n[h_1] - \mathcal{P}_n[h_1]\|_2 < \epsilon. \tag{32}$$

Together with (30), (31) and (32), this yields the assertion. $\square$

We remark all the results in this section concerning scalar Toeplitz matrices can be extended to the case of block-Toeplitz matrices.

# 4 Numerical Results

In this section, we apply the proposed numerical algorithm to solve positive definite Toeplitz systems

$$T_n \boldsymbol{x}_n = \boldsymbol{e}_n^{(1)}.$$

We remark that for the block case, one also needs to solve one more systems

$$T_n \boldsymbol{y}_n = \boldsymbol{e}_n^{(n)}.$$

All the numerical tests were done on a Compaq Evo N800v with Pentium(R) 4 Mobile CPU1.70GHz with Matlab 6.5.

For the purpose of comparison, we also give the number of iterations by PCG method without preconditioning ($I$), the with Strang's circulant preconditioner ($S$) [29], with T. Chan's circulant preconditioner ($C$) [10], with the best circulant preconditioner ($B_6$) of order 6 [9], with the Recursive-Based Preconditioner Method (RBM) ($B$) [26] and also with our preconditioner ($P$). Here "**" in the tables signifies that convergence was not attained in 1000 iterations. The stopping criteria is

$$\tau = \frac{\|r_q\|_2}{\|r_0\|_2} \leq 10^{-6},$$

where $r_q$ is the residual vector after $q$ iterations. While the initial guess for our preconditioner is $(\boldsymbol{x}_{n/2}^t, 0)^t$ and for others is the zero vector, since $\boldsymbol{x}_{n/2}$ is obtained during the recursive process in our preconditioner, while it is unknown for others preconditioners. We remark that our preconditioner is constructed recursively. For instance, when we consider the case $n = 128$, i.e., to solve $T_{128}\boldsymbol{x}_{128} = e_{128}^{(1)}$, the preconditioner is constructed by solving $T_{64}\boldsymbol{x}_{64} = \boldsymbol{e}_{64}^{(1)}$ using the preconditioned conjugate gradient method with the stopping criteria being $\tau$, and using the direct solver for $T_{32}\boldsymbol{x}_{32} = \boldsymbol{e}_{32}^{(1)}$. The initial guess for $T_{64}\boldsymbol{x}_{64} = \boldsymbol{e}_{64}^{(1)}$ is $(\boldsymbol{x}_{32}^t, 0)^t$ and for $T_{128}\boldsymbol{x}_{128} = \boldsymbol{e}_{128}^{(1)}$ is $(\boldsymbol{x}_{64}^t, 0)^t$.

In the following numerical examples, we test both well-conditioned linear systems and ill-condition systems. It is interesting to note that our preconditioner performed very well also for the later case. The readers can find more examples in [12] when the generator

|  | $\theta^4+1$ | | | | | |
|---|---|---|---|---|---|---|
| $n$ | $I$ | $S$ | $C$ | $B_6$ | $B$ | $P$ |
| 64 | 50 | 6 | 6 | 6 | 5 | 2 |
| 128 | 61 | 6 | 6 | 6 | 5 | 2 |
| 256 | 67 | 6 | 6 | 6 | 4 | 1 |
| 512 | 69 | 6 | 6 | 6 | 4 | 1 |
| 1024 | 70 | 6 | 6 | 6 | 2 | 1 |
| 2048 | 70 | 6 | 6 | 6 | 1 | 1 |
| 4096 | 70 | 6 | 6 | 6 | 1 | 1 |

Table 1: No. of Iterations for convergence for well-conditioned system.

function is a matrix and in [33] when the coefficient matrix is non-symmetric.

The first tested example is the Toeplitz matrices generated by the positive function

$$f(\theta) = \theta^4 + 1.$$

The entries $t_k$ are given by $t_0 = \pi^4/5 + 1$ and

$$t_k = \begin{cases} \pi^4/5 + 1, & k = 0 \\ (-1)^k \left( \frac{4\pi^2}{k^2} - \frac{24}{k^4} \right), & k = \pm 1, \pm 2, \ldots, \pm(n-1). \end{cases}$$

We have

$$|t_k| \leq 135(1+k)^{-2}$$

and it is a well-conditioned system. The number of iterations required for convergence is presented in Table 1. For the case of $n = 128$, we give the spectra of different preconditioned matrices in Figure 1 and the condition numbers of the preconditioned matrices are $96.22, 22.30, 20.58, 22.23, 22.31$ and $1.00$ for $I, S, C, B_6, B$ and $P$ respectively.

