Sine Transform Based Preconditioners For Symmetric Toeplitz Systems

Raymond H. Chan*, Michael K. Ng† and C.K. Wong‡

October 16, 1993
Revised January 27, 1994

Abstract

The optimal circulant preconditioner for a given matrix A is defined to be the minimizer of $||C - A||_F$ over the set of all circulant matrices C. Here $||\cdot||_F$ is the Frobenius norm. Optimal circulant preconditioners have been proved to be good preconditioners in solving Toeplitz systems with the preconditioned conjugate gradient method. In this paper, we construct optimal sine transform based preconditioner which is defined to be the minimizer of $||B - A||_F$ over the set of matrices B that can be diagonalized by sine transforms. We will prove that for general n-by-n matrices A, these optimal preconditioners can be constructed in $O(n^2)$ real operations and in $O(n)$ real operations if A is Toeplitz. We will also show that the convergence properties of these optimal sine transform preconditioners are the same as that of the optimal circulant ones when they are employed to solve Toeplitz systems. Numerical examples are given to support our convergence analysis.

Key Words. Toeplitz matrix, discrete sine transforms, preconditioning.

AMS(MOS) Subject Classifications. 65F10, 65F35, 65F99.

*Department of Mathematics, The Chinese University of Hong Kong, Shatin, Hong Kong. Research supported in part by HKRGC grant no. HKUST 178/93E.

†Department of Mathematics, The Chinese University of Hong Kong, Shatin, Hong Kong.

‡Department of Mathematics, University of Hong Kong, Pokfulam Road, Hong Kong.

1 Introduction

In this paper, we discuss the solutions to a class of symmetric positive definite systems $A\mathbf{x} = \mathbf{b}$ by the preconditioned conjugate gradient (PCG) method. The rate of convergence of the conjugate gradient (CG) method depends on the condition number $\kappa(A)$, see Axelsson and Barker [2, p.26]. In general, the smaller $\kappa(A)$ is, the faster the convergence will be. In case $\kappa(A)$ is not small, the method is always used with a symmetric positive definite matrix M to speed up the convergence rate. More precisely, instead of applying the CG method to the system $A\mathbf{x} = \mathbf{b}$, we apply the method to the transformed system $\hat{A}\hat{\mathbf{x}} = \hat{\mathbf{b}}$ where $\hat{A} = M^{-1/2}AM^{-1/2}$, $\hat{\mathbf{x}} = M^{1/2}\mathbf{x}$ and $\hat{\mathbf{b}} = M^{-1/2}\mathbf{b}$. The matrix M is called a preconditioner for A. The preconditioner M is chosen in the hope that it will minimize $\kappa(M^{-1}A)$ and allow efficient computation of the product $M^{-1}\mathbf{v}$ for any given vector $\mathbf{v}$. The preconditioner M for A can also be viewed as an approximation to A that is easily invertible.

An n-by-n matrix $T = [t_{i,j}]$ is said to be Toeplitz if $t_{i,j}{=}t_{i-j}$, i.e., T is constant along its diagonals. An n-by-n matrix C is said to be circulant if it is Toeplitz and its diagonals c_j satisfy $c_{n-j}{=}c_{-j}$ for $0 < j \leq n-1$. We remark that all circulant matrices C can be diagonalized as

$$C = F^*\Lambda F \tag{1}$$

where $F = \frac{1}{\sqrt{n}}[e^{\frac{2\pi ijk}{n}}]_{j,k=0}^{n-1}$ is the Fourier matrix. Hence, for any vector $\mathbf{v}$, the matrix-vector multiplication $C^{-1/2}\mathbf{v} = F^*\Lambda^{-1/2}F\mathbf{v}$ can be computed efficiently by the fast Fourier transform (FFT) in $O(n \log n)$ operations, see Bergland [3].

Since circulant matrices are Toeplitz matrices themselves, it is natural to consider using circulant matrices as preconditioners for Toeplitz systems. Given a Toeplitz matrix T, there are many possible circulant matrices C that one can define to be preconditioners for the system $T\mathbf{x} = \mathbf{b}$. Since the convergence rate of the PCG method depends on how good the preconditioner C approximates T, much attention has been focused on searching a circulant matrix C which is close to the matrix T in certain norms, see T. Chan [6], Tyrtyshnikov [22] and Huckle [18]. T. Chan in [6] proposed a circulant preconditioner $c(T)$ which is the minimizer of $||C-T||_F$ over all circulant matrices C. Here $||\cdot||_F$ denotes the Frobenius norm. He called $c(T)$ the optimal circulant preconditioner and showed that the first column entries c_j of $c(T)$ are given by

$$c_j = \frac{jt_{-(n-j)} + (n-j)t_j}{n}, \qquad j = 0, 1, ..., n-1,$$

where t_j are the diagonals of T.

It was shown in Chan [7] that if the underlying generating function of T is a positive function in the Wiener class, then the spectrum of $c(T)^{-1}T$ is clustered around 1. Tyrtyshnikov in [22] extended the definition of $c(\cdot)$ to any general n-by-n matrix A. Also, he

proved that $c(A)$ is symmetric positive definite whenever A is. Note that forming $c(A)$ only requires $O(n)$ operations for Toeplitz matrix A of order n, and $O(n^2)$ operations for general n-by-n matrix A. However, we remark that when A are tridiagonal Toeplitz matrix or tridiagonal block Toeplitz matrix, such as the 1-dimensional and 2-dimensional discrete Laplacian $P = \text{tridiag}[-1, 2, -1]$ and $P \otimes I + I \otimes P$ respectively, the performance of the optimal circulant preconditioners $c(A)$ are not very good, see R. Chan and T. Chan [10].

The purpose of this paper is to construct optimal sine transform based preconditioners $s(A)$ for general matrices A. They are defined to be the minimizer of $||B - A||_F$ over the set of matrices B that can be diagonalized by the sine transform matrix S. Since only sine transforms will be involved, all computations can be done in real arithmetic. We remark that the matrix-vector product $S\mathbf{v}$ can be done in $O(n \log n)$ real operations, see for instance Yip and Rao [23]. Moreover, we will see that if A is a tridiagonal Toeplitz matrix, $s(A)$ is just equal to A itself.

We note that since the Frobenius norm is a unitary-invariant norm, the minimizer $s(A)$ is given by $S\Delta S$, where Δ is a diagonal matrix with diagonal entries

$$\Delta_{j,j} = (SAS)_{j,j}, \quad j = 1, \ldots, n, \tag{2}$$

see for instance Huckle [18]. However, computing all the diagonal entries of Δ using formula (2) will require $O(n^2 \log n)$ operations. In this paper, we will show that the minimizer $s(A)$ can be obtained in $O(n^2)$ operations for general matrix A. The cost can even be reduced to $O(n)$ operations when A is a Toeplitz matrix.

