Approximate Inverse Circulant-plus-Diagonal Preconditioners for Toeplitz-plus-Diagonal Matrices

Ng, Michael K.; Pan, Jianyu

Published in:
SIAM Journal on Scientific Computing

DOI:
10.1137/080720280

Published: 21/05/2010

Document Version:
Publisher's PDF, also known as Version of record

SIAM J. SCI. COMPUT.
Vol. 32, No. 3, pp. 1442–1464

APPROXIMATE INVERSE CIRCULANT-PLUS-DIAGONAL PRECONDITIONERS FOR TOEPLITZ-PLUS-DIAGONAL MATRICES*

MICHAEL K. NG† AND JIANYU PAN‡

Abstract. We consider the solutions of Hermitian positive definite Toeplitz-plus-diagonal systems $(T+D)x = b$, where T is a Toeplitz matrix and D is diagonal and positive. However, unlike the case of Toeplitz systems, no fast direct solvers have been developed for solving them. In this paper, we employ the preconditioned conjugate gradient method with approximate inverse circulant-plus-diagonal preconditioners to solving such systems. The proposed preconditioner can be constructed and implemented efficiently using fast Fourier transforms. We show that if the entries of T decay away exponentially from the main diagonals, the preconditioned conjugate gradient method applied to the preconditioned system converges very quickly. Numerical examples including spatial regularization for image deconvolution application are given to illustrate the effectiveness of the proposed preconditioner.

Key words. approximate inverse, circulant matrices, Toeplitz-plus-diagonal matrices, convergence analysis

AMS subject classifications. 65F10, 15A23

DOI. 10.1137/080720280

1. Introduction.

In this paper, we discuss the solutions to a class of Hermitian positive definite Toeplitz-plus-diagonal systems $(T + D)x = b$ by the preconditioned conjugate gradient method. Some applications can be found in [5, 6, 15]. An application of image restoration [11, 14, 18] can be found in numerical results; see section 4.

For n-by-n Toeplitz systems $Tx = b$, fast and superfast direct solvers of complexity $O(n^2)$ and $O(n \log^2 n)$, respectively, have been developed; see, for instance, Trench [23] and Ammar and Gragg [1]. However, there exist no fast direct solvers for solving Toeplitz-plus-diagonal systems. It is mainly because the displacement rank of the matrix $T + D$ can take any value between 0 and n. Hence, fast Toeplitz solvers that are based on small displacement rank of matrices cannot be applied.

We note that given any vector x, the product $(T + D)x$ can be computed in $O(n \log n)$ operations. In fact, Tx can be obtained by FFT by first embedding T into a $2n$-by-$2n$ circulant matrix; see Strang [19]. Thus iterative methods such as the conjugate gradient method can be employed for solving these systems. The convergence rate of the conjugate gradient method depends on the spectrum of the matrix $T + D$; see Golub and van Loan [9]. However, in general, the spectrum of T, and hence of $T+D$, is not clustered, and the method will therefore converge slowly. Hence a suitable preconditioner should be chosen to speed up the convergence.

*Received by the editors April 4, 2008; accepted for publication (in revised form) April 20, 2010; published electronically May 21, 2010.

http://www.siam.org/journals/sisc/32-3/72028.html

†Centre for Mathematical Imaging and Vision and Department of Mathematics, Hong Kong Baptist University, Kowloon Tong, Hong Kong (mng@math.hkbu.edu.hk). This research was supported in part by RGC grants and HKBU FRGs.

‡Department of Mathematics, East China Normal University, Shanghai, 200241, P.R. China (jypan@math.ecnu.edu.cn). This research was supported in part by the National Natural Science Foundation of China (grants 10501013 and 10971070) and the Shanghai Municipal Natural Science Foundation (grant 09ZR1408700).

For Toeplitz systems $Tx = b$, circulant preconditioners have been proved to be successful choices under the assumption that the diagonals of T are Fourier coefficients of a positive 2π-periodic continuous function; see [15]. However, circulant preconditioners do not work for Toeplitz-plus-diagonal systems. In fact, Strang's circulant preconditioner [19] is not even defined for non-Toeplitz matrices. T. Chan's circulant preconditioner, while defined for $T + D$, will not work well when the eigenvalues of D are not clustered (see [5]). Even if we approximate T by a circulant preconditioner C, the matrix $C + D$ cannot be used as a preconditioner since the system $(C + D)z = y$ cannot be solved easily.

In [5], Chan and Ng assumed that the diagonals of T are Fourier coefficients of a non-negative piecewise continuous function f. They showed that if the essential infimum of f is attained by finitely many points in $[-\pi, \pi]$ and if f is sufficiently smooth around these points, then there exists a Hermitian positive definite banded matrix B, with bandwidth independent of n, such that the spectra of $B^{-1}(T + D)$ are uniformly bounded by a constant independent of n. Hence, for a given tolerance, the number of iterations required for convergence is independent of n. Since the banded matrix system $Bx = b$ can be solved in $O(n)$ operations, the total complexity of the method is $O(n \log n)$. The main drawback of this method is that the function f should be known in order to construct effective banded preconditioners. In this case, one possibility is to approximate the function f (see, for instance, [20]). In [2], Bai and Ng studied the use of banded preconditioners for nonsymmetric block Toeplitz-like-plus-diagonal linear systems arising from sinc-Galerkin numerical methods for partial differential equations. Since the generating functions of the corresponding Toeplitz matrices are known, banded preconditioners can be constructed straightforwardly. In [17], multigrid methods are studied for Toeplitz-plus-diagonal linear systems arising from sinc-Galerkin methods. As the generating functions are known, the proposed multigrid method is to incorporate the diagonal matrix into the interpolating process. Numerical results are reported in the paper to show the effectiveness of banded preconditioners for such application. Recently, Serra-Capizzano and Tablino-Possio [21] studied multigrid methods for structured-plus-banded uniformly bounded Hermitian positive definite linear systems and analyzed the optimality of the two-grid method. They showed that several linear systems arising from the approximation of integro-differential equations can be efficiently solved. For references about the development of Toeplitz-plus-diagonal solvers, we refer to [15].

In [10], splitting iterations for circulant-plus-diagonal systems are studied and analyzed. The splitting is based on a circulant matrix and a diagonal matrix. Both matrices can be inverted very efficiently, and therefore the splitting method can be constructed easily. Theoretical results are shown that if the eigenvalues of the circulant matrix have a positive real part, the splitting method converges to the exact solution of the system of linear equations. In [16], Ng and Bai studied a two-step preconditioning strategy based on the banded matrix approximation and the alternating direction implicit iteration for Toeplitz-plus-diagonal systems arising from the sinc-Galerkin method for boundary value problems. In [4], Benzi and Ng considered the iterative solution of weighted Toeplitz least squares problems. Their approach is based on an augmented system formulation. They studied two kinds of preconditioners for the augmented system. They considered a variant of constraint preconditioning and the Hermitian/skew-Hermitian splitting preconditioner for the augmented system. Since the weighting part and the Toeplitz part are separated, the corresponding systems involving these two preconditioners can be efficiently solved. Numerical experiments

have shown that the performance of these two preconditioners is better than that of circulant preconditioners. In [8], Lin, Ng, and Ching proposed to use factorized banded inverse preconditioners for the normal equations matrix of the (Tikhonov) regularized weighted Toeplitz least squares problems: $I + T^*D'T$, where D' is a positive nonconstant diagonal matrix. Numerical results show that the performance of their preconditioners is superior to that of circulant preconditioners. However, the construction of factorized banded inverse preconditioners for Toeplitz matrices requires to specify the bandwidth k of the lower triangular factor in the factorization and involves solving linear systems with the total cost being $O(nk^3)$ operations. For other Toeplitz-related systems, sparsity patterns should be considered and studied.

In this paper, we propose to use an approximate inverse of $C + D$ as a preconditioner for $T + D$. We show that the proposed preconditioner can be constructed and implemented efficiently by using FFTs. We also show that if T has the off-diagonal exponential decay property, then the spectra of the preconditioned matrices are clustered around one. This implies that when the conjugate gradient method is applied to solving the preconditioned system, the method converges very quickly. Numerical examples including spatial regularization for image deconvolution are given to demonstrate the effectiveness of the proposed preconditioner.

The outline of the rest of the paper is as follows. In section 2, we introduce our preconditioners. In section 3, we analyze the spectra of the preconditioned matrices and show that our proposed preconditioners are effective. Finally, numerical examples and concluding remarks are given in section 4.

2. Approximate inverse preconditioners. In this section we present our new preconditioners. Let $A = T + D$, where T is a Hermitian positive definite Toeplitz matrix and $D = \mathrm{diag}(d_1, d_2, \ldots, d_n)$ is a diagonal matrix with $d_i > 0$ $(i = 1, 2, \ldots, n)$.

