Noname manuscript No.
(will be inserted by the editor)

Circulant preconditioners for analytic functions of Toeplitz matrices

Sean Hon · Andrew Wathen

Received: date / Accepted: date

Abstract Circulant preconditioning for symmetric Toeplitz systems has been well developed over the past few decades. For a large class of such systems, descriptive bounds on convergence for the conjugate gradient method can be obtained. For (real) nonsymmetric Toeplitz systems, much work had been focused on normalising the original systems until [J. Pestana and A. J. Wathen. SIAM J. MATRIX ANAL. APPL. Vol. 36, No. 1, pp. 273-288] recently showed that theoretic guarantees on convergence for the minimal residual method can be established via the simple use of reordering. The authors further proved that a suitable absolute value circulant preconditioner can be used to ensure rapid convergence. In this paper, we show that the related ideas can also be applied to the systems defined by analytic functions of (real) nonsymmetric Toeplitz matrices. For the systems defined by analytic functions of non-Hermitian Toeplitz matrices, we also show that certain circulant preconditioners are effective. Numerical examples with the conjugate gradient method and the minimal residual method are given to support our theoretical results.

Keywords Toeplitz matrices · Functions of matrices · Circulant preconditioners · PCG · PMINRES

Mathematics Subject Classification (2000) 65F08 · 65F10 · 65F15 · 15A16 · 15B05

Sean Hon
Mathematical Institute, University of Oxford, Radcliffe Observatory Quarter, Oxford, OX2 6GG, United Kingdom
E-mail: hon@maths.ox.ac.uk

Andrew Wathen
Mathematical Institute, University of Oxford, Radcliffe Observatory Quarter, Oxford, OX2 6GG, United Kingdom
E-mail: wathen@maths.ox.ac.uk

1 Introduction

Instead of directly dealing with the $n \times n$ (real) nonsymmetric Toeplitz system $A_n\mathbf{x} = \mathbf{b}$, Pestana and Wathen suggested in [14, 11] that one can pre(or post)multiply by the "flip" matrix Y_n, defined as

$$Y_n = \begin{bmatrix} & & 1 \\ & \iddots & \\ 1 & & \end{bmatrix} \in \mathbb{R}^{n\times n},$$

to obtain the symmetric system $Y_nA_n\mathbf{x} = Y_n\mathbf{b}$. Using a suitable absolute value circulant matrix $|C_n|$ as a preconditioner in which C_n is derived from A_n in a standard way, they further proved that the eigenvalues of the preconditioned matrix $|C_n|^{-1}Y_nA_n$ are clustered at ± 1. The absolute value circulant matrix is readily computed from the diagonalisation provided by a Fast Fourier Transform. One can then use the minimal residual method (MINRES) for solving such symmetric but possibly indefinite system with a guarantee of convergence which depends only on its spectrum.

In this work, we show that Pestana and Wathen's idea of using Y_n as a reordering device and $|C_n|$ as a preconditioner can both also be applied to the systems defined by analytic functions of (real) nonsymmetric Toeplitz matrices, i.e. systems of the form $h(A_n)\mathbf{x} = \mathbf{b}$. Note that $h(A_n)$ is generally not Toeplitz. Based on those ideas, we show that one can solve the symmetric system $Y_nh(A_n)\mathbf{x} = Y_n\mathbf{b}$ using MINRES with a guarantee of convergence which depends only the eigenvalues of $Y_nh(A_n)$ instead of solving its normal equation system. In particular, we also show that $|h(C_n)|^{-1}Y_nh(A_n)$ can be decomposed into the sum of a unitary matrix, a matrix of small norm and a matrix of low rank under suitable assumptions.

Our previous work [8] showed that $g(c(A_n))$ could be an effective preconditioner for $g(A_n)$, where $g(z)$ is the trigonometric function e^z, $\sin z$ or $\cos z$, A_n is the Toeplitz matrix generated by a continuous complex-valued function defined on $[-\pi, \pi]$ and $c(A_n)$ is the optimal circulant preconditioner [4] for A_n. In this work, instead of assuming a trigonometric function, we extend it to the larger class of analytic functions.

Functions of Toeplitz matrices have some crucial applications. For example, e^{A_n} arises from the discretisation of integro-differential equations with a shift-invariant kernel [9]. Solving those equations is often required in areas such as the option pricing [5, 15]. Related work on computing the exponential of a block Toeplitz matrix arising in approximations of Markovian fluid queues can also be found in [2]. As for the matrix cosine and sine functions, an example of applications is solving the following system of second order differential equations [7]

$$\frac{d}{dt^2}y + A_n^2 y = 0,\; y(0) = y_0,\; y'(0) = y_0',$$

whose solution is given by

$$y(t) = \cos(tA_n)y_0 + A_n^{-1}\sin(tA_n)y_0'.$$

It is noted that $|h(C_n)|$ and $h(C_n)$ are circulant matrices. By the diagonalisation of a circulant matrix $C_n = U_n^* \Lambda_n U_n$, where U_n is a Fourier matrix in which the entries are given by $[U_n]_{jk} = \frac{1}{\sqrt{n}} e^{-2\pi \mathbf{i} jk/n}$ with $j, k = 0, 1, \ldots, n-1$, we have $|h(C_n)| = U_n^* |h(\Lambda_n)| U_n$ and $h(C_n) = U_n^* h(\Lambda_n) U_n$. Therefore, for any vector $\mathbf{d}$ the products $|h(C_n)|^{-1}\mathbf{d}$ and $h(C_n)^{-1}\mathbf{d}$ can be efficiently computed by Fast Fourier Transforms in $\mathcal{O}(n \log n)$ operations.

It must be noted however that fast matrix vector multiplication with the matrix $h(A_n)$ is not readily archived by circulant embedding, as for the simple case $h(z) = z$, though sparsity may still help. Indeed for e^{A_n}, the matrix vector multiplication can be computed efficiently for example by a fast algorithm in [10].

2 Preliminary results on $c(A_n)$ and A_n

In the context of iterative solvers for Toeplitz systems, we assume that the given $n \times n$ Toeplitz matrix A_n is associated with the function f via its Fourier series

$$f(\theta) = \sum_{k=-\infty}^{\infty} a_k e^{ik\theta}$$

defined on $[-\pi, \pi]$. We have

$$A_n = \begin{bmatrix} a_0 & a_{-1} & \cdots & a_{-n+2} & a_{-n+1} \\ a_1 & a_0 & a_{-1} & & a_{-n+2} \\ \vdots & a_1 & a_0 & \ddots & \vdots \\ a_{n-2} & & \ddots & \ddots & a_{-1} \\ a_{n-1} & a_{n-2} & \cdots & a_1 & a_0 \end{bmatrix},$$

where

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

are the Fourier coefficients of f. The function f is called the *generating function* of the Toeplitz matrix. If f is complex-valued, A_n is non-Hermitian for all n. If f is real-valued, A_n is Hermitian for all n. If f is real-valued and positive, A_n is Hermitian positive definite for all n. If f is real-valued and even, A_n is symmetric for all n [12].