We remark that these operation counts are the same as that of obtaining optimal circulant preconditioners $c(A)$, see T. Chan [6]. However, we emphasize that to construct $c(A)$ economically, T. Chan has used the fact that all matrices that can be diagonalized by Fourier matrix are circulant matrices which are matrices having very nice algebraic structures. Thus in order to construct $s(A)$ efficiently, one needs to find matrices having special algebraic structures to characterize all matrices that can be diagonalized by sine transforms. Recently, Boman and Koltracht [5], Bini and Benedetto [4] and Huckle [19] independently showed that matrices that can be diagonalized by sine transforms can be written as a sum of a Toeplitz matrix and a Hankel matrix. This decomposition is the crucial step that leads us to a fast algorithm for obtaining $s(A)$.

As for how good optimal sine transform based preconditioners $s(T)$ are as preconditioners for Toeplitz systems $T\mathbf{x} = \mathbf{b}$, we will show that they have the same convergence properties as the optimal circulant preconditioners $c(T)$. More precisely, we will show that if a given Toeplitz matrix T is generated by a 2π-periodic positive continuous function, then the spectrum of $s(T)^{-1}T$ is clustered around 1.

The outline of the paper is as follows. In the next section, we will exhibit a basis for the set of matrices that can be diagonalized by sine transforms. The basis is first obtained

by Boman and Koltracht [5]. Using this basis, we can then construct the optimal sine transform based preconditioner $s(A)$ for any given matrix A. We will prove that the construction of such preconditioners is of $O(n^2)$ operations for general matrices and the count reduces to $O(n)$ operations when A is a Toeplitz matrix. We show that $s(A)$ is positive definite when A is positive definite. We also show that if SAS has Property A, then $s(A)$ is the best conditioned sine transform based preconditioner, i.e.

$$\kappa(s(A)^{-1/2} A s(A)^{-1/2}) \leq \kappa(B^{-1/2} A B^{-1/2})$$

for any matrices B that can be diagonalized by the sine transform matrix S. In §3, we will give the convergence analysis of the optimal sine transform based preconditioners when they are applied to solve symmetric Toeplitz systems. Finally, numerical results and some concluding remarks are given in §4.

2 Optimal Discrete Sine Transform Preconditioner

Let S_n be the n-by-n discrete sine transform matrix with the (i,j)th entry given by

$$\sqrt{\frac{2}{n+1}} \sin(\frac{\pi i j}{n+1}), \quad 1 \leq i, j \leq n. \tag{3}$$

We note that S_n are symmetric and orthogonal, i.e. $S_n = S_n^t$ and $S_n S_n^t = I_n$. Also, for any n-vector $\mathbf{v}$, the matrix-vector multiplication $S_n \mathbf{v}$ can be computed in $O(n \log n)$ real operations, ($(n/2)\log n - n + 1$ multiplications and $2n \log n - 4n + 4$ additions), see Yip and Rao [23]. In contrast, the numbers of real multiplications and real additions required in n-dimensional fast Fourier transform (FFT) are $n \log n - 3n + 4$ and $(3n/2)\log n - (5n/2) + 4$ respectively, see Bergland [3]. The number of operations required for the fast sine transform (FST), are almost the same as that of FFT. In this paper, we consider solving linear systems by the PCG method with preconditioners that can be diagonalized by S_n. Let $\mathbf{B}_{n\times n}$ be the vector space containing all such matrices. More precisely, we let

$$\mathbf{B}_{n\times n} = \{S_n \Lambda_n S_n \mid \Lambda_n \text{ is an } n\text{−by−}n \text{ diagonal matrix}\}.$$

Recently, Boman and Koltracht [5], Bini and Benedetto [4] and Huckle [19] independently proved that a matrix belongs to $\mathbf{B}_{n\times n}$ if and only if the matrix can be expressed as a special sum of a Toeplitz matrix and a Hankel matrix. We recall that a matrix $A = [a_{i,j}]$ is said to be Toeplitz if $a_{i,j} = a_{i-j}$ and Hankel if $a_{i,j} = a_{i+j}$. The idea of their proof is to exhibit a basis for $\mathbf{B}_{n\times n}$ with each element in the basis being a sparse matrix and possessing a nice structure. The following Lemma gives the basis Boman and Koltracht considered.

Lemma 1 (Boman and Koltracht [5]) *Let Q_i, $i = 1, \ldots, n$, be n-by-n matrices with the (h,k)th entry given by*

$$Q_i(h,k) = \begin{cases} 1 & \text{if } |h-k| = i-1, \\ -1 & \text{if } h+k = i-2, \\ -1 & \text{if } h+k = 2n-i+3, \\ 0 & \text{otherwise.} \end{cases}$$

Then $\{Q_i\}_{i=1}^n$ is a basis for $\mathbf{B}_{n\times n}$.

To illustrate the sparsity and nice structure of Q_i, we display the basis for the case $n = 6$.

$$Q_1 = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}, \quad Q_2 = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix},$$

$$Q_3 = \begin{pmatrix} -1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & -1 \end{pmatrix}, \quad Q_4 = \begin{pmatrix} 0 & -1 & 0 & 1 & 0 & 0 \\ -1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & -1 \\ 0 & 0 & 1 & 0 & -1 & 0 \end{pmatrix},$$

$$Q_5 = \begin{pmatrix} 0 & 0 & -1 & 0 & 1 & 0 \\ 0 & -1 & 0 & 0 & 0 & 1 \\ -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -1 \\ 1 & 0 & 0 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 & 0 & 0 \end{pmatrix}, \quad Q_6 = \begin{pmatrix} 0 & 0 & 0 & -1 & 0 & 1 \\ 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 & -1 \\ -1 & 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 & 0 & 0 \\ 1 & 0 & -1 & 0 & 0 & 0 \end{pmatrix}. \tag{4}$$

In order to give a precise description of the vector space $\mathbf{B}_{n\times n}$, let us introduce the following notations.

Definition 1 *Let $\mathbf{z} = (z_1, \ldots, z_n)^t$ be an n-vector. Define*

$$\tilde{\mathbf{z}} \equiv (z_n, \ldots, z_1)^t$$

to be the vector with components of $\mathbf{z}$ listed in reverse order. Also, define

$$\sigma(\mathbf{z}) \equiv (z_3, \ldots, z_n, 0, 0)^t$$

to be the n-vector which is obtained by upper shifting of $\mathbf{z}$ by two entries.

Definition 2 *Let* $\mathbf{z}$ *be an* n*-vector. Define* $\mathcal{T}_n(\mathbf{z})$ *to be the* n*-by-*n *symmetric Toeplitz matrix with* $\mathbf{z}$ *as the first column and* $\mathcal{H}_n(\mathbf{z})$ *to be the* n*-by-*n *Hankel matrix with* $\mathbf{z}$ *as the first column and* $\tilde{\mathbf{z}}$ *as the last column.*

With the above notations, we are going to identify the vector space $\mathbf{B}_{n\times n}$.