Our new preconditioner is based on the approximations

$$A^{-1/2}e_j \approx K_j^{-1/2}e_j, \qquad j = 1, 2, \ldots, n,$$

where e_j is the jth column of the identity matrix I and

$$K_j = T + d_jI, \qquad j = 1, 2, \ldots, n.$$

Then we consider the following matrices to be the preconditioner:

$$B_1^{-1} = \left(\sum_{i=1}^{n} K_i^{-1/2}e_ie_i^T\right)^* \left(\sum_{i=1}^{n} K_i^{-1/2}e_ie_i^T\right) \tag{2.1}$$

or

$$\tilde{B}_1^{-1} = \left(\sum_{i=1}^{n} K_i^{-1/2}e_ie_i^T\right) \left(\sum_{i=1}^{n} K_i^{-1/2}e_ie_i^T\right)^*.$$

To construct B_1^{-1} or $\tilde{B}_1^{-1}$, we need to compute the inverse of the square roots of K_i $(i = 1, 2, \ldots, n)$; therefore we further approximate K_i by a circulant matrix. Let C be a circulant approximation (the Strang circulant preconditioner; see, for instance, [15]) for the Toeplitz matrix T. We remark that other successful preconditioners [15] can be considered and used. Let $C_i = C + d_iI$. Then we obtain the following preconditioner which is based on the circulant matrices:

$$B_2^{-1} = \left(\sum_{i=1}^{n} C_i^{-1/2}e_ie_i^T\right)^* \left(\sum_{i=1}^{n} C_i^{-1/2}e_ie_i^T\right) \tag{2.2}$$

or

$$\tilde{B}_2^{-1} = \left(\sum_{i=1}^{n} C_i^{-1/2} e_i e_i^T\right)\left(\sum_{i=1}^{n} C_i^{-1/2} e_i e_i^T\right)^*.$$

It is well known that circulant matrices can be diagonalized in $\mathcal{O}(n \log n)$ operations by making use of FFTs. Hence the product $C_i^{-1/2} y$ for any vector y can be computed by FFTs in $\mathcal{O}(n \log n)$ operations. Therefore, implementing a preconditioner based on B_2^{-1} or $\tilde{B}_2^{-1}$ would require $6n$ FFTs per iteration.

In order to further reduce computational workload, we propose to use the interpolation method to construct the preconditioner. We first choose a small number $\ell \ll n$ of values $\{\tilde{d}_j\}_{j=1}^{\ell} \subset \{d_i\}_{i=1}^{n}$, which covers (most of) the range of values of $\{d_i\}_{i=1}^{n}$. The idea is given as follows. Define the function

$$K_\lambda(x) = (\lambda + x)^{-1/2}, \qquad x \in [a, b],$$

where λ is a certain positive scalar and $a = \min_{1\le i\le n}\{d_i\} > 0$, $b = \max_{1\le i\le n}\{d_i\} > 0$. Let

$$p_\lambda(x) = \phi_1(x) K_\lambda(\tilde{d}_1) + \phi_2(x) K_\lambda(\tilde{d}_2) + \cdots + \phi_\ell(x) K_\lambda(\tilde{d}_\ell) \tag{2.3}$$

be the piecewise linear interpolation for $K_\lambda(x)$ based on the $\ell \ll n$ points

$$\left\{\tilde{d}_k, K_\lambda(\tilde{d}_k)\right\}_{k=1}^{\ell} \subset \left\{d_i, K_\lambda(d_i)\right\}_{i=1}^{n}$$

with the interpolation error satisfying

$$|K_\lambda(x) - p_\lambda(x)| \le \varepsilon_{inter} \quad \text{for } x \in [a, b] \text{ and } \lambda > 0.$$

The above interpolation error bound will be used in the analysis of preconditioners in section 3.4.

Next we precompute the eigenvalues of $\tilde{C}_j \triangleq C + \tilde{d}_j I$:

$$\tilde{C}_j = F \tilde{\Lambda}_j F^*, \qquad j = 1, 2, \ldots, \ell,$$

where F is the Fourier matrix and $\tilde{\Lambda}_j$ is a diagonal matrix whose diagonals are eigenvalues of $\tilde{C}_j$. Finally, we apply interpolation to approximate $C_i^{-1/2}$:

$$C_i^{-1/2} \approx F\left(\sum_{k=1}^{\ell} \alpha_{ik} \tilde{\Lambda}_k^{-1/2}\right) F^*, \qquad i = 1, 2, \ldots, n,$$

where

$$\alpha_{ik} = \phi_k(d_i), \quad k = 1, 2, \ldots, \ell.$$

Therefore we can get the practical preconditioner

$$\begin{aligned} B_3^{-1} &= \left(\sum_{i=1}^{n} F \sum_{k=1}^{\ell} \alpha_{ik} \tilde{\Lambda}_k^{-1/2} F^* e_i e_i^T\right)^* \left(\sum_{i=1}^{n} F \sum_{k=1}^{\ell} \alpha_{ik} \tilde{\Lambda}_k^{-1/2} F^* e_i e_i^T\right) \\ &= \left(\sum_{i=1}^{n} e_i e_i^T F \sum_{k=1}^{\ell} \alpha_{ik} \tilde{\Lambda}_k^{-1/2}\right)\left(\sum_{i=1}^{n}\sum_{k=1}^{\ell} \alpha_{ik} \tilde{\Lambda}_k^{-1/2} F^* e_i e_i^T\right) \\ &= \left(\sum_{k=1}^{\ell} D_k F \tilde{\Lambda}_k^{-1/2}\right)\left(\sum_{k=1}^{\ell} \tilde{\Lambda}_k^{-1/2} F^* D_k\right) \end{aligned} \tag{2.4}$$

or

$$\tilde{B}_3^{-1} = \left(\sum_{k=1}^{\ell} \tilde{\Lambda}_k^{-1/2} F^* D_k\right)\left(\sum_{k=1}^{\ell} D_k F \tilde{\Lambda}_k^{-1/2}\right),$$

where $D_k = \mathrm{diag}(\alpha_{1k}, \alpha_{2k}, \ldots, \alpha_{nk})$. Now computing the product $B_3^{-1}y$ or $\tilde{B}_3^{-1}y$ requires only about $\mathcal{O}(\ell n \log n)$ operations for a moderate number ℓ.

We have the following remarks for the proposed preconditioner: (i) We expect that as the number of interpolation points increases, the number of iterations required for convergence decreases. (See the numerical results in section 4.) However, the cost of forming the preconditioner grows proportionally to the number of interpolation points. In general, it is difficult to determine how to choose the interpolation points. Here we suggest that if the diagonal entries of the matrix D come from underlying function values, then we consider an approximation of underlying function to determine the number of interpolation points in the construction of preconditioners. (ii) The above construction of preconditioners can be applied to block-Toeplitz-Toeplitz-block matrices. Instead of circulant matrices, we make use of block-circulant-circulant-block matrices in the construction process. The interpolation procedure can be applied similarly.

3. Analysis of preconditioners.

3.1. Off-diagonal decay property. In this paper, we analyze the proposed preconditioners when Toeplitz matrices T have the following off-diagonal exponential decay property.

DEFINITION 3.1 (see [22]). *Let $A = [a_{i,j}]_{i,j\in\mathcal{I}}$ be a matrix, where the index set is $\mathcal{I} = \mathbb{Z}, \mathbb{N}$ or $\{1, 2, \ldots, N\}$. We say A belongs to the class $\mathcal{E}_r$ if*

$$|a_{i,j}| \le ce^{-r|i-j|} \qquad \text{for } r > 0 \tag{3.1}$$

and some constant $c > 0$.

Let us first introduce some off-diagonal decay properties of $F(X)$, where X is a Hermitian banded matrix and F is an analytic function.

Suppose F is analytic on a simply connected open region of the complex plane containing the interval $[-1, 1]$. Then there exist ellipses with foci in -1 and 1 such that F is analytic in their interiors. Let $\chi = \alpha + \beta$, where $\alpha > 1$ and $\beta > 0$ are the half axes of such an ellipse with $\alpha^2 - \beta^2 = 1$. We denote such an ellipse by $\mathbb{E}_\chi$, which is uniquely determined by χ. Then we have the following Bernstein's theorem [13].

THEOREM 3.2. *Let the function F be analytic in the interior of the ellipse $\mathbb{E}_\chi$ with $\chi > 1$ and continuous on $\mathbb{E}_\chi$. In addition, suppose $F(x)$ is real for real x. Then*

$$E_k(F) \triangleq \inf\{\|F - p\|_\infty : \deg(p) \le k\} \le \frac{2M(\chi)}{\chi^k(\chi - 1)},$$

where $\deg(p)$ *denotes the degree of the polynomial $p(x)$ and*

$$\|F - p\|_\infty = \max_{-1\le x\le 1} |F(x) - p(x)|, \qquad M(\chi) = \max_{x\in\mathbb{E}_\chi}\{|F(x)|\}.$$

The following two results are due to Benzi and Golub [3].

THEOREM 3.3. *Let B be a Hermitian, m-banded finite matrix and such that $[-1, 1]$ is the smallest interval containing $\sigma(B)$, the spectrum of B. Suppose F is*

analytic in the interior of $\mathbb{E}_\chi$ *with* $\chi > 1$ *and continuous on* $\mathbb{E}_\chi$. *If* $F(x)$ *is real for real* x, *then* $F(B) \in \mathcal{E}_r$, *that is,*

$$|(F(B))(i,j)| \le ce^{-r|i-j|}$$

with

$$c = \max\left\{\frac{2\chi M(\chi)}{\chi - 1},\ \|F(B)\|_2\right\}, \qquad r = \frac{2}{m}\ln\chi.$$

Let A be Hermitian positive definite and m-banded. Then $[\lambda_{min}(A), \lambda_{max}(A)]$ is the smallest interval containing $\sigma(A)$. If we introduce a linear affine function

$$\psi(\lambda) = \frac{2\lambda - (\lambda_{min}(A) + \lambda_{max}(A))}{\lambda_{max}(A) - \lambda_{min}(A)},$$

then $\psi([\lambda_{min}(A), \lambda_{max}(A)]) = [-1, 1]$ and hence $B = \psi(A)$ is symmetric and $[-1, 1]$ is the smallest interval containing $\sigma(B)$. If function f is analytic on $[\lambda_{min}(A), \lambda_{max}(A)]$ and $f(\lambda)$ is real for real λ, then the function $F = f \circ \psi^{-1}$ satisfies the conditions in the above theorem and hence we have the following result [3].