We first provide some lemmas concerning Toeplitz matrices. Let $\mathcal{C}[-\pi, \pi]$ be the Banach space of continuous complex-valued functions defined on $[-\pi, \pi]$ with the supremum norm $\|\cdot\|_\infty$.

Lemma 1 *[3, Lemma 1 and 3] Let $f \in \mathcal{C}[-\pi,\pi]$. Let $A_n \in \mathbb{C}^{n\times n}$ be the Toeplitz matrix generated by f and $c(A_n) \in \mathbb{C}^{n\times n}$ be the optimal circulant preconditioner for A_n, i.e. $c(A_n) = \min_{W_n \in \mathcal{M}_{U_n}} \|A_n - W_n\|_F$ where $\mathcal{M}_{U_n}$ is the set of all circulant matrices. Then we have*

$$\|A_n\|_2 \leq 2\|f\|_\infty \quad \text{and} \quad \|c(A_n)\|_2 \leq 2\|f\|_\infty \quad n = 1, 2, \ldots.$$

Lemma 2 *[3, Theorem 1] Let $f \in \mathcal{C}[-\pi,\pi]$. Let A_n be the Toeplitz matrix generated by f and $c(A_n)$ be the optimal circulant preconditioner for A_n. Then for all $\epsilon > 0$ there exist positive integers N and $M > 0$ such that for all $n > N$ we have*

$$c(A_n) - A_n = V_n - W_n,$$

where

$$\text{rank } V_n \leq 2M$$

and

$$\|W_n\|_2 \leq \epsilon.$$

3 The spectra of the preconditioned matrices

In this section, we first provide several lemmas concerning functions of matrices.

Lemma 3 *[6, Theorem 1.18] Let $h(z)$ be analytic on an open subset $\Omega \subseteq \mathbb{C}$ such that each connected component of Ω is closed under conjugation. Consider the corresponding matrix function $h(z)$ on its natural domain in $\mathbb{C}^{n\times n}$, the set $\mathcal{D} = \{A_n \in \mathbb{C}^{n\times n} : \Lambda(A_n) \subseteq \Omega\}$. Then the following are equivalent:*

(a) $h(A_n^*) = h(A_n)^*$ for all $A_n \in \mathcal{D}$.

(b) $h(\overline{A_n}) = \overline{h(A_n)}$ for all $A_n \in \mathcal{D}$.

(c) $h(\mathbb{R}^{n\times n} \cap \mathcal{D}) \subseteq \mathbb{R}^{n\times n}$.

(d) $h(\mathbb{R} \cap \Omega) \subseteq \mathbb{R}$.

Lemma 4 *[6, Theorem 4.7] Suppose $h(z)$ has a Taylor series expansion*

$$h(z) = \sum_{k=0}^{\infty} a_k (z - \alpha)^k,$$

where $a_k = \frac{h^{(k)}(\alpha)}{k!}$, with radius of convergence r. If $A_n \in \mathbb{C}^{n\times n}$ then $h(A_n)$ is defined and is given by

$$h(A_n) = \sum_{k=0}^{\infty} a_k (A_n - \alpha I_n)^k$$

if and only if the distinct eigenvalues $\lambda_1, \cdots, \lambda_s$ of A_n satisfies one of the conditions

(a) $|\lambda_i - \alpha| < r$,

(b) $|\lambda_i - \alpha| = r$ and the series for $h^{(n_i-1)}(\lambda)$, where n_i is the index of λ_i, is convergent at the point $\lambda = \lambda_i$, $i = 1, \ldots, s$.

Lemma 5 *[6, Theorem 4.8] Suppose $h(z)$ has a Taylor series expansion*

$$h(z) = \sum_{k=0}^{\infty} a_k (z-\alpha)^k,$$

where $a_k = \frac{h^{(k)}(\alpha)}{k!}$, with radius of convergence r. If $A_n \in \mathbb{C}^{n\times n}$ with $\rho(A_n - \alpha I_n) < r$ then for any matrix norm $\|\cdot\|$

$$\|h(A_n) - \sum_{k=0}^{K-1} a_k (A_n - \alpha I_n)^k\| \leq \frac{1}{K!} \max_{0\leq t\leq 1} \|(A_n - \alpha I_n)^K h^{(K)}(\alpha I_n + t(A_n - \alpha I_n))\|.$$

The following lemma shows that $Y_n h(A_n)$ is (real) symmetric when A_n is a (real) nonsymmetric Toeplitz matrix.

Lemma 6 *Given $h(z)$ is analytic on $|z| < r$. If $A_n \in \mathbb{R}^{n\times n}$ with $\rho(A_n) < r$ is (real) persymmetric, i.e. $Y_n A_n = A_n^T Y_n$, then $h(A_n)$ is also (real) persymmetric.*

Proof We start by showing that A_n^k is (real) persymmetric for any nonnegative integer k:

$$\begin{aligned} Y_n A^k &= Y_n A_n A_n^{k-1} \\ &= A_n^T Y_n A_n^{k-1} \\ &= A_n^T Y_n A_n A_n^{k-2} \\ &= (A_n^T)^2 Y_n A_n^{k-2} \\ &\vdots \\ &= (A_n^T)^k Y_n \\ &= (A_n^k)^T Y_n. \end{aligned}$$

Since $h(z)$ is analytic on $|z| < r$, it has the following Taylor series representation:

$$h(z) = \sum_{k=0}^{\infty} a_k z^k$$

with the radius of convergence r. By Lemma 4, we have

$$h(A_n) = \sum_{k=0}^{\infty} a_k A_n^k.$$

Thus,

$$\begin{aligned}
Y_n h(A_n) &= Y_n \lim_{K\to\infty} \sum_{k=0}^{K} a_k A_n^k \\
&= \lim_{K\to\infty} \sum_{k=0}^{K} a_k Y_n A_n^k \\
&= \lim_{K\to\infty} \sum_{k=0}^{K} a_k (A_n^k)^T Y_n \\
&= \lim_{K\to\infty} (\sum_{k=0}^{K} a_k (A_n^k))^T Y_n \\
&= h(A_n)^T Y_n.
\end{aligned}$$

Moreover, by Lemma 3 (c), $h(A_n)$ is real when A_n is real. The result follows. □

We are now ready to give our main results on the spectrum of $|h(c(A_n))|^{-1}h(A_n)$. Without loss of generality, we assume that $h(z)$ is represented by the following Taylor series:

$$h(z) = \sum_{k=0}^{\infty} a_k z^k.$$

Theorem 1 *Given $h(z)$ is analytic on $|z| < r$. Let $f \in \mathcal{C}[-\pi,\pi]$ with $2\|f\|_\infty < r$. Let A_n be the Toeplitz matrix generated by f and $c(A_n)$ be the optimal circulant preconditioner for A_n. Then, for all $\epsilon > 0$ there exist integers N and M such that for all $n > N$*

$$h(c(A_n)) - h(A_n) = R_n + E_n,$$

where

$$\begin{aligned}
\textit{rank } R_n &\leq 2M, \\
\|E_n\|_2 &\leq \epsilon.
\end{aligned}$$