The second and third tested examples are the Toeplitz matrices generated by the functions

$$f(\theta) = \theta^2 \quad \text{and} \quad f(\theta) = \theta^4$$

respectively. It is clearly that the entries enjoy the polynomial decay properties off the diagonal. The number of iteration is presented in Table 2. For the generating function $f(\theta) = \theta^4$. We note that PCG does not converge in 1000 iterations for both the Strang's circulant preconditioners and T. Chan's circulant preconditioners when $n \geq 1024$. The number of iterations required for convergence are presented in Table 2. ¿From the numerical results, we observe that our preconditioner is indeed efficient and its performance is better than the circulant preconditioners. For the case of $n = 128$, the spectra of different preconditioned matrices are shown in Figures 2 and 3 respectively.

| | $\theta^2$ | | | | | | $\theta^4$ | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $n$ | $I$ | $S$ | $C$ | $B_6$ | $B$ | $P$ | $I$ | $S$ | $C$ | $B_6$ | $B$ | $P$ |
| 64 | 78 | 6 | 14 | 9 | 5 | 6 | 233 | 33 | 41 | 16 | 9 | 8 |
| 128 | 170 | 6 | 17 | 9 | 5 | 5 | 929 | 53 | 79 | 18 | 9 | 8 |
| 256 | 361 | 7 | 22 | 9 | 5 | 5 | ** | 109 | 181 | 19 | 9 | 8 |
| 512 | 753 | 7 | 29 | 9 | 5 | 5 | ** | 270 | 464 | 22 | 11 | 7 |
| 1024 | ** | 7 | 38 | 9 | 5 | 5 | ** | ** | ** | 24 | 13 | 7 |
| 2048 | ** | 7 | 53 | 9 | 5 | 5 | ** | ** | ** | 26 | 13 | 7 |
| 4096 | ** | 7 | 72 | 10 | 5 | 4 | ** | ** | ** | 35 | 15 | 7 |

Table 2: No. of Iterations for convergence for ill-conditioned systems.

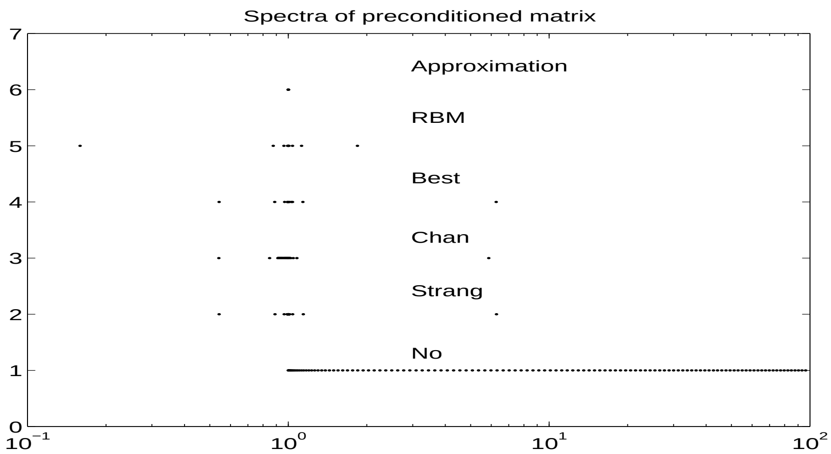


Figure 1: Spectra of the preconditioned matrices for the generating function $\theta^4+1$ with $n = 128$.