THEOREM 3.4. *Let* A *be a Hermitian positive definite,* m*-banded finite matrix, and let* f *be an analytic function on* $[\lambda_{min}(A), \lambda_{max}(A)]$ *and* $f(\lambda)$ *is real for real* λ. *Then* $f(A)$ *has the off-diagonal decay property. In particular, let* $f(x) = x^{-1}$, $x^{-1/2}$, *and* $x^{1/2}$, *respectively; then* A^{-1}, $A^{-1/2}$, *and* $A^{1/2}$ *have the off-diagonal decay property.*

We remark that the constants c and r depend not only on the function F and the bandwidth of B but also on the largest and smallest eigenvalues of B. For instance, in the special case when B is Hermitian positive definite and $F(x) = x^{-1/2}$, then

$$c = \frac{2\chi}{\chi - 1}\frac{\sqrt{2}}{\sqrt{\frac{(a-b)(\chi^2+1)}{2\chi} + a + b}} \quad \text{and} \quad r = \frac{2}{m}\ln\chi,$$

where $a = \lambda_{min}(B)$, $b = \lambda_{max}(B)$, and $1 < \chi < \frac{\sqrt{\kappa}+1}{\sqrt{\kappa}-1}$ with $\kappa = \frac{b}{a}$ (the spectral condition number of A). For details, see [3].

Moreover, it is easy to prove the following lemma.

LEMMA 3.5. *Let* $A \in \mathcal{E}_{r_1}$ *and* $B \in \mathcal{E}_{r_2}$ *be finite matrices. Then* $AB \in \mathcal{E}_r$ *for some constant* $0 < r < \min\{r_1, r_2\}$.

Now we consider the entries decay property for a special kind of matrices. Let B have the form

$$B = \begin{pmatrix} B_1 & B_2 & & & B_N \\ B_2^* & \ddots & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & B_2 \\ B_N^* & & & B_2^* & B_1 \end{pmatrix}, \tag{3.2}$$

where all blocks are of size $m \times m$. That is, B is a Hermitian block tridiagonal matrix plus two blocks in the northeast and southwest corners. By direct computations we

can see that

$$B^k = \begin{pmatrix} B_{11}^{(k)} & \cdots & B_{1,k+1}^{(k)} & 0 & \cdots & 0 & B_{1,N-k}^{(k)} & \cdots & B_{1,N}^{(k)} \\ \vdots & \ddots & & \ddots & & & & \ddots & \vdots \\ B_{k+1,1}^{(k)} & & \ddots & & \ddots & & & & B_{k,N}^{(k)} \\ 0 & \ddots & & \ddots & & \ddots & & & 0 \\ \vdots & & \ddots & & \ddots & & \ddots & & \vdots \\ 0 & & & \ddots & & \ddots & & \ddots & 0 \\ B_{N-k,1}^{(k)} & & & & \ddots & & \ddots & & B_{N-k+1,N}^{(k)} \\ \vdots & \ddots & & & & \ddots & & \ddots & \vdots \\ B_{N,1}^{(k)} & \cdots & B_{N,k}^{(k)} & 0 & \cdots & 0 & B_{N,N-k+1}^{(k)} & \cdots & B_{NN}^{(k)} \end{pmatrix}, \tag{3.3}$$

where all blocks are of size $m \times m$. That is, B^k is a block $2k$-banded matrix with two block triangles in the northeast and southwest corners.

Analogous to Theorem 3.4, we have the following result.

LEMMA 3.6. *Let B be defined by* (3.2)*, and suppose that all the conditions in Theorem 3.4 are fulfilled. Then we have*

$$|F(B)(i,j)| \le \begin{cases} ce^{-r|i-j|}, & |i-j| \le n/2, \\ ce^{-r|i-j-n|}, & |i-j| \ge n/2, \end{cases} \tag{3.4}$$

where

$$c = \max\left\{ \frac{2\chi M(\chi)}{\chi - 1}, \|F(B)\|_2 \right\} \qquad \text{and} \quad r = \frac{2}{m} \ln \chi.$$

The proof of this lemma is similar to that of Theorem 3.4. The reader is referred to [3] for the proof. This lemma shows that the entries of $F(B)$ are bounded in an exponentially decaying manner away from the main diagonal, northeast and southwest corners, respectively.

3.2. The spectra of $B_1^{-1}A$. For a given integer m, we let

$$A_m = T_m + D \qquad \text{with} \quad T_m(i,j) = \begin{cases} T(i,j), & |i-j| \le m, \\ 0, & \text{otherwise}, \end{cases} \tag{3.5}$$

i.e., T_m is the $2m$-banded matrix extracted from the Toeplitz matrix T. In this subsection, we will show that A_m can be close to A when m is sufficiently large.

In order to analyze the spectrum of the preconditioned matrix, we consider

$$K_j = T_m + d_j I, \qquad j = 1, 2, \ldots, n \tag{3.6}$$

in the following discussion. We remark in the actual computation (in section 4 of the numerical results) that we employ T in the construction of the preconditioner.

First, we show that all the matrices K_i are invertible. Since T is Hermitian positive definite with the off-diagonal decay property (3.1), it follows from (3.5) that

$$\begin{aligned}\|T - T_m\|_2 \le \|T - T_m\|_1 &= \max_{1\le j\le n} \sum_{i=1}^{n} |T(i,j) - T_m(i,j)| \\ &\le 2 \sum_{k=m+1}^{n} ce^{-rk} \\ &\le \frac{2ce^{-r(m+1)}}{1-e^{-r}}.\end{aligned}$$

Thus we may choose m large enough such that $\frac{2ce^{-r(m+1)}}{1-e^{-r}} \le \lambda_{min}(T)$. Then it holds that

$$\lambda_{min}(T_m) \ge \lambda_{min}(T) - \|T - T_m\|_2 \ge 0$$

and

$$\lambda_{min}(K_i) = \lambda_{min}(T_m + d_i I) \ge d_i,$$

which shows that K_i is invertible and therefore that the matrices B_1^{-1} and $\tilde{B}_1^{-1}$ are well defined. We note that if we assume that the bounds of $\sigma(T)$ and d_i are positive and independent of n, then C_i is also positive definite [15]. It follows that the matrices B_2^{-1} and $\tilde{B}_2^{-1}$ are also well defined.

Since $\lambda_{min}(T_m) \ge 0$, in the following discussion, we can assume that T_m is positive semidefinite. It is clear that

$$\|B_1^{-1} - A^{-1}\|_2 \le \|B_1^{-1} - A_m^{-1}\|_2 + \|A_m^{-1} - A^{-1}\|_2. \tag{3.7}$$

Since both A and A_m are Hermitian, we have

$$\|A_m^{-1} - A^{-1}\|_2 \le \|A^{-1}\|_2 \|A - A_m\|_2 \|A_m^{-1}\|_2 \le \|A^{-1}\|_2 \|A_m^{-1}\|_2 \|A - A_m\|_1.$$

Obviously, we obtain

$$\|A_m^{-1}\|_2 = 1/\lambda_{min}(T_m + D) \le 1/d_{min}(D),$$

where $d_{min}(D) = \min_{1\le i\le n}\{d_i\}$. Since $A \in \mathcal{E}_r$ for some constant $r > 0$, it follows from (3.5) and (3.1) that

$$\begin{aligned}\|A - A_m\|_1 \;=\; \max_{1\le j\le n} \sum_{i=1}^{n} |A(i,j) - A_m(i,j)| &= \max_{1\le j\le n} \sum_{i=1}^{n} |T(i,j) - T_m(i,j)| \\ &\le \frac{2ce^{-r(m+1)}}{1-e^{-r}}.\end{aligned}$$

For a given $\varepsilon > 0$, let m_1 satisfy

$$\frac{2c\|A^{-1}\|_2 e^{-r(m_1+1)}}{d_{min}(D)(1-e^{-r})} < \varepsilon/2.$$

Then for all $m \ge m_1$, we have

$$\|A_m^{-1} - A^{-1}\|_2 < \varepsilon/2. \tag{3.8}$$

For the first term on the right-hand side of (3.7), we have

$$\begin{aligned}\|B_1^{-1} - A_m^{-1}\|_2 &\le \|B_1^{-1} - A_m^{-1}\|_1 \\ &= \max_{1\le j\le n} \|(B_1^{-1} - A_m^{-1})e_j\|_1 \\ &\le \max_{1\le j\le n} \left(\|(B_1^{-1} - K_j^{-1})e_j\|_1 + \|(K_j^{-1} - A_m^{-1})e_j\|_1\right). \end{aligned} \tag{3.9}$$

Next we need to estimate the upper bounds of

$$\|(B_1^{-1} - K_j^{-1})e_j\|_1 \quad \text{and} \quad \|(K_j^{-1} - A_m^{-1})e_j\|_1.$$

LEMMA 3.7. *Let A_m and K_j be defined by* (3.5) *and* (3.6)*, respectively. Then for a given $\varepsilon > 0$, there exist a constant c_2 and an integer m_2 such that for all $N \ge m_2$, we have*

$$\left\|\left(K_j^{-1} - A_m^{-1}\right) e_j\right\|_1 \le c_2 \Delta_N d_j + \varepsilon/6,$$

where $\Delta_N d_j = \max_{j-N\le k\le j+N} |d_k - d_j|$.