Proof Since $h(z)$ is analytic on $|z| < r$, it has the following Taylor series representation:

$$h(z) = \sum_{k=0}^{\infty} a_k z^k$$

with the radius of convergence $r = (\lim_{k\to\infty} |\frac{a_{k+1}}{a_k}|)^{-1}$. By the assumption $2\|f\|_\infty < r$ and Lemma 1, we have

$$\begin{aligned}
r &> 2\|f\|_\infty \\
&> \|A_n\|_2 \\
&> \rho(A_n) \\
&= \max_i \lambda_i(A_n) \\
&\geq \lambda_i(A_n) \quad \text{for} \quad i = 1, 2, \ldots, n,
\end{aligned}$$

where $\lambda_i(A_n)$ denotes the i-th eigenvalue of A_n. Therefore, by Lemma 4, we have

$$h(A_n) = \sum_{k=0}^{\infty} a_k A_n^k.$$

Similarly, we have

$$h(c(A_n)) = \sum_{k=0}^{\infty} a_k c(A_n)^k.$$

We now decompose

$$\begin{aligned} & h(c(A_n)) - h(A_n) \\ = & \underbrace{h(c(A_n)) - \sum_{k=0}^{K} a_k c(A_n)^k}_{G_1} + \underbrace{\sum_{k=0}^{K} a_k c(A_n)^k - \sum_{k=0}^{K} a_k A_n^k}_{B} + \underbrace{\sum_{k=0}^{K} a_k A_n^k - h(A_n)}_{G_2}. \end{aligned}$$

We first get the measure of $\|G_1 + G_2\|_2$. Using Lemmas 5, we have

$$\begin{aligned} & \|G_1 + G_2\|_2 \\ \leq & \|h(c(A_n)) - \sum_{k=0}^{K} a_k c(A_n)^k\|_2 + \|h(A_n) - \sum_{k=0}^{K} a_k A_n^k\|_2 \\ \leq & \frac{\max_{0\leq t\leq 1} \|c(A_n)^{K+1} h^{(k+1)}(tc(A_n))\|_2}{(K+1)!} + \frac{\max_{0\leq t\leq 1} \|A_n^{K+1} h^{(k+1)}(tA_n)\|_2}{(K+1)!} \\ \leq & \frac{\|c(A_n)^{K+1}\|_2}{(K+1)!} \max_{0\leq t\leq 1} \|h^{(k+1)}(tc(A_n))\|_2 + \frac{\|A_n^{K+1}\|_2}{(K+1)!} \max_{0\leq t\leq 1} \|h^{(k+1)}(tA_n)\|_2 \end{aligned}$$

Now, by Lemma 1,

$$\begin{aligned} \max_{0\leq t\leq 1} \|h^{(k+1)}(tA_n)\|_2 &= \max_{0\leq t\leq 1} \|\sum_{k=0}^{\infty} \frac{(K+k+1)!}{k!} a_{K+k+1} (tA_n)^k\|_2 \\ &\leq \max_{0\leq t\leq 1} \sum_{k=0}^{\infty} \frac{(K+k+1)!}{k!} |a_{K+k+1}| \|tA_n\|_2^k \\ &\leq \sum_{k=0}^{\infty} \frac{(K+k+1)!}{k!} |a_{K+k+1}| \|A_n\|_2^k \\ &\leq \sum_{k=0}^{\infty} \frac{(K+k+1)!}{k!} |a_{K+k+1}| (2\|f\|_\infty)^k \end{aligned}$$

We now show that $\sum_{k=0}^{\infty} \frac{(K+k+1)!}{k!} |a_{K+k+1}| (2\|f\|_\infty)^k$ is a convergent series using the ratio test. By the assumption $2\|f\|_\infty < r = (\lim_{k\to\infty} |\frac{a_{k+1}}{a_k}|)^{-1}$, we know that

$$\lim_{k\to\infty} |\frac{a_{K+k+2}}{a_{K+k+1}}|(\frac{K+k+2}{k+1})(2\|f\|_\infty) = \lim_{k\to\infty} |\frac{a_{k+1}}{a_k}|(2\|f\|_\infty)$$
$$< (\frac{1}{r})(2\|f\|_\infty)$$

which is a constant less than 1 because $2\|f\|_\infty < r$.

Therefore, by the ratio test, $\sum_{k=0}^{\infty} \frac{(K+k+1)!}{k!}|a_{K+k+1}|(2\|f\|_\infty)^k$ is convergent. We conclude that $\sum_{k=0}^{\infty} \frac{(K+k+1)!}{k!}|a_{K+k+1}|(2\|f\|_\infty)^k := m_{(2\|f\|_\infty)}$ is independent of n.

Hence,

$$\max_{0\le t\le 1} \|h^{(k+1)}(tA_n)\|_2 \le m_{(2\|f\|_\infty)}.$$

Similarly, we get

$$\max_{0\le t\le 1} \|h^{(k+1)}(tc(A_n))\|_2 \le m_{(2\|f\|_\infty)}.$$

Therefore,

$$\begin{aligned}
&\|G_1 + G_2\|_2 \\
&\le \frac{\|c(A_n)^{K+1}\|_2}{(K+1)!} \max_{0\le t\le 1} \|h^{(k+1)}(tc(A_n))\|_2 + \frac{\|A_n^{K+1}\|_2}{(K+1)!} \max_{0\le t\le 1} \|h^{(k+1)}(tA_n)\|_2 \\
&\le \frac{\|c(A_n)\|_2^{K+1}}{(K+1)!} m_{(2\|f\|_\infty)} + \frac{\|A_n\|_2^{K+1}}{(K+1)!} m_{(2\|f\|_\infty)} \\
&\le \frac{(2\|f\|_\infty)^{K+1}}{(K+1)!} m_{(2\|f\|_\infty)} + \frac{(2\|f\|_\infty)^{K+1}}{(K+1)!} m_{(2\|f\|_\infty)} \\
&= \frac{(2\|f\|_\infty)^{K+1}}{(K+1)!} (2m_{(2\|f\|_\infty)}) := \epsilon_K
\end{aligned}$$

which converges to zero as K goes to infinity. Therefore, for a given $\epsilon_K > 0$, there exists an integer K such that for all $k > K$,

$$\|G_1 + G_2\|_2 \le \epsilon_K, \tag{1}$$

where k is the number of terms in the Taylor series of $c(A_n)$ (or A_n).