Lemma 2 $\mathbf{B}_{n\times n} = \{\mathcal{T}_n(\mathbf{z}) - \mathcal{H}_n(\sigma(\mathbf{z})) \mid \mathbf{z} = (z_1, \ldots, z_n)^t \in \mathbf{R}^n\}$.

Proof: Let $\mathbf{e}_i$ be the ith unit vector in $\mathbf{R}^n$. Then by Lemma 1, Q_i can be rewritten as

$$Q_i = \mathcal{T}_n(\mathbf{e}_i) - \mathcal{H}_n(\sigma(\mathbf{e}_i)).$$

Therefore, an n-by-n matrix B_n belongs to $\mathbf{B}_{n\times n}$ if and only if there exist $z_1, \ldots, z_n \in \mathcal{R}$ such that

$$\begin{aligned} B_n = \sum_{j=1}^{n} z_j Q_j &= \sum_{j=1}^{n} z_j[\mathcal{T}_n(\mathbf{e}_j) - \mathcal{H}_n(\sigma(\mathbf{e}_j))] \\ &= \mathcal{T}_n(\sum_{j=1}^{n} z_j\mathbf{e}_j) - \mathcal{H}_n(\sigma(\sum_{j=1}^{n} z_j\mathbf{e}_j)) \\ &= \mathcal{T}_n(\mathbf{z}) - \mathcal{H}_n(\sigma(\mathbf{z})) \end{aligned}$$

with $\mathbf{z} = \sum_{j=1}^{n} z_j\mathbf{e}_j$. □

For any Toeplitz matrix T_n with $\mathbf{t} = (t_0, \ldots, t_{n-1})^t$ as the first column, Boman and Koltracht [5] recently considered using the matrices $K_n = \mathcal{T}_n(\mathbf{t}) - \mathcal{H}_n(\sigma(\mathbf{t}))$ as preconditioners for solving symmetric Toeplitz systems $T_n\mathbf{x} = \mathbf{b}$. We remark from Lemma 2 that K_n can be diagonalized by the sine transform matrix S_n. In [5], the preconditioner K_n is shown to be positive definite whenever T_n is. Also, if t_j are Fourier coefficients of a positive function in the Wiener class (i.e. $\sum_{j=0}^{\infty} |t_j| < \infty$), then the conjugate gradient method applied to the preconditioned matrix $K_n^{-1}T_n$ has a superlinear convergence rate. We note that if T_n is a tridiagonal Toeplitz matrix, then $\mathcal{H}_n(\sigma(\mathbf{t}))$, the Hankel part of K_n, is a zero matrix and hence the preconditioner K_n is equal to the matrix T_n itself. It follows that tridiagonal Toeplitz systems can be solved in one iteration by the PCG method with preconditioners K_n.

However, we remark that their approach of constructing sine transform based preconditioners for Toeplitz matrices cannot be extended to general symmetric matrices. In this paper, we are going to proposal sine transform based preconditioners $s(A_n)$ that are defined for any n-by-n symmetric matrices A_n. Furthermore, if A_n is a symmetric tridiagonal Toeplitz matrix, then our $s(A_n)$ is also equal to A_n itself.

Since preconditioners can be viewed as approximations to the given matrix A_n, it is reasonable to consider preconditioners which minimize $||B_n - A_n||$ over all $B_n \in \mathbf{B}_{n\times n}$ for some matrix norm $||\cdot||$. We choose our preconditioner $s(A_n)$ to be the minimizer of $||B_n - A_n||_F$ in the Frobenius norm. According to the terminology used in T. Chan [6], we call $s(A_n)$ *the optimal sine transform based preconditioner*. We will show that $s(A_n)$ can be obtained in $O(n^2)$ operations for general matrix. The cost can even be reduced to $O(n)$ operations when A_n is a Toeplitz matrix. We remark that the cost of constructing $s(A_n)$ is the same as that of optimal circulant preconditioner $c(A_n)$.

For the sake of presentation, let us illustrate the procedure of constructing $s(A_n)$ by considering the simple case $n = 6$. By Lemma 2, $s(A_6)$ is of the following form:

$$s(A_6) = \begin{pmatrix} z_1 & z_2 & z_3 & z_4 & z_5 & z_6 \\ z_2 & z_1 & z_2 & z_3 & z_4 & z_5 \\ z_3 & z_2 & z_1 & z_2 & z_3 & z_4 \\ z_4 & z_3 & z_2 & z_1 & z_2 & z_3 \\ z_5 & z_4 & z_3 & z_2 & z_1 & z_2 \\ z_6 & z_5 & z_4 & z_3 & z_2 & z_1 \end{pmatrix} - \begin{pmatrix} z_3 & z_4 & z_5 & z_6 & 0 & 0 \\ z_4 & z_5 & z_6 & 0 & 0 & 0 \\ z_5 & z_6 & 0 & 0 & 0 & z_6 \\ z_6 & 0 & 0 & 0 & z_6 & z_5 \\ 0 & 0 & 0 & z_6 & z_5 & z_4 \\ 0 & 0 & z_6 & z_5 & z_4 & z_3 \end{pmatrix},$$

where $\mathbf{z} = (z_1, z_2, z_3, z_4, z_5, z_6)^t$ is the unknown vector to be found. Let a_{ij} be the (i,j)th entry of A_6. Minimizing $\|s(A_6) - A_6\|_F^2$ by setting

$$\frac{\partial}{\partial z_i}\|s(A_6) - A_6\|_F^2 = 0, \qquad \text{for } i = 1, \ldots, 6,$$

we see that $\mathbf{z}$ satisfies the following linear system

$$\begin{pmatrix} 6 & 0 & -2 & 0 & -2 & 0 \\ 0 & 10 & 0 & -4 & 0 & -4 \\ -2 & 0 & 10 & 0 & -4 & 0 \\ 0 & -4 & 0 & 10 & 0 & -4 \\ -2 & 0 & -4 & 0 & 10 & 0 \\ 0 & -4 & 0 & -4 & 0 & 10 \end{pmatrix} \begin{pmatrix} z_1 \\ z_2 \\ z_3 \\ z_4 \\ z_5 \\ z_6 \end{pmatrix}$$

$$= \begin{pmatrix} a_{11} + a_{22} + a_{33} + a_{44} + a_{55} + a_{66} \\ a_{12} + a_{23} + a_{34} + a_{45} + a_{56} + a_{21} + a_{32} + a_{43} + a_{54} + a_{65} \\ a_{13} + a_{24} + a_{35} + a_{46} + a_{31} + a_{42} + a_{53} + a_{64} - a_{11} - a_{66} \\ a_{14} + a_{25} + a_{36} + a_{41} + a_{52} + a_{63} - a_{12} - a_{21} - a_{56} - a_{65} \\ a_{15} + a_{26} + a_{51} + a_{62} - a_{13} - a_{22} - a_{31} - a_{46} - a_{55} - a_{64} \\ a_{16} + a_{61} - a_{14} - a_{23} - a_{32} - a_{41} - a_{36} - a_{45} - a_{54} - a_{63} \end{pmatrix}. \tag{5}$$

We observe that the ith entry of the right hand side vector in (5) is obtained by adding or subtracting those a_{hk} for which the (h,k)th position of Q_i is 1 or -1 respectively (c.f. (4)).