Proof. It is easy to see that K_j is a Hermitian m-banded matrix. By Theorem 3.4, we know that K_j^{-1} has the off-diagonal decay property (3.1), that is, there exist constants $\tilde{c} > 0$ and $\tilde{r} > 0$ such that

$$|K_j^{-1}(i,k)| \le \tilde{c} e^{-\tilde{r}|i-k|} \quad \text{for } i,k = 1,2,\ldots,n.$$

Then

$$\begin{aligned}\left(K_j^{-1} - A_m^{-1}\right) e_j &= A_m^{-1}(D - d_j I)K_j^{-1} e_j \\ &= A_m^{-1} \begin{pmatrix} d_1 - d_j & & & \\ & d_2 - d_j & & \\ & & \ddots & \\ & & & d_n - d_j \end{pmatrix} \begin{pmatrix} K_j^{-1}(1,j) \\ K_j^{-1}(2,j) \\ \vdots \\ K_j^{-1}(n,j) \end{pmatrix} \\ &= A_m^{-1} \begin{pmatrix} (d_1 - d_j)K_j^{-1}(1,j) \\ (d_2 - d_j)K_j^{-1}(2,j) \\ \vdots \\ (d_n - d_j)K_j^{-1}(n,j) \end{pmatrix}. \end{aligned}$$

For a given $\varepsilon > 0$, let m_2 satisfy

$$\frac{\tilde{c}\|A_m^{-1}\|_1 e^{-\tilde{r}(m_2+1)}}{1 - e^{-\tilde{r}}} \max_{1\le k\le n} |d_k - d_j| \le \varepsilon/12.$$

Then for all $N \ge m_2$, we have

$$\begin{aligned}&\left\|\left(K_j^{-1} - A_m^{-1}\right) e_j\right\|_1 \\ &\quad = \left\|A_m^{-1}(D - d_j I)K_j^{-1} e_j\right\|_1 \\ &\quad \le \|A_m^{-1}\|_1 \sum_{k=1}^{n} |d_k - d_j| \left|K_j^{-1}(k,j)\right| \\ &\quad \le \tilde{c}\|A_m^{-1}\|_1 \sum_{k=1}^{n} |d_k - d_j| e^{-\tilde{r}|j-k|} \end{aligned}$$

$$
\begin{aligned}
&= \tilde{c}\|A_m^{-1}\|_1 \left(\sum_{k=1}^{j-N-1} |d_k - d_j| e^{-\tilde{r}|k-j|} + \sum_{k=j-N}^{j+N} |d_k - d_j| e^{-\tilde{r}|k-j|} + \right. \\
&\qquad \left. \sum_{k=j+N+1}^{n} |d_k - d_j| e^{-\tilde{r}|k-j|} \right) \\
&\le \tilde{c}\|A_m^{-1}\|_1 \left(\frac{e^{-\tilde{r}N}}{1-e^{-\tilde{r}}} \max_{1\le k\le j-N-1} |d_k - d_j| + \frac{2e^{-\tilde{r}}}{1-e^{-\tilde{r}}} \max_{j-N\le k\le j+N} |d_k - d_j| + \right. \\
&\qquad \left. \frac{e^{-\tilde{r}N}}{1-e^{-\tilde{r}}} \max_{j+N+1\le k\le n} |d_k - d_j| \right) \\
&\le \frac{2\tilde{c}\|A_m^{-1}\|_1 e^{-\tilde{r}}}{1-e^{-\tilde{r}}} \max_{j-N\le k\le j+N} |d_k - d_j| + \frac{\varepsilon}{6} \\
&= \ c_2 \Delta_N d_j + \frac{\varepsilon}{6},
\end{aligned}
$$

where

$$
\Delta_N d_j = \max_{j-N\le k\le j+N} |d_k - d_j| \quad \text{and} \quad c_2 = \frac{2\tilde{c}\|A_m^{-1}\|_1 e^{-\tilde{r}}}{1-e^{-\tilde{r}}}.
$$

The result follows. □

LEMMA 3.8. *Let B_1 and K_j be defined by* (2.1) *and* (3.6)*, respectively. Then for a given $\varepsilon > 0$, there exist a constant c_3 and an integer m_3 such that for all $N \ge m_3$, we have*

$$
\left\| \left(B_1^{-1} - K_j^{-1} \right) e_j \right\|_1 \le c_3 \Delta_N d_j + \varepsilon/6,
$$

where $\Delta_N d_j = \max_{j-N\le k\le j+N} |d_k - d_j|$.

Proof. It is obvious that

$$
K_i K_j = (T_m + d_i I)(T_m + d_j I) = (T_m + d_j I)(T_m + d_i I) = K_j K_i,
$$

which means that K_i and K_j commute for all $i, j = 1, 2, \ldots, n$. Then it follows that

$$
\begin{aligned}
\left(B_1^{-1} - K_j^{-1} \right) e_j &= \left(\sum_{i=1}^{n} e_i e_i^T K_i^{-1/2} \right) \left(\sum_{i=1}^{n} K_i^{-1/2} e_i e_i^T \right) e_j - K_j^{-1} e_j \\
&= \left(\sum_{i=1}^{n} e_i e_i^T K_i^{-1/2} \right) K_j^{-1/2} e_j - K_j^{-1} e_j \\
&= \sum_{i=1}^{n} e_i e_i^T \left(K_i^{-1/2} - K_j^{-1/2} \right) K_j^{-1/2} e_j \\
&= \sum_{i=1}^{n} e_i e_i^T K_i^{-1/2} K_j^{-1/2} \left(K_i^{1/2} + K_j^{1/2} \right)^{-1} (K_j - K_i) K_j^{-1/2} e_j \\
&\triangleq \sum_{i=1}^{n} e_i e_i^T (d_j - d_i) H_{ij} e_j,
\end{aligned}
$$

where

$$
H_{ij} = K_i^{-1/2} K_j^{-1/2} \left(K_i^{1/2} + K_j^{1/2} \right)^{-1} K_j^{-1/2} = \left(K_i K_j + K_i^{1/2} K_j^{3/2} \right)^{-1}.
$$

It follows from Theorem 3.4, Lemma 3.5, and Theorem 2.3 in [22], which is due to Jaffard [12], that H_{ij} has the off-diagonal decay property (3.1), that is, there exist constants $c_{ij} > 0$ and $r_{ij} > 0$ such that

$$|H_{ij}(k,l)| \le c_{ij} e^{-r_{ij}|k-l|} \qquad \text{for } k,l = 1,2,\ldots,n.$$

Therefore

$$|e_i^T(d_j - d_i)H_{ij}e_j| = |d_j - d_i| \cdot |H_{ij}(i,j)| \le c_{ij}|d_j - d_i|e^{-r_{ij}|i-j|} \quad \text{for } i = 1,2,\ldots,n.$$

Let $\hat{c}_j = \max_{1\le i\le n}\{c_{ij}\} > 0$ and $\hat{r}_j = \min_{1\le i\le n}\{r_{ij}\} > 0$. Then

$$\left\|\left(B_1^{-1} - K_j^{-1}\right)e_j\right\|_1 = \left\|\sum_{i=1}^n e_i e_i^T(d_j - d_i)H_{ij}e_j\right\|_1 = \sum_{i=1}^n |e_i^T(d_j - d_i)H_{ij}e_j| \le \hat{c}_j \sum_{i=1}^n |d_j - d_i|e^{-\hat{r}_j|i-j|}.$$

The rest of the proof is similar to that of Lemma 3.7. ∎

From Lemmas 3.7 and 3.8, we immediately have the following conclusion.

COROLLARY 3.9. *Let A_m, K_j, and B_1 be defined by* (3.5), (3.6), *and* (2.1), *respectively. Then for a given $\varepsilon > 0$, there exist a constant $c_4 = c_2 + c_3$ and an integer $m_4 = \max\{m_2, m_3\}$ such that for all $N \ge m_4$, we have*

$$\|(B_1^{-1} - K_j^{-1})e_j\|_1 + \|(K_j^{-1} - A_m^{-1})e_j\|_1 \le c_4\Delta_N d_j + \varepsilon/3,$$

where $\Delta_N d_j = \max_{j-N\le k\le j+N}|d_k - d_j|$.

If D is generated by some function $d(x) \in C^1[0,1]$, i.e., $d_i = d(x_i)$ with $x_i = i/n$, $i = 1,2,\ldots,n$, then we have

$$\Delta_N d_j = \max_{j-N\le k\le j+N}|d_k - d_j| \le \xi(x_{j+N} - x_{j-N}) = \frac{2N\xi}{n},$$

where $\xi = \max_{0\le x\le 1} d'(x)$. Let n_1 be large enough such that $2c_4N\xi/n_1 \le \varepsilon/6$. Then for all $n \ge n_1$, we have $c_4\Delta_N d_j < \varepsilon/6$, which, together with (3.8), (3.9), and Corollary 3.9, leads to the following theorem.

THEOREM 3.10. *Let T be a Hermitian positive definite matrix and D be a positive diagonal matrix. Assume T has the off-diagonal decay property* (3.1) *and that the entries of D satisfy $d_i = d(i/n)(i = 1,2,\ldots,n)$ with $d(x) \in C^1[0,1]$. Then for a given $\varepsilon > 0$, there exist integers m and n_1 such that for all $n \ge n_1$, we have*

$$\|B_1^{-1} - A^{-1}\|_2 \le \varepsilon,$$

where B_1 is defined in (2.1).