We next show that B can be decomposed into a sum of a matrix of fixed rank and a matrix of small norm. Firstly, by Lemma 2, for all $\epsilon > 0$ there exist integers N_1 and $M_2 > 0$ such that for all $n > N_1$, we have

$$c(A_n) - A_n = V_n - W_n,$$

where

$$V_n = \begin{bmatrix} & & & \Diamond & \cdots & \Diamond \\ & & & & \ddots & \vdots \\ & & & & & \Diamond \\ \Diamond & & & & & \\ \vdots & \ddots & & & & \\ \Diamond & \cdots & \Diamond & & & \end{bmatrix}$$

with diamonds representing the non-zero entries,

$$\text{rank } V_n \leq 2M_1$$

and

$$\|W_n\|_2 \leq \epsilon.$$

Rewrite B as

$$\begin{aligned} B &= \sum_{k=0}^{K} a_k c(A_n)^k - \sum_{k=0}^{K} a_k A_n^k \\ &= \sum_{k=1}^{K} a_k (c(A_n)^k - A_n^k) \\ &= \sum_{k=1}^{K} a_k (\sum_{j=0}^{k-1} c(A_n)^j (c(A_n) - A_n) A_n^{k-1-j}) \\ &= \sum_{k=1}^{K} a_k (\sum_{j=0}^{k-1} c(A_n)^j (V_n - W_n) A_n^{k-1-j}) \\ &= \underbrace{\sum_{k=1}^{K} a_k (\sum_{j=0}^{k-1} c(A_n)^j V_n A_n^{k-1-j})}_{R_n} + \underbrace{\sum_{k=1}^{K} a_k (\sum_{j=0}^{k-1} c(A_n)^j W_n A_n^{k-1-j})}_{J}. \end{aligned}$$

Using Lemmas 1, we can estimate the norm of J:

$$\begin{aligned}
\|J\|_2 &= \|\sum_{k=1}^{K} a_k \sum_{j=0}^{k-1} c(A_n)^j W_n A_n^{k-1-j}\|_2 \\
&\leq \sum_{k=1}^{K} |a_k| \| \sum_{j=0}^{k-1} c(A_n)^j W_n A_n^{k-1-j}\|_2 \\
&\leq \|W_n\|_2 \sum_{k=1}^{K} |a_k| \sum_{j=0}^{k-1} \|c(A_n)\|_2^j \|A_n\|_2^{k-1-j} \\
&\leq \|W_n\|_2 \sum_{k=1}^{K} |a_k| \sum_{j=0}^{k-1} (2\|f\|_\infty)^j (2\|f\|_\infty)^{k-1-j} \\
&= \|W_n\|_2 \underbrace{\sum_{k=1}^{K} |a_k| \sum_{j=0}^{k-1} k (2\|f\|_\infty)^{k-1}}_{m_0} \\
&\leq m_0 \epsilon, \qquad (2)
\end{aligned}$$

where m_0 is a constant independent of n.

We now show that rank $R_n \leq 2KM_1$ by first investigating the structure of R_n. Similar to the approach used in the proof of Lemma 3.11 in [13], simple computations yield

$$c(A_n)^\alpha V_n A_n^\beta = \begin{bmatrix} \Diamond \cdots \Diamond & & \Diamond \cdots \Diamond \\ \vdots \;\; \Diamond \;\; \vdots & & \vdots \;\; \Diamond \;\; \vdots \\ \Diamond \cdots \Diamond & & \Diamond \cdots \Diamond \\ & & \\ \Diamond \cdots \Diamond & & \Diamond \cdots \Diamond \\ \vdots \;\; \Diamond \;\; \vdots & & \vdots \;\; \Diamond \;\; \vdots \\ \Diamond \cdots \Diamond & & \Diamond \cdots \Diamond \end{bmatrix},$$

where the diamonds represent the non-zero entries which appear only in the four $(\alpha+1)M_1$ by $(\beta+1)M_1$ blocks in the corners, provided that n is larger than $2\max(\alpha+1, \beta+1)M_1$. Since the rank of

$$R_n = \sum_{k=1}^{K} a_i (\sum_{j=0}^{k-1} c(A_n)^j V_n A_n^{k-1-j})$$

is determined by that of $\sum_{j=0}^{K-1} c(A_n)^j V_n A_n^{K-1-j}$ which is a block matrix with only four non-zero KM_1 by KM_1 blocks in its corners, it follows that the rank of R_n is less than or equal to $2KM_1$ if we assume $n > 2\underbrace{KM_1}_{M}$.

Hence, we pick

$$N := \max\{N_1, 2M\},$$

and, combining (1) and (2), it follows that for all $n > N$ we have

$$\|\underbrace{G_1 + J + G_2}_{E_n}\|_2 \leq m_0\epsilon + \epsilon_K.$$

The result follows. □

Corollary 1 *Given $h(z)$ is analytic on $|z| < r$. Let $f \in \mathcal{C}[-\pi, \pi]$ with $2\|f\|_\infty < r$. Let A_n be the Toeplitz matrix generated by f and $c(A_n)$ be the optimal circulant preconditioner for A_n. If $\|h(c(A_n))^{-1}\|_2$ is bounded for $n = 1, 2, \ldots,$ then for all $\epsilon > 0$ there exist positive integers N and M such that for all $n > N$*

$$|h(c(A_n))|^{-1}h(A_n) = Q_n + \widehat{R_n} + \widehat{E_n},$$

where Q_n is unitary,

$$\text{rank } \widehat{R_n} \leq 2M,$$

$$\|\widehat{E_n}\|_2 \leq \epsilon.$$

Proof As $h(c(A_n))$ is a circulant matrix we write $h(c(A_n)) = U_n^* h(\Lambda_n) U_n$ where $h(\Lambda_n)$ is the diagonal matrix with the eigenvalues of $h(c(A_n))$. We then have

$$\begin{aligned} |h(c(A_n))| &= U_n^*|h(\Lambda_n)|U_n \\ &= U_n^* h(\Lambda_n) U_n \underbrace{U_n^* \text{sign}(h(\Lambda_n))^{-1} U_n}_{Q_n} \\ &= h(c(A_n))Q_n, \end{aligned} \tag{3}$$

where $\text{sign}(h(\Lambda_n)) = \text{diag}(\frac{h(\Lambda_i)}{|h(\Lambda_i)|})$, and Q_n is unitary.

By Theorem 1, we know that for all $\epsilon > 0$, there exist positive integers N and M such that for all $n > N$

$$h(c(A_n)) - h(A_n) = R_n + E_n,$$

where

$$\text{rank } R_n \leq 2M,$$

$$\|E_n\|_2 \leq \epsilon.$$

By the assumption that $\|h(c(A_n))^{-1}\|_2 < m_0$ for $n = 1, 2, \ldots,$ where m is a positive constant independent of n, we have

$$\begin{aligned} h(c(A_n))^{-1}h(A_n) &= I_n + h(c(A_n))^{-1}(h(A_n) - h(c(A_n))) \\ &= I_n + h(c(A_n))^{-1}(-R_n) + h(c(A_n))^{-1}(-E_n). \end{aligned}$$

Further using (4), we have

$$\begin{aligned}|h(c(A_n))|^{-1}h(A_n) &= Q_n h(c(A_n))^{-1}h(A_n)\\ &= Q_n + \underbrace{Q_n h(c(A_n))^{-1}(-R_n)}_{\widehat{R}_n} + \underbrace{Q_n h(c(A_n))^{-1}(-E_n)}_{\widehat{E}_n}.\end{aligned}$$

Since Q_n is unitary, we know

$$\mathrm{rank}(\widehat{R}_n) = \mathrm{rank}(Q_n h(c(A_n))^{-1}R_n) = \mathrm{rank}(R_n) \leq 2M$$

and

$$\|\widehat{E}_n\|_2 = \|Q_n h(c(A_n))^{-1}E_n\|_2 = \|h(c(A_n))^{-1}E_n\|_2 \leq m_0\epsilon.$$

The result follows. □

Assuming f is a real-valued, by Lemma 3 (a), we know that $h(A_n)$ is Hermitian as A_n is Hermitian. Similarly, $h(c(A_n))$ is Hermitian as $c(A_n)$ is Hermitian. Hence, we consider the following two cases: when $h(A_n)$ is Hermitian indefinite, MINRES can be used with $|h(c(A_n))|$ as the preconditioner. In the special case where both $h(A_n)$ and $h(C_n)$ are Hermitian positive definite, $|h(c(A_n))|$ becomes $h(c(A_n))$ and CG can be applied.