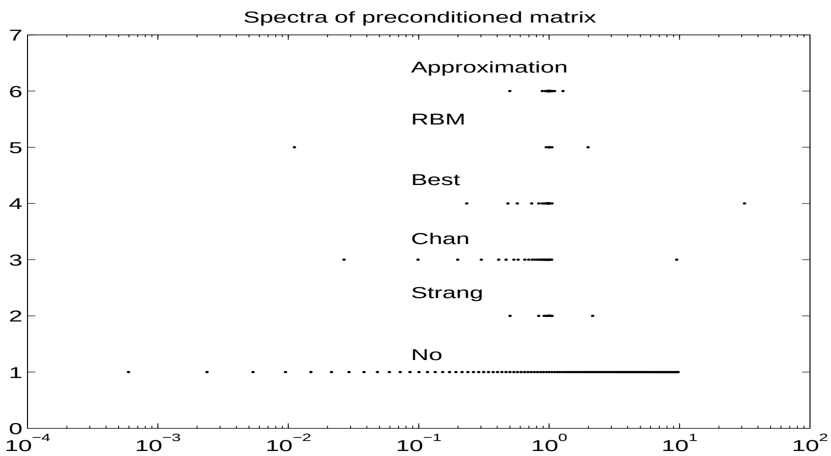


Figure 2: Spectra of the preconditioned matrices for the generating function $\theta^2$ with $n = 128$.

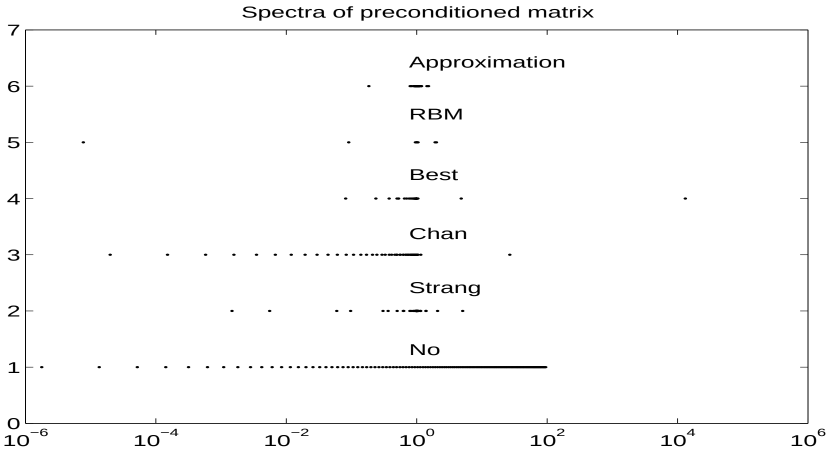


Figure 3: Spectra of the preconditioned matrices for the generating function $\theta^4$ with $n = 128$.

# 5 Concluding Remarks

In this paper, we consider the solutions of positive definite Toeplitz systems. We introduce an interesting property of a Toeplitz matrix with the relation to the Gohberg-Semencul formula. We construct an approximate inverse of Toeplitz matrices which can be used as a preconditioner. We prove that if a sequence of Toeplitz matrices is generated by a positive bounded function, then the spectrum of the preconditioned matrices is uniformly clustering around one. Hence the conjugate gradient methods when applied to solving the preconditioned Toeplitz systems will converge very fast. In the numerical experiments, we consider both the well-conditioned systems and the ill-conditioned linear systems. Numerical results indicated that the proposed preconditioner is efficient in both cases and is superior to other existing preconditioners in the literature.

# References

[1] G. Ammar and W. Gragg. Superfast solution of real positive definite Toeplitz systems. *SIAM Journal on Matrix Analysis and Applications*, 9:61–76, 1988.

[2] A. Ben-Artzi and T. Shalom. On inversion of Toeplitz and close to Toeplitz matrices. *Linear Algebra and Its Applications*, 75:173–192, 1986.

[3] R. Bitmead and B. Anderson. Asymptotically fast solution of Toeplitz and related systems of linear equations. *Linear Algebra and Its Applications*, 34:103–116, 1980.

[4] R. Brent, F. Gustavson, and D. Yun. Fast solution of Toeplitz systems of equations and computation of padé approximants. *Journal of Algorithm*, 1:259–295, 1980.

[5] R. Chan. Circulant preconditioners for Hermitian Toeplitz systems. *SIAM Journal on Matrix Analysis and Applications*, 10:542–550, 1989.

[6] R. Chan and M. Ng. Conjugate gradient methods for Toeplitz systems. *SIAM Review*, 38:427–482, 1996.

[7] R. Chan, D. Potts, and G.Steidl. Preconditioners for non-Hermitian Toeplitz systems. *Numerical Linear Algebra with Applications*, 8:83–98, 2001.

[8] R. Chan and G. Strang. Toeplitz equations by conjugate gradients with circulant preconditioner. *SIAM Journal on Scientific Computing*, 10(1):104–119, 1989.

[9] R. Chan, A. Yip, and M. Ng. The best circulant preconditioners for Hermitian Toeplitz systems. *SIAM Journal on Numerical Analysis*, 38:876–896, 2001.