For general n, if we let $\mathbf{1}_n$ be the n-vector with all entries being one and $\circ$ be the Hadamard product, then a straightforward computation as the one we did above shows that the right hand side vector is given by

$$\mathbf{r}_n = (\mathbf{1}_n^t(Q_1 \circ A_n)\mathbf{1}_n, \mathbf{1}_n^t(Q_2 \circ A_n)\mathbf{1}_n, \ldots, \mathbf{1}_n^t(Q_n \circ A_n)\mathbf{1}_n)^t. \tag{6}$$

If A_n has no special structure, then clearly, $\mathbf{r}_n = (r_1, \ldots, r_n)^t$ can be computed in $O(n^2)$ operations because Q_i are sparse with only $O(n)$ nonzero entries each. We note however that if A_n is a Toeplitz matrix with first row $(t_0, t_1, \ldots, t_{n-1})$, then $\mathbf{r}_n$ can be obtained in $O(n)$ operations. This can be seen from the following algorithm when n is even. For odd n, similar algorithm can be derived.

Algorithm 1:

$r_1 = nt_0$
$r_2 = 2(n-1)t_1$
$w_1 = -t_0$
$v_1 = -2t_1$
for $k = 2 : \frac{n}{2}$
  $r_{2k-1} = 2(n-2k+2)t_{2k-2} + 2w_{k-1}$
  $w_k = w_{k-1} - 2t_{2k-2}$
  $r_{2k} = 2(n-2k+1)t_{2k-1} + 2v_{k-1}$
  $v_k = v_{k-1} - 2t_{2k-1}$
end

We now go back to the solution of the linear system (5). We first reorder the unknowns z_i of $\mathbf{z}$ in such a way that the odd index entries and even index entries appear respectively in the upper half and lower half of the resulting vector. For simplicity, this leads to the following definition.

Definition 3 *Let P_n be the n-by-n permutation matrix with the (i,j)th entry given by*

$$[P_n]_{i,j} = \begin{cases} 1 & \text{if } 1 \le i \le \lceil \frac{n}{2} \rceil \text{ and } j = 2i-1, \\ 1 & \text{if } \lceil \frac{n}{2} \rceil < i \le n \text{ and } j = 2i - 2\lceil \frac{n}{2} \rceil, \\ 0 & \text{otherwise.} \end{cases}$$

In particular, P_6 is given by

$$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

After permutation, (5) becomes a block diagonal system,

$$\begin{pmatrix} 6 & -2 & -2 & 0 & 0 & 0 \\ -2 & 10 & -4 & 0 & 0 & 0 \\ -2 & -4 & 10 & 0 & 0 & 0 \\ 0 & 0 & 0 & 10 & -4 & -4 \\ 0 & 0 & 0 & -4 & 10 & -4 \\ 0 & 0 & 0 & -4 & -4 & 10 \end{pmatrix} \begin{pmatrix} z_1 \\ z_3 \\ z_5 \\ z_2 \\ z_4 \\ z_6 \end{pmatrix} = P_6 \mathbf{r}_6. \tag{7}$$

The following theorem proves that in general if $\mathbf{r}_n$ is known in advance, then solving the block diagonal matrix can be done in $O(n)$ operations.

Theorem 1 *Let $A_n = [a_{jk}]$ be an n-by-n symmetric matrix and $s(A_n)$ be the minimizer of $\|B_n - A_n\|_F$ over all $B_n \in \mathbf{B}_{n\times n}$. Denote U_m to be the m-by-m matrix with all its entries being one, and $\mathbf{e}_1$ to be the first unit vector of length $\lceil \frac{n}{2} \rceil$. Then*

$$s(A_n) = \mathcal{T}_n(\mathbf{z}) - \mathcal{H}_n(\sigma(\mathbf{z}))$$

with

$$\mathbf{z} = \frac{1}{2(n+1)} P_n^t \begin{pmatrix} U_{\frac{n}{2}} + I_{\frac{n}{2}} + \mathbf{e}_1\mathbf{e}_1^t & 0 \\ 0 & 2U_{\frac{n}{2}} + I_{\frac{n}{2}} \end{pmatrix} P_n \mathbf{r}_n \tag{8}$$

if n is even; and

$$\mathbf{z} = \frac{1}{2(n+1)} P_n^t \begin{pmatrix} 2U_{\frac{n+1}{2}} + I_{\frac{n+1}{2}} + \mathbf{e}_1\mathbf{e}_1^t & 0 \\ 0 & U_{\frac{n-1}{2}} + I_{\frac{n-1}{2}} \end{pmatrix} P_n \mathbf{r}_n \tag{9}$$

if n is odd.

Proof: Here we just give the proof for the case n is even. The proof for odd n is similar. To minimize $\|B_n - A_n\|_F^2$ over $\mathbf{B}_{n\times n}$, we set

$$\frac{\partial}{\partial z_i} \|s(A_n) - A_n\|_F^2 = 0, \qquad \text{for } i = 1, \ldots, n.$$

We obtain a linear system that has the same structure as that in (5). Permutating the system by P_n yields

$$\begin{pmatrix} DKD + \frac{n+1}{2}\mathbf{e}_1\mathbf{e}_1^t & 0 \\ 0 & K \end{pmatrix} P_n \mathbf{z} = P_n \mathbf{r}_n.$$

Here K is an $\frac{n}{2}$-by-$\frac{n}{2}$ Toeplitz matrix given by

$$K = \mathcal{T}_{\frac{n}{2}}([2(n-1), -4, -4, \ldots, -4]^t)$$

and $D = \mathrm{diag}(\frac{1}{2}, 1, 1, \ldots, 1)$ is an $\frac{n}{2}$-by-$\frac{n}{2}$ diagonal matrix (c.f. (7)). Note that K can be rewritten as

$$K = 2(n+1)I_{\frac{n}{2}} - 4U_{\frac{n}{2}}. \tag{10}$$

Applying Sherman-Morrison formula, see [17, p.3], we can express K^{-1} as,

$$K^{-1} = \frac{1}{n+1}U_{\frac{n}{2}} + \frac{1}{2(n+1)}I_{\frac{n}{2}}.$$

Similarly by rewriting

$$DKD + \frac{n+1}{2}\mathbf{e}_1\mathbf{e}_1^t = 2(n+1)I_{\frac{n}{2}} - (2\mathbf{1}_{\frac{n}{2}} - \mathbf{e}_1)(2\mathbf{1}_{\frac{n}{2}} - \mathbf{e}_1)^t + (n+1)\mathbf{e}_1\mathbf{e}_1^t$$

and applying Sherman-Morrison formula we have

$$\frac{1}{2(n+1)}(DKD + \frac{n+1}{2}\mathbf{e}_1\mathbf{e}_1^t)(U_{\frac{n}{2}} + I_{\frac{n}{2}} + \mathbf{e}_1\mathbf{e}_1^t) = I_{\frac{n}{2}}.$$