Corollary 2 *Given $h(z)$ is analytic on $|z| < r$. Let $f \in \mathcal{C}[-\pi, \pi]$ with $2\|f\|_\infty < r$. Let A_n be the Toeplitz matrix generated by f and $c(A_n)$ be the optimal circulant preconditioner for A_n. If $\|h(c(A_n))^{-1}\|_2$ is bounded for $n = 1, 2, \dots,$ then for all $\epsilon > 0$, there exist positive integers N and M such that for all $n > N$*

$$[|h(c(A_n))|^{-1}h(A_n)]^*|h(c(A_n))|^{-1}h(A_n) = I_n + \overline{R_n} + \overline{E_n},$$

where

$$\begin{aligned}\mathit{rank}\ \overline{R_n} &\leq 4M,\\ \|\overline{E_n}\|_2 &\leq \epsilon.\end{aligned}$$

Proof By Corollary 1, we know that for all $\epsilon > 0$ there exist positive integers N and M such that for all $n > N$

$$|h(c(A_n))|^{-1}h(A_n) = Q_n + \widehat{R_n} + \widehat{E_n},$$

where Q_n is unitary,

$$\begin{aligned}\mathrm{rank}\ \widehat{R_n} &\leq M,\\ \|\widehat{E_n}\|_2 &\leq \epsilon.\end{aligned}$$

We then have

$$\begin{aligned}&[|h(c(A_n))|^{-1}h(A_n)]^*|h(c(A_n))|^{-1}h(A_n)\\ &= (Q_n + \widehat{R_n} + \widehat{E_n})^*(Q_n + \widehat{R_n} + \widehat{E_n})\\ &= Q_n^*Q_n + \underbrace{\widehat{R_n}^*(I_n + \widehat{R_n} + \widehat{E_n}) + (I_n + \widehat{E_n}^*)\widehat{R_n}}_{\overline{R_n}}\\ &\quad + \underbrace{\widehat{E_n} + \widehat{E_n}^* + \widehat{E_n}^*\widehat{E_n}}_{\overline{E_n}}\\ &= I_n + \overline{R_n} + \overline{E_n}.\end{aligned}$$

It immediately follows that rank $\overline{R_n} \leq 4M$ and $\|\overline{E_n}\|_2 \leq \epsilon^2 + 2\epsilon$. □

When $h(A_n)$ is non-Hermitian, we consider its normal equation system as shown in Corollary 2. Since $[|h(c(A_n))|^{-1}h(A_n)]^*|h(c(A_n))|^{-1}h(A_n)$ is Hermitian positive definite, we can CG in this case.

In the special case when $h(A_n)$ is (real) nonsymmetric, it is not necessary to normalise the original system via the following corollary:

Corollary 3 *Given $h(z)$ is analytic on $|z| < r$. Let $f \in \mathcal{C}[-\pi, \pi]$ with real Fourier coefficients and $2\|f\|_\infty < r$. Let A_n be the real Toeplitz matrix generated by f and $c(A_n)$ be the optimal circulant preconditioner for A_n. If $\|h(c(A_n))^{-1}\|_2$ is bounded for $n = 1, 2, \ldots$, then for all $\epsilon > 0$ there exist positive integers N and M such that for all $n > N$*

$$|h(c(A_n))|^{-1}Y_n h(A_n) = Q_n + \widehat{R_n} + \widehat{E_n},$$

where Q_n is orthogonal and (real) symmetric,

$$\text{rank } \widehat{R_n} \leq 2M,$$
$$\|\widehat{E_n}\|_2 \leq \epsilon.$$

Proof As $h(c(A_n))$ is a circulant matrix we write $h(c(A_n)) = U_n^* h(\Lambda_n) U_n$ where $h(\Lambda_n)$ is the diagonal matrix with the eigenvalues of $h(c(A_n))$. We then have

$$\begin{aligned} |h(c(A_n))| &= U_n^*|h(\Lambda_n)|U_n \\ &= U_n^* h(\Lambda_n) U_n (\underbrace{U_n^* \text{sign}(h(\Lambda_n)) U_n}_{\widetilde{C_n}})^{-1} \\ &= h(c(A_n))\widetilde{C_n}^{-1}, \end{aligned} \tag{4}$$

where $\text{sign}(h(\Lambda_n)) = \text{diag}(\frac{h(\Lambda_i)}{|h(\Lambda_i)|})$, and $\widetilde{C_n}$ is orthogonal.

By Theorem 1, we know that for all $\epsilon > 0$, there exist positive integers N and M such that for all $n > N$

$$h(c(A_n)) - h(A_n) = R_n + E_n,$$

where

$$\text{rank } R_n \leq 2M,$$
$$\|E_n\|_2 \leq \epsilon.$$

By (4) and the assumption that $\|h(c(A_n))^{-1}\|_2 < m_0$ for $n = 1, 2, \ldots,$ where m_0 is a positive constant independent of n, we have

$$\begin{aligned} |h(c(A_n))|^{-1}Y_n h(A_n) &= Y_n|h(c(A_n))|^{-1}h(A_n) \\ &= Y_n|h(c(A_n))|^{-1}(h(c(A_n)) - R_n - E_n) \\ &= \underbrace{Y_n\widetilde{C_n}}_{Q_n} + \underbrace{Y_n|h(c(A_n))|^{-1}(-R_n)}_{\widehat{R}_n} + \underbrace{Y_n|h(c(A_n))|^{-1}(-E_n)}_{\widehat{E}_n}. \end{aligned}$$

As $\widetilde{C_n}$ is a real circulant matrix, $Q_n = Y_n\widetilde{C_n}$ is (real) symmetric. Also, as $\widetilde{C_n}$ is orthogonal, we can show that Q_n is orthogonal:

$$\begin{aligned} Q_n^T Q_n &= (Y_n\widetilde{C_n})^T(Y_n\widetilde{C_n}) \\ &= \widetilde{C_n}^T(Y_n^T Y_n)\widetilde{C_n} \\ &= \widetilde{C_n}^T\widetilde{C_n} \\ &= I_n. \end{aligned}$$

We have

$$\mathrm{rank}(\widehat{R}_n) = \mathrm{rank}(Y_n|h(c(A_n))|^{-1}R_n) = \mathrm{rank}(R_n) \leq 2M$$

and

$$\|\widehat{E}_n\|_2 = \|Y_n|h(c(A_n))|^{-1}E_n\|_2 = \|h(c(A_n))^{-1}E_n\|_2 \leq m_0\epsilon.$$

The result follows. □

In the above results, the condition on the uniform boundedness of $\|h(c(A_n))^{-1}\|_2$ has been required. In fact, the validity of this condition depends on the composition of the analytic function $h(z)$ and the generating function f of the Toeplitz matrix A_n. The following examples illustrate the point.