[10] T. Chan. An optimal circulant preconditioner for Toeplitz systems. *SIAM Journal on Scientific Computing*, 9:766–771, 1988.

[11] W. Ching. Iterative methods for queuing and manufacturing Systems, *Springer Monographs in Mathematics*, Springer, London, 2001.

[12] W. Ching, M. Ng, and Y. Wen. Block diagonal and schur complement preconditioners for block Toeplitz systems. *revised and re-submitted to SIAM Journal on Matrix Analysis and Applications.*

[13] F. de Hoog. A new algorithm for solving Toeplitz systems of equations. *Linear Algebra and Its Applications*, 88-89:123–138, 1987.

[14] J. Durbin. The fitting of time series models,. *International Statistical Review*, 28:233–244, 1960.

[15] I. Gohberg and A. Semencul. On the inversion of finite Toeplitz matrices and their continuous analogs. *Matematicheskie Issledovaniya*, 2:201–233, 1972.

[16] M. Gutknecht and M. Hochbruck. The stability of inversion formulas for Toeplitz matrices. *Linear Algebra and Its Applications*, 223-224:307–324, 1995.

[17] G. Heinig. On the reconstruction of Toeplitz matrix inverses from columns. *Linear Algebra and Its Applications*, 350:199–212, 2002.

[18] G. Heinig and K. Rost. *Algebraic Methods for Toeplitz-like Matrices and Operators.* Birkhäuser Verlag, Basel, 1984.

[19] T. Huckle. *Superfast Solution of Linear Equations with Low Displacement Rank.* High Performance Algorithms for Structured Matrix Problems, Adv. Theory Comput. Comput. Math., II, Nova Sci. Publ., 1998.

[20] S. Jaffard. Propriétés des matrices "bien localisées" près deleur diagonale et quelques applications. *Ann. Inst. H. Poincaré Anal. Non Linéaire,*, 7:461–476, 1990.

[21] G. Labahn and T. Shalom. Inversion of Toeplitz matrices with only two standard equations. *Linear Algebra and Its Applications*, 175:143–158, 1992.

[22] N. Levinson. The Wiener RMS (root mean square) error criterion in filter design and prediction. *Journal of Mathematical Physics*, 25:261–278, 1946.

[23] F. Lin and W. Ching. Inverse Toeplitz preconditioners for Hermitian Toeplitz systems. *Numerical Linear Algebra with Applications*, 12(2-3):221–229, 2005.

[24] M.Morf. Doubling algorithms for Toeplitz and related equations. *Proceedings International Conference on Acoustics, Speech, and Signal Processing*, 3:954–959, 1980.

[25] M. Ng, K. Rost, and Y. Wen. On inversion of Toeplitz matrices. *Linear Algebra and Its Applications*, 348, 2002.

[26] M. Ng, H. Sun, and X. Jin. Recursive-based PCG methods for Toeplitz systems with nonnegative generating functions. *SIAM Journal on Scientific Computing*, 24:1507–1529, 2003.

[27] D. Noutsos and P. Vassalos. New band Toeplitz preconditioners for ill-conditioned symmetric positive definite Toeplitz systems. *SIAM Journal on Matrix Analysis and Applications*, 23:728–743, 2002.

[28] S. Serra. Optimal, quasi-optimal and superlinear band Toeplitz preconditioners for asymptotically ill-conditioned positive definite Toeplitz systems. *Mathematics of Computation*, 66:651–665, 1997.

[29] G. Strang. A proposal for Toeplitz matrix calculations. *Studies in Applied Mathematics*, 74:171–176, 1986.

[30] T. Strohmer. Four short stories about Toeplitz matrix calculations. *Linear Algebra and Its Applications*, 343/344:321–344, 2002.

[31] W. Trench. An algorithm for the inversion of finite Toeplitz matrices. *SIAM Journal on Applied Mathematics*, 12:515–522, 1964.

[32] Y. Wen, M. Ng, W. Ching, and H. Liu. A note on the stability of Toeplitz matrix inversion formulas. *Applied Mathematics Letters*, 17, 2004.

[33] Y. Wen, W. Ching, and M. Ng. Fast solvers for queueing systems with negative customers. *The first international conference on performance evaluation methodologies and tools*, to appear in ACM proceedings, 2007.