Combining these together with the fact that P_n is orthogonal, (8) follows. □

Before going on, let us first emphasize the relationship between the first column of matrices $B \in \mathbf{B}_{n\times n}$ and their eigenvalues. For any matrix $B \in \mathbf{B}_{n\times n}$, we have $B = S\Lambda S$ where Λ is the eigenvalue matrix of B. If D denotes the diagonal matrix whose diagonal is equal to the first column of S_n, then we have $S\mathbf{e}_1 = D\mathbf{1}_n$. Therefore the relation is given by

$$D^{-1}S_nB\mathbf{e}_1 = \Lambda\mathbf{1}_n. \tag{11}$$

Hence, any matrix in $\mathbf{B}_{n\times n}$ is determined by its first column. In particular, eigenvalues of the minimizer $s(A_n)$ can be computed in $O(n\log n)$ operations. The following corollary gives the explicit formula for the entries of the first column of $s(A_n)$. The proof follows directly from the expressions (8) and (9) and therefore we omit it.

Corollary 1 *Let $A_n = [a_{jk}]$ be an n-by-n symmetric matrix and $s(A_n)$ be the minimizer of $\|B_n - A_n\|_F$ over all $B_n \in \mathbf{B}_{n\times n}$. Denote $\mathbf{z}$ to be the first column of $s(A_n)$. If s_o and s_e are defined respectively to be the sum of the odd and even index entries of $\mathbf{r}_n$, then we have*

$$\begin{aligned}
[\mathbf{z}]_1 &= \frac{1}{2(n+1)}(2[\mathbf{r}_n]_1 - [\mathbf{r}_n]_3) \\
[\mathbf{z}]_i &= \frac{1}{2(n+1)}([\mathbf{r}_n]_i - [\mathbf{r}_n]_{i+2}) \qquad i = 2, \ldots, n-2
\end{aligned}$$

with

$$\begin{aligned}[\mathbf{z}]_{n-1} &= \frac{1}{2(n+1)}(s_o + [\mathbf{r}_n]_{n-1}) \\ [\mathbf{z}]_n &= \frac{1}{2(n+1)}(2s_e + [\mathbf{r}_n]_n)\end{aligned}$$

if n is even; and

$$\begin{aligned}[\mathbf{z}]_{n-1} &= \frac{1}{2(n+1)}(s_e + [\mathbf{r}_n]_n) \\ [\mathbf{z}]_n &= \frac{1}{2(n+1)}(2s_o + [\mathbf{r}_n]_n)\end{aligned}$$

if n is odd.

From Corollary 1 and (6), we see that $s(A_n)$ can be obtained in $O(n^2)$ operations for general symmetric matrix A_n and $O(n)$ operations for band matrix A_n. Using Algorithm 1, we further see that only $O(n)$ operations is required if A_n is a symmetric Toeplitz matrix. In the following, we give some spectral properties of $s(A_n)$.

Theorem 2 *Let A_n be an n-by-n symmetric matrix. Then $s(A_n)$ is symmetric. Moreover, we have*

$$\lambda_{\min}(A_n) \leq \lambda_{\min}(s(A_n)) \leq \lambda_{\max}(s(A_n)) \leq \lambda_{\max}(A_n) \ , \tag{12}$$

where $\lambda_{\max}(\cdot)$ and $\lambda_{\min}(\cdot)$ denote the largest and the smallest eigenvalues respectively. In particular,

$$||s(A_n)||_2 \leq ||A_n||_2 \tag{13}$$

and if A_n is positive definite, then $s(A_n)$ is also positive definite.

Proof: The proof is similar to that of Theorem 1 in Chan, Jin and Yeung [9] or that of Theorem 2 in Huckle [18]. □

Next we consider matrices A_n having Property A, i.e. there exists a permutation matrix P such that

$$PA_nP^t = \begin{pmatrix} D_1 & E_2 \\ E_1 & D_2 \end{pmatrix},$$

where D_1 and D_2 are square diagonal matrices and E_1 and E_2 are arbitrary matrices. In [11], Chan and Wong proved that if FA_nF^* has Property A, then the optimal circulant preconditioner $c(A_n)$ minimizes the condition number $\kappa(C^{-1/2}A_nC^{-1/2})$ over all positive definite circulant matrices C. Here similarly if $S_nA_nS_n$ has Property A, then we can prove that $s(A_n)$ minimizes $\kappa(B^{-1/2}A_nB^{-1/2})$ over all positive definite $B \in \mathbf{B}_{n\times n}$. The proof of the following theorem is similar to that of Theorem 1 in [11] and therefore will be omitted.

Theorem 3 *Let A_n be an n-by-n symmetric positive definite matrix. If the matrix $S_n A_n S_n$ has Property A, then $s(A_n)$ minimizes $\kappa(B^{-1/2}AB^{-1/2})$ over all symmetric positive definite matrices $B \in \mathbf{B}_{n\times n}$.*

3 Application in Solving Toeplitz Systems

In this section, we consider applying the optimal sine transform based preconditioners $s(T_n)$ to solving a class of symmetric Toeplitz systems $T_n\mathbf{x} = \mathbf{b}$ by the preconditioned conjugate gradient method. Our main result is that the spectra of these preconditioned matrices $s(T_n)^{-1}T_n$ are clustered around 1. Hence the conjugate gradient method when applied to solving the preconditioned systems $s(T_n)^{-1}T_n\mathbf{x} = s(T_n)^{-1}\mathbf{b}$ converges sufficiently fast.

In the following, we assume that the Toeplitz matrices T_n are generated by 2π-periodic continuous real-valued even functions. We emphasize that this class of symmetric Toeplitz matrices arises in some practical problems. Typical examples of generating functions are the kernels of the Wiener-Hopf equations, see Gohberg and Fel'dman [15, p.82], the function which gives amplitude characteristic of the recursive digital filters, see Chui and Chan [14], the spectral density functions in stationary stochastic process , see Grenander and Szegö [16, p.171] and the point-spread functions in image deblurring, see Oppenheim [20, p.200]. In the following discussions, we denote $\mathbf{C}_{2\pi}$ to be the set of 2π-periodic continuous real-valued even functions. For all $f \in \mathbf{C}_{2\pi}$, let

$$t_k(f) = \frac{1}{2\pi}\int_0^{2\pi} f(\theta)e^{-ik\theta}d\theta, \quad k = 0, \pm 1, \pm 2, \cdots$$

be the Fourier coefficients of f. Since f is even and real-valued, we have

$$t_k(f) = t_{-k}(f), \quad k = 0, \pm 1, \pm 2, \cdots.$$

For simplicity, we write $t_k(f) = t_k$. We also let $\mathbf{f}_n$ to be an n-vector with entries given by the first n Fourier coefficients of f, i.e.

$$\mathbf{f}_n = (t_0, t_1, t_2, \ldots, t_{n-1})^t.$$

The following Lemma gives the relation between f and the spectrum $\lambda(\mathcal{T}_n(\mathbf{f}_n))$ of $\mathcal{T}_n(\mathbf{f}_n)$.

