Fast Iterative Solvers for Toeplitz-plus-Band Systems

Raymond H. Chan and Kwok-Po Ng
Department of Mathematics
University of Hong Kong
Hong Kong

October 1991
Revised June 1992

Abstract. We consider the solutions of Hermitian Toeplitz-plus-band systems $(A_n + B_n)x = b$, where A_n are n-by-n Toeplitz matrices and B_n are n-by-n band matrices with band-width independent of n. These systems appear in solving integro-differential equations and signal processing. However, unlike the case of Toeplitz systems, no fast direct solvers have been developed for solving them. In this paper, we employ preconditioned conjugate gradient method with band matrices as preconditioners. We prove that if A_n is generated by a non-negative piecewise continuous function and B_n is positive semidefinite, then there exists a band matrix C_n, with band-width independent of n, such that the spectra of $C_n^{-1}(A_n + B_n)$ are uniformly bounded by a constant independent of n. In particular, we show that the solution of $(A_n + B_n)x = b$ can be obtained in $O(n \log n)$ operations.

Abbreviated Title. Toeplitz-plus-Band Systems

Key words. Toeplitz matrix, band matrix, generating function, preconditioned conjugate gradient method

AMS(MOS) subject classifications. 65F10, 65F15

1 Introduction

In this paper, we consider the solution of systems of the form $(A_n + B_n)x = b$, where A_n is an n-by-n Hermitian Toeplitz matrix (i.e. the entries of A_n are the same along its diagonals) and B_n is an n-by-n Hermitian band matrix with band-width independent of n. These systems appear in solving Fredholm integro-differential equations of the form

$$L\{x(\theta)\} + \int_{\alpha}^{\beta} K(\phi - \theta)x(\phi)d\phi = b(\theta).$$

Here $x(\theta)$ is the unknown function to be found, $K(\theta)$ is a convolution kernel, L is a differential operator and $b(\theta)$ is a given function. After discretization, K will lead to a Toeplitz matrix, L a band matrix and $b(\theta)$ the right hand side vector, see Delves and Mohamed [6, p.343]. Toeplitz-plus-band matrices also appear in signal processing literature and have been referred to as periheral innovation matrices, see Carayannis et. al. [2].

For Toeplitz systems $A_n x = b$, fast and superfast direct solvers of complexity $O(n^2)$ and $O(n \log^2 n)$ respectively have been developed, see for instance Trench [10] and Ammar and Gragg [1]. However, there exists no fast direct solvers for solving Toeplitz-plus-band systems. It is mainly because the displacement rank of the matrix $A_n + B_n$ 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 $(A_n + B_n)x$ can be computed in $O(n \log n)$ operations. In fact, $A_n x$ can be obtained by Fast Fourier Transform by first embedding A_n into a $2n$-by-$2n$ circulant matrix, see Strang [9]. 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 specturm of the matrix $A_n + B_n$, see Golub and van Loan [8]. However, in general, the specturm of A_n, and hence of $A_n + B_n$, is not clustered and the method will therefore converge slowly. Hence a suitable preconditioner should be chosen to speed up the convergence.

For Toeplitz systems $A_n x = b$, circulant preconditioners have been proved to be successful choices under the assumption that the diagonals of A_n are Fourier coefficients of a positive 2π-periodic continuous function. In that case, Chan and Yeung [3] proved that the convergence rate of the method

is superlinear. However, circulant preconditioners do not work for Toeplitz-plus-band systems. In fact, Strang's circulant preconditioner [9] is not even defined for non-Toeplitz matrices. T. Chan's circulant preconditioner, while defined for $A_n + B_n$, will not work well when the eigenvalues of B_n are not clustered, see the numerical results in §4. Even if we approximate A_n by a circulant preconditioner M_n, the matrix $M_n + B_n$ cannot be used as a preconditioner since the system $(M_n + B_n)z = y$ cannot be solved easily.

In this paper, we use band matrices C_n as preconditioners. We will assume that B_n is an arbitrary Hermitian positive semidefinite band matrix with band-width independent of n and the diagonals of A_n are Fourier coefficients of a non-negative piecewise continuous function f. In that case, $A_n + B_n$ will be Hermitian positive definite. We prove 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 band matrix C_n, with band-width independent of n, such that the spectra of $C_n^{-1}(A_n + B_n)$ 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 band matrix system $C_n x = b$ can be solved in $O(n)$ operations, the total complexity of the method is $O(n \log n)$.