For $h(z) = \sin z$, from

$$\|(\sin c(A_n))^{-1}\|_2 = \max_i |\frac{1}{\sin\lambda_i}|$$

where λ_i is the i-th eigenvalue of $c(A_n)$, we know that $\|(\sin c_n[f])^{-1}\|_2$ could be arbitrarily large since $\sin\lambda_i$ could be close to zero. Therefore, we have needed to assume that $\|(\sin c(A_n))^{-1}\|_2$ is bounded for $n = 1, 2, \ldots$.

As for $h(z) = e^z$, the condition always holds as we can easily show that

$$\|(e^{c(A_n)})^{-1}\|_2 \leq e^{2\|f\|_\infty} \quad n = 1, 2, \ldots.$$

4 Extension to multilevel cases

In this section, we briefly discuss that our results can be extended to the multilevel Toeplitz cases. The related proofs can be done in a similar fashion as in the 1D Toeplitz case. Namely, with the appropriate assumptions, we can show that the preconditioned system defined by an analytic function of a multilevel Toeplitz can be decomposed into the sum of a matrix with small norm and a matrix of low rank for a sufficiently large dimension.

Consider the simple block-Toeplitz-Toeplitz-block (BTTB) matrix as an example, $A_{n,m} = H_n \otimes K_m$ where $H_n \in \mathbb{R}^{n\times n}$ and $K_m \in \mathbb{R}^{m\times m}$ are both real Toeplitz. Introducing the matrix $Y_{n,m} = Y_n \otimes Y_m$, it can be easily shown that $Y_{n,m}A_{n,m} = (Y_n \otimes Y_m)(H_n \otimes K_m) = (Y_nH_n) \otimes (Y_mK_m)$. Thus $Y_{n,m}A_{n,m}$ is symmetric as both Y_nH_n and Y_mK_m are symmetric. Together with an effective absolute value block-circulant-circulant-block (BCCB) preconditioner, we can use MINRES for $Y_{n,m}A_{n,m}$ without considering the normal equations system. The well-known difficulty that the low rank part now has been multiplied by the block dimension is as exactly the same issue in our situation as it does for the more standard well-known situation.

Indeed, this symmetrising technique applies to more general real multilevel Toeplitz matrices since they are in general persymmetric. Consequently, by Lemma 6, analytic functions of real multilevel Toeplitz matrices are also persymmetric. Multiplying these matrices by the matrix Y_n with the appropriate dimension can result the modified yet symmetric matrices. Again, MINRES with effective absolute value multilevel circulant preconditioners can be used.

5 Numerical results

In this section, we demonstrate the effectiveness of our proposed preconditioner $|h(c(A_n))|$ for $h(A_n)\mathbf{x} = \mathbf{b}$ using CG, MINRES and GMRES. Throughout all numerical tests, e^{A_n} is computed by the MATLAB built-in function **expm** whilst $\sin A_n$, $\cos A_n$ and other matrix functions are computed by **funm**. The vector $\mathbf{b}$ is generated by the function **ones(n,1)** and the initial guess is the zero vector. Also, we use the function **pcg** to solve the Hermitian positive definite systems. For Hermitian indefinite systems, we use the function **minres**. As a comparison, GMRES is also used for some systems and it is executed by **gmres**. The stopping criterion used is

$$\frac{\|\mathbf{r}_j\|_2}{\|\mathbf{b}\|_2} < 10^{-7},$$

where $\mathbf{r}_j$ is the residual vector after j iterations.

Example 1. We first consider a Grcar matrix A_n, namely

$$A_n = \begin{bmatrix} 1 & 1 & 1 & 1 & & & \\ -1 & \ddots & \ddots & \ddots & \ddots & & \\ & \ddots & \ddots & \ddots & \ddots & 1 \\ & & \ddots & \ddots & \ddots & 1 \\ & & & \ddots & \ddots & 1 \\ & & & & -1 & 1 \end{bmatrix}.$$

Table 1 (a) and (b) show the numerical results for A_n using MINRES or GMRES. The preconditioners seem efficient for speeding up the rate of convergence of both Krylov subspace methods.

Table 1 Numbers of iterations with (a) MINRES for $Y_n A_n$ and (b) GMRES for A_n with Grcar matrix A_n.

(a)

n	I_n	$\lvert c(A_n)\rvert$
128	49	13
256	49	12
512	49	11
1024	47	11

(b)

n	I_n	$c(A_n)$
128	94	6
256	158	6
512	218	6
1024	213	5

Example 2. We now consider $h(A_n)$ with $h(z) = z^2 + z + 1$ and A_n being a Grcar matrix. Table 2 (a) and (b) show the numerical results for $h(A_n)$. In Figure 1 (a) and (b), we also show the spectra of $Y_n h(A_n)$ before or after applying the preconditioner $|h(c(A_n))|$ when $n = 512$. Again, we observe the clusters of eigenvalues at ± 1.

Table 2 Numbers of iterations with (a) MINRES for $Y_n h(A_n)$ and (b) GMRES for $h(A_n)$ with $h(z) = z^2 + z + 1$ and A_n being a Grcar matrix.

(a)

n	I_n	$\lvert h(c(A_n))\rvert$
128	144	16
256	167	15
512	194	14
1024	190	13

(b)

n	I_n	$h(c(A_n))$
128	128	9
256	256	8
512	512	8
1024	1024	7

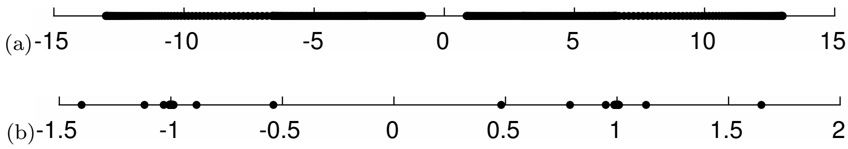

Fig. 1 The spectrum of $Y_n h(A_n)$ when $n = 512$ with $h(z) = z^2 + z + 1$ and A_n being a Grcar matrix under the following condition: (a) without a preconditioner; (b) with the preconditioner $|h(c(A_n))|$.

Example 3. We also consider the system defined by hyperbolic sine function of Toeplitz matrix. Table 3 (a) and (b) show the numerical results for $\sinh A_n$ with A_n being Grcar matrix. The convergence rate appears accelerated with our proposed preconditioners.

Table 3 Numbers of iterations with (a) MINRES for $Y_n \sinh A_n$ and (b) GMRES for $\sinh A_n$ with A_n being a Grcar matrix.

(a)

n	I_n	$\lvert\sinh c(A_n)\rvert$
64	105	24
128	172	22
256	391	20
512	808	19

(b)

n	I_n	$\sinh c(A_n)$
64	64	14
128	124	13
256	240	11
512	486	10

Example 4. The next example is the (real) nonsymmetric Toeplitz matrix generated by $f(\theta) = e^{i\theta} + 2e^{-i\theta}$, namely

$$A_n = \begin{bmatrix} & 2 & & \\ 1 & & \ddots & \\ & \ddots & & 2 \\ & & 1 & \end{bmatrix}.$$

In Figure 2, we still observe that the eigenvalues are highly clustered at ± 1 as expected.