Lemma 3 (Grenander and Szegö [16, pp.63-65]) *Let $f \in \mathbf{C}_{2\pi}$ with the minimum and maximum values given by $f_{\min}$ and $f_{\max}$ respectively. Then $\lambda(\mathcal{T}_n(\mathbf{f}_n)) \subseteq [f_{\min}, f_{\max}]$. In particular, we have*

$$||\mathcal{T}_n(\mathbf{f}_n)||_2 \leq ||f||_\infty$$

where $||\cdot||_\infty$ denotes the supremum norm.

In the following, we first prove that if f is an even function in the Wiener class, then the spectrum of $\mathcal{T}_n(\mathbf{f}_n) - s(\mathcal{T}_n(\mathbf{f}_n))$ is clustered around zero. Then we extend the clustering result from the Wiener class to $\mathbf{C}_{2\pi}$. We remark that a function f is in the Wiener class if its Fourier coefficients are absolutely summable, i.e.

$$\sum_{k=-\infty}^{\infty} |t_k| < \infty.$$

It is clear that if f is an even function in the Wiener class, then $f \in \mathbf{C}_{2\pi}$.

In the analysis of the spectra of the preconditioned matrices, we first write $\mathcal{T}_n(\mathbf{f}_n) - s(\mathcal{T}_n(\mathbf{f}_n))$ as

$$\mathcal{T}_n(\mathbf{f}_n) - s(\mathcal{T}_n(\mathbf{f}_n)) = \mathcal{H}_n(\sigma(\mathbf{f}_n)) + \mathcal{T}_n(\mathbf{f}_n) - \mathcal{H}_n(\sigma(\mathbf{f}_n)) - s(\mathcal{T}_n(\mathbf{f}_n)). \tag{14}$$

The clustering of the spectrum of $\mathcal{H}_n(\sigma(\mathbf{f}_n))$ has already been proved by Boman and Koltracht [5].

Lemma 4 (**Boman and Koltracht [5]**) *Let f be an even function in the Wiener class. Then for all $\epsilon > 0$, there exist $N, M > 0$ such that for all $n > N$, at most M eigenvalues of $\mathcal{H}_n(\sigma(\mathbf{f}_n))$ have absolute value larger than ϵ.*

According to this lemma and equation (14), it suffices to show that the spectra of $\mathcal{T}_n(\mathbf{f}_n) - \mathcal{H}_n(\sigma(\mathbf{f}_n))$ and $s(\mathcal{T}_n(\mathbf{f}_n))$ are asymptotically the same.

Lemma 5 *Let f be an even function in the Wiener class, then*

$$\lim_{n\to\infty} ||s(\mathcal{T}_n(\mathbf{f}_n)) - [\mathcal{T}_n(\mathbf{f}_n) - \mathcal{H}_n(\sigma(\mathbf{f}_n))]||_2 = 0.$$

Proof: For simplicity, we only consider the case where $n = 2m$. The case where n is odd can be proved similarly. We first note from Theorem 1 and Corollary 1 that our optimal sine transform based preconditioner $s(\mathcal{T}_n(\mathbf{f}_n))$ can be expressed as follows:

$$s(\mathcal{T}_n(\mathbf{f}_n)) = \mathcal{T}_n(\mathbf{z}_n) - \mathcal{H}_n(\sigma(\mathbf{z}_n)).$$

Here the kth entry of the n-vector $\mathbf{z}_n$ is given by

$$[\mathbf{z}_n]_k = \begin{cases} t_0 + \dfrac{2}{n+1}\displaystyle\sum_{j=(k+1)/2}^{m-1} t_{2j}, & k = 1, \\ t_1 + \dfrac{2}{n+1}\displaystyle\sum_{j=k/2}^{m-1} t_{2j+1}, & k = 2, \\ \left(\dfrac{n-k+3}{n+1}\right) t_{k-1} + \dfrac{2}{n+1}\displaystyle\sum_{j=(k+1)/2}^{m-1} t_{2j}, & 3 \le k \le n \text{ and where } k \text{ is odd}, \\ \left(\dfrac{n-k+3}{n+1}\right) t_{k-1} + \dfrac{2}{n+1}\displaystyle\sum_{j=k/2}^{m-1} t_{2j}, & 3 \le k \le n \text{ and where } k \text{ is even}. \end{cases}$$

It is clear that $\mathcal{T}_n(\mathbf{f}_n) - \mathcal{H}_n(\sigma(\mathbf{f}_n)) - s(\mathcal{T}_n(\mathbf{f}_n))$ is a symmetric Toeplitz-plus-Hankel matrix. After some manipulations, it can be re-written as

$$\mathcal{T}_n(\mathbf{f}_n) - \mathcal{H}_n(\sigma(\mathbf{f}_n)) - s(\mathcal{T}_n(\mathbf{f}_n)) = \mathcal{T}_n(\mathbf{x}_1) - \mathcal{H}_n(\sigma(\mathbf{x}_1)) - \mathcal{T}_n(\mathbf{x}_2) + \mathcal{H}_n(\sigma(\mathbf{x}_2)),$$

where

$$\mathbf{x}_1 = \frac{1}{n+1}\,(0, 0, t_2, 2t_3, \ldots, (n-3)t_{n-2}, (n-2)t_{n-1})^t$$

and

$$\mathbf{x}_2 = \frac{2}{n+1}\left(\sum_{j=1}^{m-1} t_{2j}, \sum_{j=1}^{m-1} t_{2j+1}, \sum_{j=2}^{m-1} t_{2j}, \sum_{j=2}^{m-1} t_{2j+1}, \ldots, t_{n-2}, t_{n-1}, 0, 0\right)^t.$$

As $\mathcal{T}_n(\mathbf{x}_1)$, $\mathcal{T}_n(\mathbf{x}_2)$, $\mathcal{H}_n(\sigma(\mathbf{x}_1))$ and $\mathcal{H}_n(\sigma(\mathbf{x}_2))$ are symmetric matrices, we obtain

$$||\mathcal{T}_n(\mathbf{x}_1)||_2 \le ||\mathcal{T}_n(\mathbf{x}_1)||_1 \le 2||\mathbf{x}_1||_1;$$

$$||\mathcal{T}_n(\mathbf{x}_2)||_2 \le ||\mathcal{T}_n(\mathbf{x}_2)||_1 \le 2||\mathbf{x}_2||_1;$$

$$||\mathcal{H}_n(\sigma(\mathbf{x}_1))||_2 \le 2||\mathbf{x}_1||_1$$

and

$$||\mathcal{H}_n(\sigma(\mathbf{x}_2))||_2 \le 2||\mathbf{x}_2||_1.$$

For all $\epsilon > 0$, since f is in the Wiener class, we can always find positive integers N_1, N_2 and an $N_3 > N_2$ such that

$$\frac{1}{N_1}\sum_{j=1}^{\infty} |t_j| \le \frac{\epsilon}{6}; \quad \sum_{j=N_2+1}^{\infty} |t_j| \le \frac{\epsilon}{24} \quad \text{and} \quad \frac{1}{N_3}\sum_{j=1}^{N_2} j|t_j| \le \frac{\epsilon}{24}.$$

Thus, for all $n > \max\{N_1, N_3\}$, we have $||\mathbf{x}_1||_1 \leq \epsilon/6$ and

$$||\mathbf{x}_2||_1 \leq \frac{2}{N_3} \sum_{j=1}^{N_2} j(|t_{2j}| + |t_{2j+1}|) + 2 \sum_{j=N_2+1}^{\infty} (|t_{2j}| + |t_{2j+1}|) \leq \frac{\epsilon}{6}.$$

Hence the result follows. □

We now extend the result in Lemma 4 to the class of 2π-periodic continuous even functions.