Assume that T is the finite section of size n of an infinite Toeplitz matrix and that $\|A\|_2$ is uniformly bounded. Then it follows from Theorem 3.10 and the fact

$$\|B_1^{-1}A - I\|_2 = \|B_1^{-1} - A^{-1}\|_2\|A\|_2$$

that the spectrum of $B_1^{-1}A$ is clustered around one. Therefore, if the preconditioned conjugate gradient method is applied to the preconditioned system $B_1^{-1}Ax = B_1^{-1}b$, it will converge rapidly.

Analogously, we can obtain the same results for the preconditioner $\tilde{B}_1$.

3.3. The spectra of $B_2^{-1}A$. The second step is to consider the difference between B_1^{-1} and B_2^{-1}. Let $K = T_m + dI$, where $d > 0$ is constant. It is easy to see that K can be regarded as a block tridiagonal matrix, that is,

$$K = \begin{pmatrix} K_0 & K_1 & & & \\ K_1^T & K_0 & K_1 & & \\ & \ddots & \ddots & \ddots & \\ & & K_1^T & K_0 & K_1 \\ & & & K_1^T & K_0 \end{pmatrix},$$

where $K_0, K_1 \in \mathbb{R}^{m\times m}$. Then the Strang preconditioner for K is defined by

$$C = \begin{pmatrix} K_0 & K_1 & & & K_1^T \\ K_1^T & K_0 & K_1 & & \\ & \ddots & \ddots & \ddots & \\ & & K_1^T & K_0 & K_1 \\ K_1 & & & K_1^T & K_0 \end{pmatrix}.$$

Let $F(x) = x^{-1/2}$. It follows from [3] and Theorem 3.2 that there exists a polynomial p_k with $\deg(p_k) \le k$ such that

$$\|K^{-1/2} - p_k(K)\|_2 = \max_{x\in\sigma(K)} |x^{-1/2} - p_k(x)| \le \|F - p_k\|_\infty \le \frac{2M(\chi_K)}{\chi_K - 1} \cdot \frac{1}{\chi_K^k},$$
$$\|C^{-1/2} - p_k(C)\|_2 = \max_{x\in\sigma(C)} |x^{-1/2} - p_k(x)| \le \|F - p_k\|_\infty \le \frac{2M(\chi_C)}{\chi_C - 1} \cdot \frac{1}{\chi_C^k},$$

where

$$1 < \chi_C < \frac{\sqrt{\kappa_C}+1}{\sqrt{\kappa_C}-1}, \quad 1 < \chi_K < \frac{\sqrt{\kappa_K}+1}{\sqrt{\kappa_K}-1},$$

and κ_C and κ_K are the condition numbers of C and K, respectively. Therefore, for a given $\varepsilon_{ap} > 0$, there exists an integer N_{ap} such that for all $k \ge N_{ap}$, we have

$$\|K^{-1/2} - p_k(K)\|_2 \le \varepsilon_{ap} \quad \text{and} \quad \|C^{-1/2} - p_k(C)\|_2 \le \varepsilon_{ap}. \tag{3.10}$$

Now we want to show that

$$B_2^{-1} - B_1^{-1} = E_1 + M_1,$$

where E_1 is of small norm and $\operatorname{rank}(M_1) \le 4km$. Obviously we have

$$B_2^{-1} - B_1^{-1} = (B_2^{-1} - \hat{B}_2) + (\hat{B}_2 - \hat{B}_1) + (\hat{B}_1 - B_1^{-1}), \tag{3.11}$$

where

$$\hat{B}_2 = \left(\sum_{i=1}^n p_k(C_i)e_ie_i^T\right)^* \left(\sum_{i=1}^n p_k(C_i)e_ie_i^T\right) \tag{3.12}$$

and

$$\hat{B}_1 = \left(\sum_{i=1}^n p_k(K_i)e_ie_i^T\right)^* \left(\sum_{i=1}^n p_k(K_i)e_ie_i^T\right). \tag{3.13}$$

LEMMA 3.11. *Let $\hat{B}_2$ and $\hat{B}_1$ be defined by* (3.12) *and* (3.13)*, respectively. Then we have*

$$\operatorname{rank}(\hat{B}_2 - \hat{B}_1) \le 4km$$

for some k.

Proof. We first investigate the structure of $p_k(C) - p_k(K)$. By direct computations, we can see that

$$C^\alpha (C-K)K^\beta = \begin{pmatrix} * & \cdots & * & * & \cdots & * \\ \vdots & & \vdots & \vdots & & \vdots \\ * & \cdots & * & * & \cdots & * \\ & & & & & \\ * & \cdots & * & * & \cdots & * \\ \vdots & & \vdots & \vdots & & \vdots \\ * & \cdots & * & * & \cdots & * \end{pmatrix} \quad \text{for integer values } \alpha \text{ and } \beta,$$

that is, $C^\alpha(C-K)K^\beta$ is a block matrix with four blocks in its four corners, respectively, and each block is of size $(\alpha+1)m \times (\beta+1)m$. Hence

$$C^k - K^k = \sum_{i=0}^{k-1} (C^{k-i}K^i - C^{k-i-1}K^{i+1}) = \sum_{i=0}^{k-1} C^{k-i-1}(C-K)K^i,$$

which is also a block matrix with four blocks in its four corners, respectively, and each block is of size $km \times km$. It follows that

$$p_k(S) - p_k(K) = \sum_{i=0}^{k} a_i (C^i - K^i)$$

has the same form as that of $C^k - K^k$.

Let $K = K_i$ and $C = C_i$. Then the above conclusion holds for all $i = 1, 2, \ldots, n$. Therefore

$$\sum_{i=1}^{n} p_k(C_i) e_i e_i^T - \sum_{i=1}^{n} p_k(K_i) e_i e_i^T = \sum_{i=1}^{n} (p_k(C_i) - p_k(K_i)) e_i e_i^T$$

has the same structure as that of $C^k - K^k$, which means that its rank is less than or equal to $2km$. Since

$$\hat{B}_2 - \hat{B}_1 = \left(\sum_{i=1}^{n} e_i e_i^T p_k(S_i) \right) \left(\sum_{i=1}^{n} p_k(S_i) e_i e_i^T - \sum_{i=1}^{n} p_k(K_i) e_i e_i^T \right) + \left(\sum_{i=1}^{n} p_k(S_i) e_i e_i^T - \sum_{i=1}^{n} p_k(K_i) e_i e_i^T \right)^* \sum_{i=1}^{n} p_k(K_i) e_i e_i^T,$$

it follows that $\operatorname{rank}(\hat{B}_2 - \hat{B}_1) \le 4km$. ∎

LEMMA 3.12. *There exist some constants c_5 and c_6 such that*

$$\|\hat{B}_1 - B_1^{-1}\|_1 \le c_5 \varepsilon_{ap} \qquad \text{and} \quad \|B_2^{-1} - \hat{B}_2\|_1 \le c_6 \varepsilon_{ap}.$$

Proof. Denote

$$\tilde{K}_i \triangleq K_i^{-1/2} - p_k(K_i);$$

then $\tilde{K}_i$ is Hermitian, and it follows from (3.10) that

$$|\tilde{K}_i(k,j)| \le \|\tilde{K}_i\|_2 \le \varepsilon_{ap}. \tag{3.14}$$

Since $K_i^{-1/2}$ has the off-diagonal decay property and $p_k(K_i)$ is $2km$-banded, we can see that $\tilde{K}_i$ also has the off-diagonal decay property, that is, there exist constants $\tilde{c}_i$ and $\tilde{r}_i > 0$ such that

$$|\tilde{K}_i(k,j)| \le \tilde{c}_i e^{-\tilde{r}_i|k-j|}, \qquad i = 1, 2, \ldots, n. \tag{3.15}$$

Let $\tilde{c} = \max_{1\le i\le n} \tilde{c}_i > 0$ and $\tilde{r} = \min_{1\le i\le n} \tilde{r}_i > 0$, and denote

$$H \triangleq \sum_{i=1}^{n} \tilde{K}_i e_i e_i^T = \sum_{i=1}^{n} (K_i^{-1/2} - p_k(K_i)) e_i e_i^T.$$

Then it follows from (3.14) and (3.15) that

$$|H(i,j)| \le \varepsilon_{ap} \qquad \text{and} \quad |H(i,j)| \le \tilde{c} e^{-\tilde{r}|i-j|}. \tag{3.16}$$

For a given $\varepsilon_{ap} > 0$, there exists an integer N such that

$$\sum_{i=N}^{\infty} \tilde{c} e^{-\tilde{r}i} \le \varepsilon_{ap}.$$

From (3.16), we have

$$\begin{aligned} \|He_j\|_1 \;=\; \sum_{i=1}^{n} |H(i,j)| &= \sum_{i=j-N+1}^{j+N-1} |H(i,j)| + \sum_{i=1}^{j-N} |H(i,j)| + \sum_{i=j+N}^{n} |H(i,j)| \\ &\le \sum_{i=j-N+1}^{j+N-1} \varepsilon_{ap} + 2\sum_{i=N}^{\infty} \tilde{c} e^{-\tilde{r}i} \\ &\le (2N-1)\varepsilon_{ap}. \end{aligned}$$

Therefore

$$\left\| \sum_{i=1}^{n} K_i^{-1/2} e_i e_i^T - \sum_{i=1}^{n} p_k(K_i) e_i e_i^T \right\|_1 = \|H\|_1 = \max_{1\le j\le n} \|He_j\|_1 \le (2N+1)\varepsilon_{ap}.$$