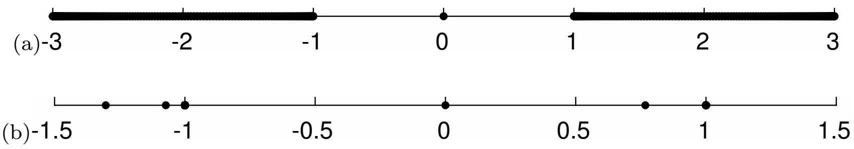

Fig. 2 The spectrum of $Y_n A_n$ with A_n generated by $f(\theta) = e^{i\theta} + 2e^{-i\theta}$ when $n = 512$ under the following condition: (a) without a preconditioner; (b) with the preconditioner $|c(A_n)|$.

Example 5. We consider the matrix exponentials with A_n generated by $f(\theta) = \theta^2$. Table 4 shows the numbers of iterations required with CG for e^{A_n} with or without the preconditioner. It is apparent that the proposed preconditioners are effective for speeding up the rate of convergence of CG.

Table 4 Numbers of iterations with CG for e^{A_n} with A_n generated by $f(\theta) = \theta^2$.

n	I_n	$e^{c(A_n)}$
128	258	9
256	465	8
512	673	8
1024	835	8

In Figure 3, we further show the spectra of the preconditioned matrices with different n. We observe that the highly clustered spectrum is independent of n. In Figure 3 (i) and (ii), the contrast between the spectra of the systems is shown. In Figure 3 (iii), we show the zoom-in spectrum of (ii) and observe that the eigenvalues are highly clustered at 1. Due to the highly clustered eigenvalues of the preconditioned system, a fast convergence rate for CG is expected (see for example [1]).

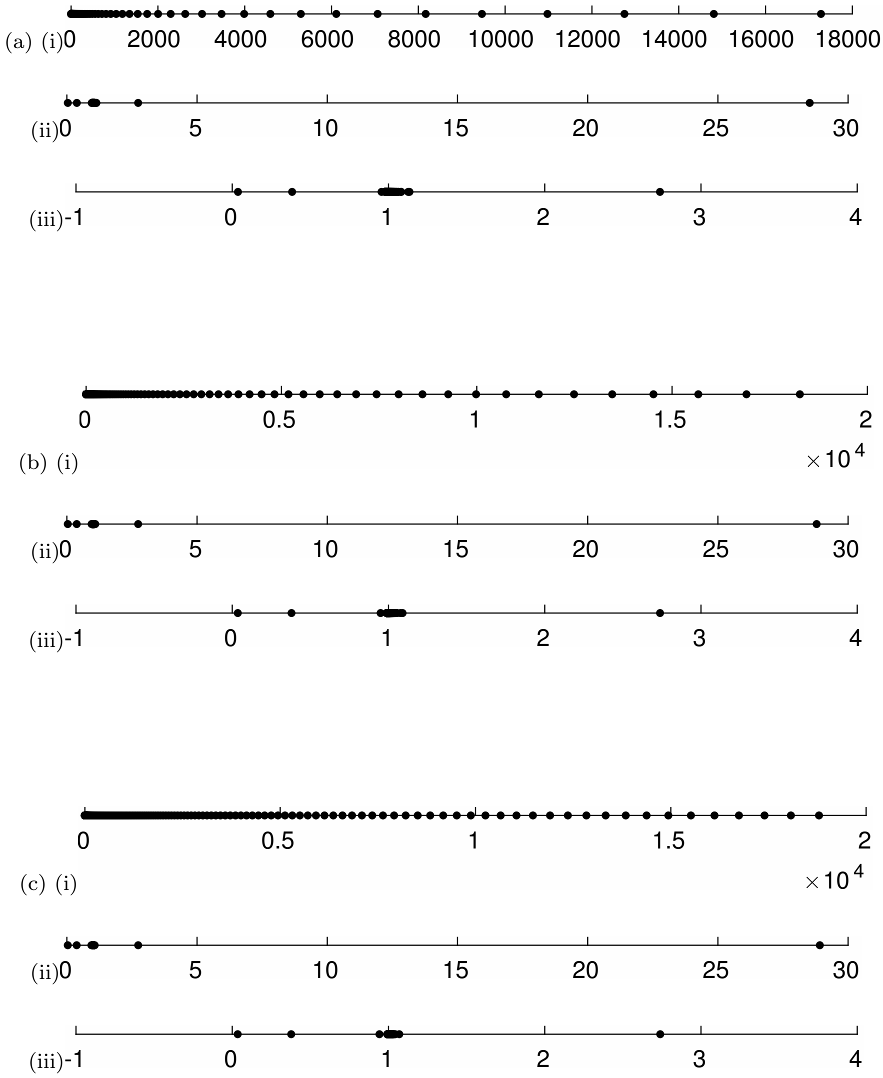

Fig. 3 The spectrum of e^{A_n} with A_n generated by $f(\theta) = \theta^2$ when (a) $n = 128$, (b) $n = 256$ or (c) $n = 512$ under the following condition: (i) without a preconditioner; (ii) with the preconditioner $e^{c(A_n)}$. (iii) The zoom-in spectrum of (ii).

Example 6. Next, we consider symmetric indefinite Toeplitz matrices. Table 5 shows the numerical results for A_n generated by $f(\theta) = \theta^2 - \pi$. In Figure 4 (a) and (b), we also show the spectra of the system before and after applying the preconditioners. We observe the clusters of eigenvalues at ± 1 which are independent of n.

Table 5 Numbers of iterations with (a) MINRES and (b) GMRES for A_n with A_n generated by $f(\theta) = \theta^2 - \pi$.

(a)

n	I_n	$\lvert c(A_n)\rvert$
128	76	11
256	159	11
512	321	10
1024	648	10

(b)