Lemma 6 *Let $f \in \mathbf{C}_{2\pi}$. Then for all $\epsilon > 0$, there exist $N, M > 0$ such that for all $n > N$, at most M eigenvalues of $\mathcal{T}_n(\mathbf{f}_n) - s(\mathcal{T}_n(\mathbf{f}_n))$ have absolute value larger than ϵ.*

Proof: The idea of our proof is to use the Weierstrass Theorem to approximate any real-valued even function in $\mathbf{C}_{2\pi}$ by trigonometric polynomials. Let $f \in \mathbf{C}_{2\pi}$. Then for any $\epsilon > 0$, there exist an integer $M > 0$ and a trigonometric polynomial

$$p(\theta) = \sum_{k=-M}^{M} c_k e^{ik\theta}$$

with $c_k = c_{-k}$ such that $g = f - p$ and

$$||g||_\infty \leq \epsilon, \tag{15}$$

see Cheney [13, p.144]. For all $n > 2M$, by Lemmas 4 and 5, we write

$$\begin{aligned} & \mathcal{T}_n(\mathbf{f}_n) - s(\mathcal{T}_n(\mathbf{f}_n)) \\ = \; & \mathcal{T}_n(\mathbf{g}_n) - s(\mathcal{T}_n(\mathbf{g}_n)) + \mathcal{T}_n(\mathbf{p}_n) - s(\mathcal{T}_n(\mathbf{p}_n)) \\ = \; & \mathcal{T}_n(\mathbf{g}_n) - s(\mathcal{T}_n(\mathbf{g}_n)) + \mathcal{H}_n(\sigma(\mathbf{p}_n)) + \mathcal{T}_n(\mathbf{p}_n) - \mathcal{H}_n(\sigma(\mathbf{p}_n)) - s(\mathcal{T}_n(\mathbf{p}_n)). \end{aligned} \tag{16}$$

We note that the first two terms in right hand side of (16) are matrices of small norm. In fact by (13), Lemma 3 and (15), we have

$$||\mathcal{T}_n(\mathbf{g}_n) - s(\mathcal{T}_n(\mathbf{g}_n))||_2 \leq ||\mathcal{T}_n(\mathbf{g}_n)||_2 + ||s(\mathcal{T}_n(\mathbf{g}_n))||_2 \leq ||g||_\infty + ||g||_\infty \leq 2\epsilon.$$

Since p is a real-valued even function and also in the Wiener class, Lemmas 4 and 5 imply that both matrices $\mathcal{H}_n(\sigma(\mathbf{p}_n))$ and $\mathcal{T}_n(\mathbf{p}_n) - \mathcal{H}_n(\sigma(\mathbf{p}_n)) - s(\mathcal{T}_n(\mathbf{p}_n))$ have spectra clustered around zero. Hence the result follows. □

From (12) and Lemma 3, we can easily show that the smallest eigenvalue of $s(\mathcal{T}_n(\mathbf{f}_n))$ is uniformly bounded from below when f is positive. Using the identity

$$s(\mathcal{T}_n(\mathbf{f}_n))^{-1}\mathcal{T}_n(\mathbf{f}_n) = I_n + s(\mathcal{T}_n(\mathbf{f}_n))^{-1}[(\mathcal{T}_n(\mathbf{f}_n) - s(\mathcal{T}_n(\mathbf{f}_n))],$$

we obtain the following main theorem.

Theorem 4 *Let $f \in \mathbf{C}_{2\pi}$ be positive. Then for all $\epsilon > 0$, there exist $N, M > 0$ such that for all $n > N$, at most M eigenvalues of $s(\mathcal{T}_n(\mathbf{f}_n))^{-1}\mathcal{T}_n(\mathbf{f}_n) - I_n$ have absolute value larger than ϵ.*

It follows from Theorem 4 that the conjugate gradient method when applied to solving $\mathcal{T}_n(\mathbf{f}_n)\mathbf{x} = \mathbf{b}$, converges superlinearly, see for instance Chan [8]. Finally we consider the cost of solving Toeplitz systems. It is known that the cost per iteration in the preconditioned conjugate gradient method is about $5n$ operations plus the cost of computing $\mathcal{T}_n(\mathbf{f}_n)\mathbf{y}$ and $s(\mathcal{T}_n(\mathbf{f}_n))^{-1}\mathbf{d}$ for some vectors $\mathbf{y}$ and $\mathbf{d}$, see Axelsson and Barker [2, p.23]. Both matrix-vector multiplications $\mathcal{T}_n(\mathbf{f}_n)\mathbf{y}$ and $s(\mathcal{T}_n(\mathbf{f}_n))^{-1}\mathbf{d}$ can be done by using fast sine transforms, see for instance Boman and Koltracht [5]. The cost is of $O(n \log n)$ operations. Hence the cost per iteration is of order $O(n \log n)$ operations. As the method converges superlinearly, the number of iterations required for convergence remains bounded. Hence the total cost of solving the Toeplitz system $\mathcal{T}_n(\mathbf{f}_n)\mathbf{x} = \mathbf{b}$ is in $O(n \log n)$ operations. We emphasize that all the computations can be done in real arithmetic.

4 Numerical Examples and Concluding Remarks

In this section, we compare our optimal sine transform based preconditioners $s(T_n)$ with sine transform based preconditioners derived by Boman and Koltracht [5], Strang's circulant preconditioners [21] and T. Chan's circulant preconditioners [6]. We test their performances on four even functions defined on $[-\pi, \pi]$. They are

(i) $\sum_{k=-\infty}^{\infty}(1 + |k|)^{-1.1}e^{ik\theta}$,

(ii) $\theta^4 + 1$,

(iii) θ^2 and

(iv) $\sum_{k=-\infty}^{\infty}(1 + |k|)^{-1}e^{ik\theta}$.

We note that the first three functions are continuous, but the fourth one is not. Also the third function has a zero at $\theta = 0$. The Toeplitz matrices T_n are formed by evaluating the Fourier coefficients of the test functions.