Analogously, we can show that

$$\left\| \sum_{i=1}^{n} e_i e_i^T K_i^{-1/2} - \sum_{i=1}^{n} e_i e_i^T p_k(K_i) \right\|_1 = \|H\|_\infty = \max_{1\le i\le n} \|e_i^T H\|_1 \le (2N+1)\varepsilon_{ap},$$

and it holds that

$$
\begin{aligned}
&\|\hat{B}_1 - B_1^{-1}\|_1 \\
&= \left\| \left(\sum_{i=1}^{n} K_i^{-1/2} e_i e_i^T \right)^* \left(\sum_{i=1}^{n} K_i^{-1/2} e_i e_i^T \right) - \left(\sum_{i=1}^{n} p_k(K_i) e_i e_i^T \right)^* \left(\sum_{i=1}^{n} p_k(K_i) e_i e_i^T \right) \right\|_1 \\
&\le \left\| \sum_{i=1}^{n} e_i e_i^T K_i^{-1/2} \right\|_1 \left\| \sum_{i=1}^{n} K_i^{-1/2} e_i e_i^T - \sum_{i=1}^{n} p_k(K_i) e_i e_i^T \right\|_1 \\
&\quad + \left\| \sum_{i=1}^{n} e_i e_i^T K_i^{-1/2} - \sum_{i=1}^{n} e_i e_i^T p_k(K_i) \right\|_1 \left\| \sum_{i=1}^{n} p_k(K_i) e_i e_i^T \right\|_1 \\
&\le c_5 \varepsilon_{ap},
\end{aligned}
$$

where c_5 is some constant.

Now we turn to estimate $\|B_2^{-1} - \hat{B}_2\|_1$. Let $\tilde{S}_i \triangleq S_i^{-1/2} - p_k(S_i)$. Then by (3.10) we have

$$|\tilde{S}_i(k,j)| \le \|\tilde{S}_i\|_2 \le \varepsilon_{ap}.$$

Let $F(x) = x^{-1/2}$. By Lemma 3.6 we can see that $S_i^{-1/2}$ has the entries decay property (3.4). It is easy to see that $p_k(S_i)$ has the same special structure as B^k in (3.3). Hence $\tilde{S}_i$ also has the entries decay property (3.4). The rest of the proof is analogous to that of the first part. ∎

By Lemmas 3.11 and 3.12, we obtain the following theorem.

THEOREM 3.13. *Let B_1^{-1} and B_2^{-1} be defined by* (2.1) *and* (2.2)*, respectively. Then*

$$B_2^{-1} - B_1^{-1} = E_1 + M_1,$$

where $E_1 = (B_2^{-1} - \hat{B}_2) + (\hat{B}_1 - B_1^{-1})$ is of small norm and $M_1 = \hat{B}_2 - \hat{B}_1$ is of low rank.

Remark. For a given ε_{ap}, there exist an integer N_{ap} such that (3.8) holds true for $K_i = T_m + d_i I$. But the integer N_{ap} may be different for different K_i $(i = 1, 2, \ldots, n)$, and it depends on how large χ_{K_i} is. In order to make sure that (3.8) holds true for all K_i, we need to choose the largest N_{ap}, which is the value of k in Lemma 3.11. However, it may depend on n because it depends on χ_{K_i} $(i = 1, 2, \ldots, n)$, which has the lower bound

$$\min_{1 \le i \le n} \left\{ \frac{\sqrt{\text{cond}(K_i)} + 1}{\sqrt{\text{cond}(K_i)} - 1} \right\}.$$

If the above value tends to 1 as n tends to infinity, then the largest N_{ap} may also tend to infinity. Therefore, in order to show that k in Lemma 3.11 is independent on n, we need to assume that there exists an upper bound of $\text{cond}(K_i) = \lambda_{max}(K_i)/\lambda_{min}(K_i)$ that is independent on n. For instance, if we assume that (1) the eigenvalues of $T_m \in \mathbb{C}^{n \times n}$ satisfy $c_1 \le \lambda(T_m) \le c_2$ (i.e., the Toeplitz matrices are generated by a positive function), where $c_1 > 0$ and $c_2 > 0$ are independent on n, and (2) d_i satisfy $a \le d_i \le b$, where $a > 0$ and $b > 0$ are independent on n, then the results in Theorem 3.13 hold.

3.4. The spectra of $B_3^{-1}A$. The final step is to analyze the spectra of $B_3^{-1}A$.

We recall that the polynomials $\phi_k(x)$ $(k = 1, 2, \ldots, \ell)$ in (2.3) are independent on λ. Let $T_m = U^* \Lambda_{T_m} U$ be the spectral decomposition of T_m, where

$$\Lambda_{T_m} = \operatorname{diag}(\hat{\lambda}_1, \hat{\lambda}_1, \ldots, \hat{\lambda}_n).$$

Define

$$\hat{K}_i = \alpha_{i1}(T_m + \tilde{d}_1 I)^{-1/2} + \alpha_{i2}(T_m + \tilde{d}_2 I)^{-1/2} + \cdots + \alpha_{i\ell}(T_m + \tilde{d}_\ell I)^{-1/2},\ i = 1, 2, \ldots, n,$$

with $\alpha_{ik} = \phi_k(d_i)$ $(k = 1, 2, \ldots, \ell)$. Then we have

$$\begin{aligned}
\left\| K_i^{-1/2} - \hat{K}_i \right\|_2 &= \left\| U^*(\Lambda_{T_m} + d_i I)^{-1/2} U - \sum_{k=1}^{\ell} \alpha_{ik} U^* (\Lambda_{T_m} + \tilde{d}_k I)^{-1/2} U \right\|_2 \\
&= \left\| (\Lambda_{T_m} + d_i I)^{-1/2} - \sum_{k=1}^{\ell} \alpha_{ik} (\Lambda_{T_m} + \tilde{d}_k I)^{-1/2} \right\|_2 \\
&= \max_{1 \le j \le n} \left| K_{\hat{\lambda}_j}(d_i)^{-1/2} - p_{\hat{\lambda}_j}(d_i) \right| \\
&\le \varepsilon_{inter}.
\end{aligned}$$

Analogously to the second part of Lemma 3.12, we can easily show the following lemma.

LEMMA 3.14. *Let B_2^{-1} and B_3^{-1} be defined by* (2.2) *and* (2.4)*, respectively. Then there exists a constant $c_7 > 0$ such that*

$$\|B_3^{-1} - B_2^{-1}\|_1 \le c_7 \varepsilon_{inter}.$$

THEOREM 3.15. *Let B_3^{-1} be defined by* (2.4)*. Then*

$$B_3 - A = E + M,$$

where E is of a small norm and M is of a low rank.

Proof. It is easy to see that

$$\begin{aligned}
B_3 - A &= -B_3 \left(B_3^{-1} - A^{-1} \right) A \\
&= -B_3 \left((B_3^{-1} - B_2^{-1}) + (B_2^{-1} - B_1^{-1}) + (B_1^{-1} - A^{-1}) \right) A \\
&\triangleq E + M,
\end{aligned}$$

where

$$E = -B_3 \left((B_3^{-1} - B_2^{-1}) + E_1 + (B_1^{-1} - A^{-1}) \right) A \qquad \text{and} \quad M = -B_3 M_1 A.$$

Then by Theorems 3.10 and 3.13 and Lemma 3.14, we can obtain the result. ∎

To summarize the results in this section, we show that if T has the off-diagonal exponential decay property, then the spectra of the preconditioned matrices are clustered around one. This implies that when the conjugate gradient method is applied to solving the preconditioned system, the method converges very quickly (see [6]). In the next section, numerical examples are given to demonstrate the effectiveness of the proposed preconditioner.

4. Numerical examples. In this section, we present numerical experiments to illustrate the performance of the proposed preconditioner. We first consider the linear system

$$(T + D)x = b,$$

where T is a Hermitian positive definite Toeplitz matrix with generating function $f(x)$ and D is a diagonal matrix defined by

$$D = f_{max} \cdot \mathrm{diag}\left(0, \frac{1}{n}, \ldots, \frac{n-1}{n}\right).$$

The eigenvalues of D are distributed uniformly in the interval $[0, f_{max}]$. Such systems are tested in [5]. In our experiments, we set the right-hand side vector to be $b = [1, 1, \ldots, 1]^T$, and we choose the zero vector as the initial guess. The stopping criterion is $\|r_k\|_2/\|b\|_2 < 10^{-7}$, where r_k is the residual vector after k iterations.

We first test Toeplitz matrices generated by $|a_{i,j}| = e^{-0.01|i-j|}$ (see (3.1)). Moreover, three different generating functions of Toeplitz matrices were tested. They are $f_1(x) = x^4$, $f_2(x) = \cosh(x)$, and

$$f_3(x) = \begin{cases} x^2, & |x| \le \pi/2, \\ 1, & |x| > \pi/2. \end{cases}$$

The above generating functions are first defined on $[-\pi, \pi]$ and then extended periodically to the whole real line. These generated Toeplitz matrices are symmetric and positive definite. As the diagonal entries of these Toeplitz matrices are given by the Fourier coefficients of the generating functions, these Toeplitz matrices have the off-diagonal decay property [15]. The decay rates of the Fourier coefficients of the first and second generating functions are $O(1/k^2)$, where k refers to the index of the k-th Fourier coefficient. Since f_3 has a jump at $\pi/2$, its Fourier coefficients decay in $O(1/k)$ manner. For the functions $f_1(x)$ and $f_3(x)$, the generated Toeplitz matrices are ill-conditioned as they have a zero at $x = 0$; see, for instance, [15]. The number of iterations required for convergence is expected to be large.