n	I_n	$c(A_n)$
128	74	6
256	150	7
512	299	6
1024	595	6

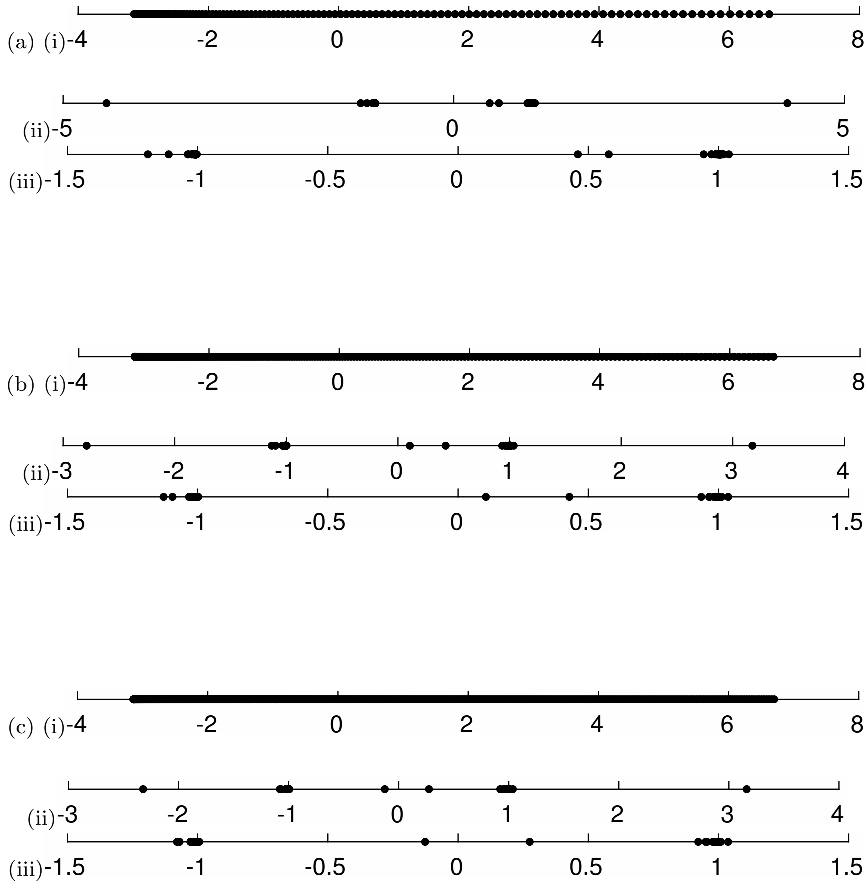

Fig. 4 (i) The spectrum of A_n and that of with A_n is generated by $f(\theta) = \theta^2 - \pi$ when (a) $n = 128$, (b) $n = 256$ or (c) $n = 512$ under the following condition: (i) without a preconditioner; (ii) with the preconditioner $|c(A_n)|$. (iii) The zoom-in spectrum of (ii).

Example 7. In this case, we consider $\cos A_n$ with the symmetric indefinite A_n generated by $f(\theta) = \theta^2 - \pi$. Table 6 shows the numerical results with MINRES for $\cos A_n$ with or without preconditioners $|\cos c(A_n)|$. Again, we observe significant reduce in the numbers of iterations needed for MINRES with the proposed preconditioners used. In Figure 5, we also show the spectra of $|\cos c(A_n)|^{-1} \cos A_n$ when $n = 512$ and observe the clusters of eigenvalues at ± 1.

Table 6 Numbers of iterations with (a) MINRES and (b) GMRES for $\cos A_n$ with A_n generated by $f(\theta) = \theta^2 - \pi$.

(a)

n	I_n	$\|\cos c(A_n)\|$
128	58	24
256	110	24
512	212	24
1024	417	22

(b)

n	I_n	$\cos c(A_n)$
128	58	12
256	110	13
512	212	11
1024	417	11

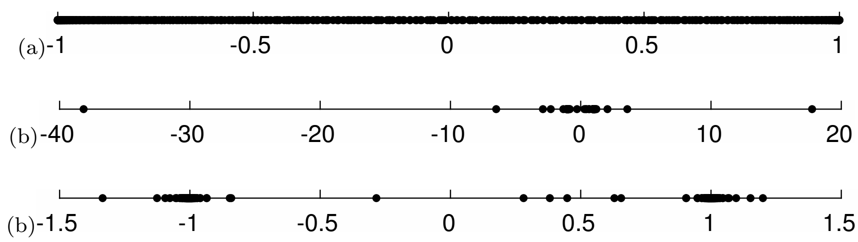

Fig. 5 The spectrum of $\cos A_n$ when $n = 512$ with A_n generated by $f(\theta) = \theta^2 - \pi$ under the following condition: (a) without a preconditioner; (b) with the preconditioner $|\cos c(A_n)|$. (c) The zoom-in spectrum of (b).

We recall that the cost of GMRES increases for every iteration, whereas MINRES has a constant cost per iteration. Thus whilst in many examples, GMRES requires fewer iterations than MINRES, there is not necessarily a reduction in work. The rapid convergence guarantees established here for MINRES are a true indication of the success of our preconditioning approach; it is not perhaps surprising that related preconditioning strategies such as those we have used with GMRES are also successful. It is now widely appreciated that a good preconditioner will lead to work well with efficient iterative methods; indeed that it is the quality of the preconditioner rather than the choice of iterative method which usually determines the success or otherwise of a solver.

References

1. Owe Axelsson and Gunhild Lindskog. On the rate of convergence of the preconditioned conjugate gradient method. *Numerische Mathematik*, 48:499–524, 1986.
2. D.A. Bini, S. Dendievel, G. Latouche, and B. Meini. Computing the exponential of large block-triangular block-Toeplitz matrices encountered in fluid queues. *Linear Algebra and its Applications*, 502(Supplement C):387 – 419, 2016. Structured Matrices: Theory and Applications.
3. Raymond H. Chan and Man-Chung Yeung. Circulant preconditioners for complex Toeplitz matrices. *SIAM Journal on Numerical Analysis*, 30(4):1193–1207, 1993.
4. T. F. Chan. An optimal circulant preconditioner for Toeplitz systems. *SIAM Journal on Scientific and Statistical Computing*, 9(4):766–771, 1988.
5. D.J. Duffy. *Finite Difference Methods in Financial Engineering: A Partial Differential Equation Approach*. The Wiley Finance Series. Wiley, 2013.
6. Nicholas J. Higham. *Functions of matrices*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2008. Theory and computation.
7. Nicholas J. Higham and Peter Kandolf. Computing the action of trigonometric and hyperbolic matrix functions. *SIAM Journal on Scientific Computing*, 39(2):A613–A627, 2017.
8. Sean Hon. Optimal preconditioners for systems defined by functions of Toeplitz matrices. *Under review*.
9. D. Kressner and R. Luce. Fast computation of the matrix exponential for a Toeplitz matrix. *ArXiv e-prints*, July 2016.
10. Spike T. Lee, Hong-Kui Pang, and Hai-Wei Sun. Shift-invert Arnoldi approximation to the Toeplitz matrix exponential. *SIAM Journal on Scientific Computing*, 32(2):774–792, 2010.
11. Eleanor McDonald, Sean Hon, Jennifer Pestana, and Andy Wathen. *Preconditioning for Nonsymmetry and Time-Dependence*, pages 81–91. Springer International Publishing, Cham, 2017.
12. Michael K. Ng. *Iterative methods for Toeplitz systems*. Numerical Mathematics and Scientific Computation. Oxford University Press, New York, 2004.
13. Michael K. Ng and Jianyu Pan. Approximate inverse circulant-plus-diagonal preconditioners for Toeplitz-plus-diagonal matrices. *SIAM Journal on Scientific Computing*, 32(3):1442–1464, 2010.
14. J. Pestana and A. J. Wathen. A preconditioned MINRES method for nonsymmetric Toeplitz matrices. *SIAM Journal on Matrix Analysis and Applications*, 36(1):273–288, 2015.
15. E.W. Sachs and A.K. Strauss. Efficient solution of a partial integro-differential equation in finance. *Applied Numerical Mathematics*, 58(11):1687 – 1703, 2008.