In the test, we used the vector of all ones as the right hand side vector and the zero vector as the initial guess. The stopping criterion is $||\mathbf{e}_q||_2/||\mathbf{e}_0||_2 < 10^{-7}$, where $\mathbf{e}_q$ is the residual vector after q iterations. All computations are done by Matlab on a SUN sparc workstation. Tables 1-4 show the numbers of iterations required for convergence with different choices of preconditioners. In the tables, I denotes no preconditioner was used, S_C, S_B, C_S and C_T are respectively our optimal sine transform based preconditioners, Boman and Koltracht's preconditioners, Strang's circulant preconditioners and T. Chan's circulant preconditioners.

From the numerical results, we see that in all tests, our optimal sine transform based preconditioners S_C performs almost the same as Boman's, Strang's and T. Chan's ones. However, for the test function (iii), the number of iterations of our optimal sine transform based preconditioners S_C is less than that of circulant preconditioners (c.f. Table 3).

n	I	S_C	S_B	C_S	C_T
16	8	6	6	4	7
32	11	6	5	5	6
64	14	5	5	5	5
128	17	5	5	5	5
256	21	5	5	5	5
512	22	5	5	5	5

Table 1. Numbers of iterations for test function (i).

n	I	S_C	S_B	C_S	C_T
16	8	6	6	8	8
32	19	6	5	7	8
64	36	5	5	6	5
128	54	5	5	6	5
256	66	5	5	6	5
512	70	5	5	6	5

Table 2. Number of iterations for test function (ii).

n	I	S_C	S_B	C_S	C_T
16	8	4	5	7	8
32	16	4	5	7	10
64	37	5	5	7	11
128	83	5	6	7	14
256	176	5	6	8	17
512	370	5	6	8	22

Table 3. Number of iterations for test function (iii).

n	I	S_C	S_B	C_S	C_T
16	8	6	6	4	7
32	11	6	5	5	6
64	16	6	5	5	6
128	19	6	5	5	5
256	21	6	5	5	5
512	24	6	5	5	5

Table 4. Number of iterations for test function (iv).

In this paper, we have proposed and analyzed the optimal sine transform based preconditioners $s(A)$ for general symmetric matrices A. For Toeplitz or near-Toeplitz systems arising from the discretization of elliptic problems with Dirichlet boundary conditions, we anticipate them to be better preconditioners than circulant ones. A typical example is the 1-dimensional Laplacian tridiag$[-1, 2, -1]$. In this case, our preconditioner is exact whereas the condition number of the system preconditioned by the optimal circulant preconditioner is of $O(n^{3/2})$, see R. Chan and T. Chan [10]. Recently, Chan and Wong [12] proved that when the optimal sine transform based preconditioners are used in solving the elliptic problems, the condition number of these preconditioned matrices are bounded independent of the sizes of the discretization matrices.

We remark that using the approach in this paper, one can also construct the optimal cosine transform based preconditioner which is defined to be the minimizer of $||R - A||_F$ over the set of matrices R that can be diagonalized by the cosine transform matrix. The cost of construction will also be the same as that of $s(A)$.

References

[1] G. Ammar and W. Gragg, *Superfast Solution of Real Positive Definite Toeplitz System*, SIAM J. Matrix Appl., V9 (1988), pp. 61–76.

[2] O. Axelsson and V. Barker, *Finite Element Solution of Boundary Value Problems: Theory and Computation*, Academic Press, Orlando, Fl., 1983.

[3] G. Bergland, *A Fast Fourier Transform Algorithm for Real-valued Series*, Comm. ACM, V11 (1968), pp. 703–710.

[4] D. Bini and F. Di Benedetto, *A new Preconditioner for the Parallel Solution of Positive Definite Toeplitz Systems*, in Second Ann. Symp. Parallel Algorithms and Architecture, Crete, Greece, 1990, pp. 220–223.

[5] E. Boman and I. Koltracht, *Fast Transform Based Preconditioners for Toeplitz Equations*, Preprint, August 1993.

[6] T. Chan, *An Optimal Circulant Preconditioner for Toeplitz Systems*, SIAM J. Sci. Statist. Comput., V9 (1988), pp. 766–771.

[7] R. Chan, *The Spectrum of a Family of Circulant Preconditioned Toeplitz Systems*, SIAM J. Num. Anal., V26 (1989), pp. 503–506.

[8] R. Chan, *Circulant Preconditioners for Hermitian Toeplitz Systems*, SIAM J. Matrix Anal. Appl., V10 (1989), pp. 542–506.

[9] R. Chan, X. Jin and M. Yeung, *The Circulant Operator in the Banach Algebra of Matrices*, Lin. Alg. Appl. V149 (1991), pp. 41–53.

[10] R. Chan and T. Chan, *Circulant Preconditioners for Elliptic Problems*, J. of Num. Lin. Alg. Appl., V1 (1992), pp. 77–101.

[11] R. Chan and C. Wong, *Best Conditioned Circulant Preconditioners*, Lin. Alg. Appl., to appear.

[12] R. Chan and C. Wong, *Sine Transform Based Preconditioners for Elliptic Problems*, submitted.

[13] E. Cheney, *Introduction to Approximation Theory*, Mcgraw-Hill Book Co., New York, 1966.

[14] C. Chui and A. Chan, *Application of Approximation Theory Methods to Recursive Digital Filter Design*, IEEE Trans. Acoustic. Speech Signal Process., V30 (1982), pp. 18–24.

[15] I. Gohberg and I. Fel'dman, *Convolution Equations and Projection Methods for Their Solution*, Transl. Math. Monographs, V41 (1974), Amer. Math. Soc. Providence, RI.

[16] U. Grenander and G. Szegö, *Toeplitz Forms and their Applications*, 2nd Ed., Chelsea Pub. Co., New York, 1984.

[17] G. Golub and C. Van Loan, *Matrix Computations* , 2nd Ed., The Johns Hopkins University Press, Maryland, 1989.

[18] T. Huckle, *Circulant and Skew-circulant Matrices for Solving Toeplitz Matrix Problems*, SIAM J. Matrix Anal. Appl., V13 (1992), pp. 767–777.

[19] T. Huckle, *Fast Transforms for Tridiagonal Linear Equations*, Bit, to appear.

[20] A. Oppenheim and R. Schafer, *Discrete-time Signal Processing*, Prentice-Hall, Englewood Cliffs, NJ, 1989.

[21] G. Strang, *A Proposal for Toeplitz Matrix Calculations*, Stud. Appl. Math., V74 (1986), pp. 171–176.

[22] E. Tyrtyshnikov, *Optimal and Super-optimal Circulant Preconditioners*, SIAM J. Matrix Anal. Appl., V13 (1992), pp. 459–473.

[23] P. Yip and K. Rao, *Fast Decimation-in-time Algorithms for a Family of Discrete Sine and Cosine Transforms*, Circuits, Syst., Signal Process, V3 (1984), pp. 387–408.