TABLE 4.1
Numerical results for (3.1) *with* $r = 0.01$.

n	32	64	128	256	512	1024	2048
CG	29	41	63	98	160	256	392
PCG(4)	12	16	19	21	26	26	27
PCG(8)	10	13	15	17	21	22	24
PCG(16)	9	11	12	14	16	19	20
PCG(32)	10	10	11	11	13	14	15
PCG(T)	25	33	42	52	60	65	69

The iteration numbers of different kinds of methods are listed in Tables 4.1, 4.2, 4.3, and 4.4, where CG denotes the conjugate gradient method, PCG(T) denotes the preconditioned CG method with the T. Chan circulant preconditioner, and PCG(ℓ) denotes the preconditioned CG method with B_3 as the preconditioner and ℓ is the number of interpolation points. Here the interpolation points $\{\tilde{d}_j\}_{j=1}^{\ell}$ are chosen uniformly located in the interval $[d_{min}, d_{max}]$. The construction and implementation

TABLE 4.2
Numerical results for $f_1(x)$.

n	32	64	128	256	512	1024	2048
CG	26	36	50	68	91	122	162
PCG(4)	10	13	16	21	27	36	47
PCG(8)	8	9	12	15	19	25	33
PCG(16)	7	9	9	11	14	18	23
PCG(32)	7	9	8	9	10	13	16
PCG(T)	23	31	40	53	70	91	119
B	12	14	14	15	15	15	15

TABLE 4.3
Numerical results for $f_2(x)$.

n	32	64	128	256	512	1024	2048
CG	21	25	29	32	34	36	36
PCG(4)	8	9	10	11	11	12	12
PCG(8)	6	7	8	8	9	9	9
PCG(16)	6	6	7	7	7	7	7
PCG(32)	6	6	6	6	6	6	6
PCG(T)	18	21	23	25	27	27	28
B	8	9	9	10	10	10	10

cost of the preconditioner is still of $O(\ell n \log n)$ operations; see section 2. For Toeplitz matrices generated by functions f_1, f_2, and f_3, banded preconditioners B can be constructed [5]; see Tables 4.2, 4.3, and 4.4. However, the generating function for Toeplitz matrices in Table 4.1 is unknown, and banded preconditioners cannot be defined in a suitable manner. We observe from the tables that the performance of the proposed preconditioner is quite good, compared with the other preconditioners. In Tables 4.3, and 4.4, the performance of the proposed preconditioner is better than the other preconditioners. Also, when the number of interpolation points is larger, the number of iterations required for convergence is smaller.

To further demonstrate the usefulness of the proposed preconditioner, we consider an example of block-Toeplitz-Toeplitz-block systems, where the coefficient matrices are equal to $(T \otimes T + D)$ and T the Toeplitz matrices generated by f_3. Here the size of block is n, and the number of blocks is also n. Also D is constructed similarly as above, and the eigenvalues of D are distributed uniformly in the interval $[0, f_{max}^2]$. We note that the cost of using banded preconditioners is expensive since the corresponding bandwidth is very large. The numerical results are listed in Table 4.5. We see from the table that the standard circulant preconditioner does not work,

TABLE 4.4
Numerical results for $f_3(x)$.

n	32	64	128	256	512	1024	2048
CG	18	23	30	39	50	63	81
PCG(4)	9	9	10	12	15	19	23
PCG(8)	8	8	9	10	11	13	17
PCG(16)	8	8	8	9	9	11	13
PCG(32)	8	8	9	9	9	9	10
PCG(T)	16	19	24	30	38	47	59
B	12	14	14	15	15	15	15

TABLE 4.5
Numerical results for block-Toeplitz-Toeplitz-block matrices.

n^2	8^2	16^2	32^2	64^2	128^2	256^2
CG	23	38	64	111	192	318
PCG(4)	12	16	23	36	58	93
PCG(8)	12	15	20	29	45	73
PCG(16)	11	14	18	25	36	58
PCG(32)	12	14	17	22	31	47
PCG(T)	20	30	47	75	121	193

but the proposed preconditioner works quite promisingly. We remark that although a banded preconditioner can be constructed in this case (the generating function of T is known), the cost of using such a banded preconditioner is very expensive. The related computational cost is $O(n^4)$ as the bandwidth is $O(n)$. However, the cost of the proposed preconditioner is $O(\ell n^2 \log n)$ by using FFTs efficiently.

4.1. Spatial regularization for image deconvolution. We use an image deconvolution as an example. For simplicity, let us first consider the convolution of a discrete signal x of length n with a convolution vector h of the form

$$h = [h_{-m+1}, h_{-m+2}, \ldots, h_0, \ldots, h_{m-2}, h_{m-1}]^T.$$

The resulting vector b is of length $2m + n - 2$, and the convolution operation can be expressed in matrix notation as $b = H_{m,n}x$, where $H_{m,n}$ is a column circulant matrix of the form

$$H_{m,n} = \begin{pmatrix} h_{-m+1} & & & 0 \\ h_{-m+2} & h_{-m+1} & & \\ \vdots & \ddots & \ddots & \\ h_0 & & \ddots & h_{-m+1} \\ \vdots & \ddots & & \vdots \\ h_{m-2} & & \ddots & \\ h_{m-1} & \ddots & & h_0 \\ & h_{m-1} & \ddots & \\ & & \ddots & \vdots \\ 0 & & & h_{m-1} \end{pmatrix}. \tag{4.1}$$

Since $H_{m,n}$ is a column circulant matrix, the normal equations matrix $H_{m,n}^T H_{m,n}$ is a Toeplitz matrix. The aim is to compute x. This is known as a discrete deconvolution; see [15] for details.

For two-dimensional imaging application, similar normal equations can be formed; see [7] for details. Here the matrix A is a block column circulant matrix with column circulant blocks. More precisely,

$$A = \begin{pmatrix} A_{-m+1} & & & 0 \\ A_{-m+2} & A_{-m+1} & & \\ \vdots & \ddots & \ddots & \\ A_0 & & \ddots & A_{-m+1} \\ \vdots & \ddots & & \vdots \\ A_{m-2} & & \ddots & \\ A_{m-1} & \ddots & & A_0 \\ & A_{m-1} & \ddots & \\ & & \ddots & \vdots \\ 0 & & & A_{m-1} \end{pmatrix} \tag{4.2}$$

with each subblock A_j being a $2m + n - 2$-by-n matrix of the form given by (4.1). We note that $A^T A$ will be an n-by-n block Toeplitz matrix with n-by-n Toeplitz blocks. The two-dimensional deconvolution problem has n^2 unknowns since A has n^2 columns.

It is well known that deconvolution algorithms can be extremely sensitive to noise. In [7], Chan, Ng, and Plemmons used Tikhonov regularization and solved the least squares problem

$$\min \left\| \begin{pmatrix} b \\ 0 \end{pmatrix} - \begin{pmatrix} A \\ \mu I \end{pmatrix} x \right\|_2, \tag{4.3}$$

where μ is the regularization parameter and I is the identity matrix. The solution x of (4.3) can be obtained by solving the normal equations

$$(\mu^2 I + A^T A)x = A^T b. \tag{4.4}$$

We note in (4.4) that we regularize the same amount (μ) of each entry of x. In [11, 18, 14], different amounts of regularization can be applied to different entries of the restored image x. For instance, we can regularize more on flat regions of the restored image and regularize less on the edges of the restored image. Here we solve the following least squares problem:

$$\min \left\| \begin{pmatrix} b \\ 0 \end{pmatrix} - \begin{pmatrix} A \\ \mu D \end{pmatrix} x \right\|_2, \tag{4.5}$$

where D is a diagonal matrix with each diagonal value is the amount of regularization to each pixel of the restored image x. Now we solve the following normal equations:

$$(\mu D + A^T A)x = A^T b. \tag{4.6}$$

In the test, the original image is the 256-by-256 satellite image shown in Figure 4.1(a). Figure 4.1(b) is the out-of-focus blurring function which corresponds to construct the blurring matrix $H_{m,n}$ or to form T_n. Here the diagonal value of D_n is computed in terms of the total variation of the first order difference of the image

$$[D_n]_{k,k} = \frac{1}{\sqrt{d_x^2(k) + d_y^2(k)} + 10^{-6}},$$

FIG. 4.1. *(a) The original satellite image; (b) the out-of-focus blurring function.*

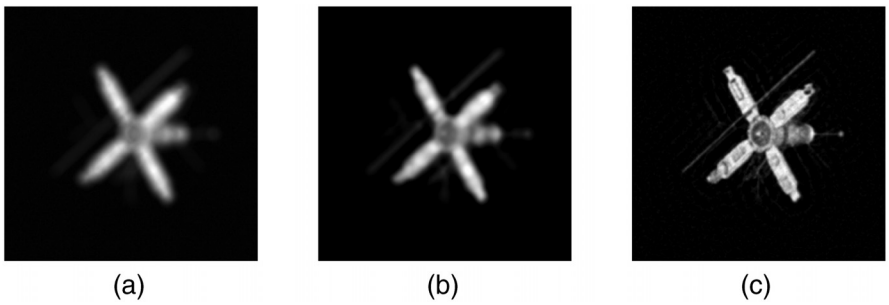

FIG. 4.2. *(a) The blurred and noisy image with* $\sigma = 0.01$*; (b) the restored image using the spatial invariant model; (c) the restored image using the spatial variant model with PCG(4) in Table* 4.6.

where $d_x(k)$ and $d_y(k)$ are the first order differences in the x-direction and y-direction at the k pixel location, respectively. Also the number 10^{-6} is used to avoid the zero appearing in the denumerator in the formula. It is clear that when the diagonal value is small (e.g., at the edge), the amount of regularization should be small; or when the diagonal value is large (e.g., at the flat region), the amount of regularization should be large. Figures 4.2, 4.3, and 4.4 are the restoring results for $\sigma = 0.01$, 0.05, and 0.1, respectively, where σ denotes the different levels of Gaussian white noise given by

$$\sigma = \frac{\|\hat{n}\|_2}{\|x^\star\|_2},$$

where $\hat{n}$ denotes the noise and $x^\star$ is the original image. In each figure, (a) refers to the blurred and noisy image, and (b) and (c) are the restored images obtained by solving (4.4) and (4.6), respectively. The regularization parameter μ in (4.4) or (4.6) is chosen to give the least relative error of the restored images to the original image. It is clear from the figures that the visual quality of the restored images in Figures 4.2(c), 4.3(c), and 4.4(c) by solving (4.6) is much better than that in Figures 4.2(b), 4.3(b), and 4.4(b) by solving (4.4).

The main aim of this experiment is to show the performance of the proposed preconditioner for the spatial variant system in (4.6). We solve (4.6) with different preconditioners, and the results are listed in Table 4.6. The stopping criterion of the

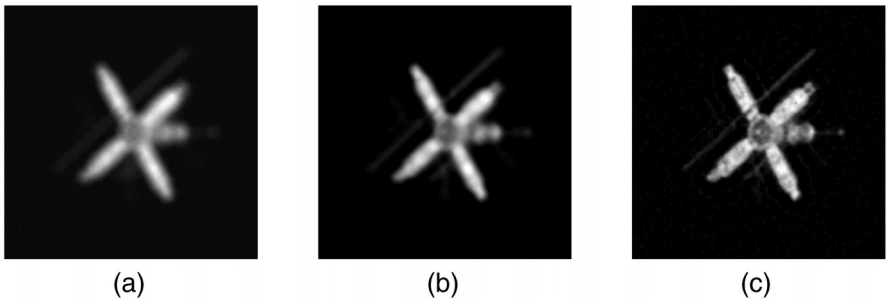

FIG. 4.3. *(a) The blurred and noisy image with $\sigma = 0.05$; (b) the restored image using the spatial invariant model; (c) the restored image using the spatial variant model with PCG(4) in Table* 4.6.

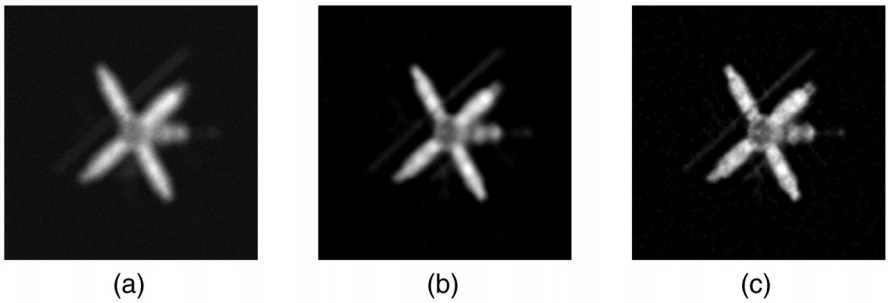

FIG. 4.4. *(a) The blurred and noisy image with $\sigma = 0.1$; (b) the restored image using the spatial invariant model; (c) the restored image using the spatial variant model with PCG(4) in Table* 4.6.

PCG method depends on the relative error of the computed image defined as follows:

$$\text{relerr}^{(k)} = \ln\left(\frac{\|x^{(k)} - x^\star\|_2}{\|x^\star\|_2}\right), \quad k = 1, 2, \ldots,$$

where $x^{(k)}$ is the k-th iterate of PCG iteration. Here we measure the difference between the computed image and the original image, and we expect that the smallest attained value of the PCG method will be controlled by the noise added to the image restoration problem. Therefore, we set the stopping criteria as -1.6, -1.3, and -1.2 for different noise levels in the table. We see from the table that when the proposed preconditioning technique is used, it converges very quickly. However, the convergence rate is very slow without using a preconditioner or with use of the T. Chan circulant preconditioner.

TABLE 4.6
Number of iterations required for different preconditioners for solving (4.6).

σ	0.01	0.05	0.1
CG	> 200	> 200	> 200
PCG(T)	> 200	> 200	> 200
PCG(2)	39	46	48
PCG(4)	7	9	11
PCG(8)	4	4	5
stopping criterion	$\text{relerr}^{(k)} < -1.6$	$\text{relerr}^{(k)} < -1.3$	$\text{relerr}^{(k)} < -1.2$

4.2. Concluding remarks. In summary, we considered the solutions of Hermitian Toeplitz-plus-diagonal systems $(T + D)x = b$, where T are Toeplitz matrices and D are diagonal matrices, and we studied approximate inverse circulant-plus-diagonal preconditioners for such systems. Both theoretical and numerical results show that the proposed preconditioner works quite well.

REFERENCES

[1] G. S. Ammar and W. B. Gragg, *Superfast solution of real positive definite Toeplitz systems*, SIAM J. Matrix Anal. Appl., 9 (1988), pp. 61–76.

[2] Z. Bai and M. Ng, *Preconditioners for nonsymmetric block-Toeplitz-like-plus-diagonal linear system*, Numer. Math., 96 (2003), pp. 197–220.

[3] M. Benzi and G. H. Golub, *Bounds for the entries of matrix functions with applications to preconditioning*, BIT, 39 (1999), pp. 417–438.

[4] M. Benzi and M. K. Ng, *Preconditioned iterative methods for weighted Toeplitz least squares problems*, SIAM J. Matrix Anal. Appl., 27 (2006), pp. 1106–1124.

[5] R. H. Chan and K.-P. Ng, *Fast iterative solvers for Toeplitz-plus-band systems*, SIAM J. Sci. Comput., 14 (1993), pp. 1013–1019.

[6] R. H. Chan and M. K. Ng, *Conjugate gradient methods for Toeplitz systems*, SIAM Rev., 38 (1996), pp. 427–482.

[7] R. Chan, M. Ng, and R. Plemmons, *Generalization of Strang's Preconditioner with Applications to Toeplitz Least Squares Problems*, Numer. Linear Algebra Appl., 3 (1996), pp. 45–64.

[8] F.-R. Lin, M. K. Ng, and W.-K. Ching, *Factorized banded inverse preconditioners for matrices with Toeplitz structure*, SIAM J. Sci. Comput., 26 (2005), pp. 1852–1870.

[9] G. H. Golub and C. Van Loan, *Matrix Computations*, 3rd ed., The Johns Hopkins University Press, Baltimore, MD, 1996.

[10] M. Ho and M. Ng, *Splitting iterations for circulant-plus-diagonal systems*, Numer. Linear Algebra Appl., 12 (2005), pp. 779–792.

[11] M. Hong, M. Kang, and A. Katsaggelos, *An iterative weighted regularized algorithm for improving the resolution of video sequences*, in Proceedings of the 1997 International Conference on Image Processing, Washington, DC, 1997, pp. 474–477.

[12] S. Jaffard, *Propriétés des matrices "bien localisées" près de leur diagonale et quelques applications*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 7 (1990), pp. 461–476.

[13] G. Meinardus, *Approximation of Functions: Theory and Numerical Methods*, Springer-Verlag, New York, 1967.

[14] J. Nagy, P. Pauca, R. Plemmons, and T. Torgersen, *Space-varying restoration of optical images*, J. Opt. Soc. Amer. A, 14 (1997), pp. 3162–3174.

[15] M. Ng, *Iterative Methods for Toeplitz Systems*, Oxford University Press, Oxford, UK, 2004.

[16] M. Ng and Z. Bai, *A hybrid preconditioner of banded matrix approximation and alternating direction implicit iteration for symmetric sinc-Galerkin linear systems*, Linear Algebra Appls., 366 (2003), pp. 317–335.

[17] M. Ng, S. Serra-Capizzano, and C. Tablino-Possio, *Multigrid methods for symmetric sinc-Galerkin systems*, Numer. Linear Algebra Appl., 12 (2005), pp. 261–269.

[18] S. Reeves, *Optimal space-varying regularization in iterative image restoration*, IEEE Trans. Image Process., 3 (1994), pp. 319–324.

[19] G. Strang, *A proposal for Toeplitz matrix calculations*, Stud. Appl. Math., 74 (1986), pp. 171–176.

[20] S. Serra, *How to choose the best iterative strategy for symmetric Toeplitz systems?*, SIAM J. Numer. Anal., 36 (1999), pp. 1078–1103.

[21] S. Serra-Capizzano and Cristina Tablino-Possio, *A Note on Multigrid Methods for (Multilevel) Structured-plus-banded Uniformly Bounded Hermitian Positive Definite Linear Systems*, preprint, http://arxiv.org/pdf/0804.3016v1.

[22] T. Strohmer, *Four short stories about Toeplitz matrix calculations*, Linear Algebra Appl., 343–344 (2002), pp. 321–344.

[23] W. F. Trench, *An Algorithm for the inversion of finite Toeplitz matrices*, SIAM J. Appl. Math., 12 (1964), pp. 515–522.