The outline of the rest of the paper is as follows. In §2, we introduce our preconditioners C_n and study the spectral properties of $A_n + B_n$ and C_n. In §3, we show that the spectra of $C_n^{-1}(A_n + B_n)$ are uniformly bounded by constants independent of n. Finally, numerical examples and concluding remarks are given in §4.

2 Construction of the Preconditioner C_n

To begin with, let $\mathcal{C}^+$ be the set of all non-negative piecewise continuous functions defined on $[-\pi, \pi]$. For all $f \in \mathcal{C}^+$, let

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

be the Fourier coefficients of f. Since f is real-valued,

$$t_{-k}[f] = \bar{t}_k[f], \quad k = 0, \pm 1, \pm 2, \cdots.$$

Let $A_n[f]$ be the n-by-n Hermitian Toeplitz matrix with the (j,l)th entry given by $t_{j-l}[f]$. The function f is called the generating function of the matrices $A_n[f]$. We recall that a point θ_0 is said to be a zero of f with order ν if $f(\theta_0) = 0$ and ν is the smallest positive integer such that $f^{(\nu)}(\theta_0) \neq 0$ and $f^{(\nu+1)}(\theta)$ is continuous in a neighborhood of θ_0.

In the following, we denote the essential infimum and the essential supremum of f by $f_{\min}$ and $f_{\max}$ respectively. We will assume that f attains $f_{\min}$ at finitely many points in $[-\pi, \pi]$ and that f is smooth around these points. More precisely, we assume that $f(\theta) - f_{\min}$ has finitely many zeros in $[-\pi, \pi]$ and that the orders of these zeros are finite and positive. Notice that the matrix $A_n[f]$ is unchanged when f is redefined at finitely many points. Thus we can always assume without loss of generality that f is continuous at those minimum points.

From the assumptions, we see that $f_{\max} \neq f_{\min}$. Then by using the fact that

$$u^* A_n[g] u = \frac{1}{2\pi} \int_{-\pi}^{\pi} |\sum_{j=1}^{n} u_j e^{i(j-1)\theta}|^2 g(\theta) d\theta \tag{1}$$

for any $g \in \mathcal{C}^+$ and any n-vector $u = (u_1, \cdots, u_n)^*$, Chan [4, Lemma 1] proved that

$$\lambda_{\min}(A_n[f]) > f_{\min}. \tag{2}$$

Here $\lambda_{\min}(A_n[f])$ is the smallest eigenvalue of $A_n[f]$. Since f is non-negative, $A_n[f]$ is positive definite for all n. In the following, we will assume that the band matrices B_n are Hermitian positive semidefinite matrices with band-width $2\omega + 1$ and that ω is independent of n. Clearly the matrix $A_n[f] + B_n$ is positive definite for all n.

For all $n > 0$, our preconditioners C_n are defined as

$$C_n \equiv A_n[b_\mu] + B_n + f_{\min} \cdot I_n. \tag{3}$$

Here

$$b_\mu(\theta) = (2 - 2\cos\theta)^\mu = (2\sin(\frac{\theta}{2}))^{2\mu},$$

and it has a unique zero of order 2μ at $\theta = 0$. We remark that $A_n[b_\mu]$ is a symmetric band Toeplitz matrix of band-width $(2\mu+1)$ and its diagonals are given by the Pascal triangle, see Chan [4]. Clearly, C_n is a symmetric band matrix of band-width

$$2\ell + 1 = \max\{2\mu + 1, 2\omega + 1\}.$$

Moreover, since the minimum of b_μ is 0, it follows from (2) and (3) that

$$\lambda_{\min}(C_n) \geq \lambda_{\min}(A_n[b_\mu]) + \lambda_{\min}(B_n) + f_{\min} > f_{\min} \geq 0.$$

In particular, the preconditioner C_n is positive definite for all n. We note that in [4, Theorem 2], we have shown that $A_n[b_\mu(\theta)] + f_{\min} \cdot I_n$ is a good preconditioner for $A_n[f]$. Thus intuitively, we expect C_n so defined to be a good preconditioner for $A_n[f] + B_n$.

3 Condition Number of the Preconditioned Matrix

In this section, we show that the spectra of $C_n^{-1}(A_n + B_n)$ are uniformly bounded by constants independent of n. We first consider generating functions f in $\mathcal{C}^+$ where $f(\theta) - f_{\min}$ has only one zero at θ_0. Let the order of θ_0 be ν. We note that $f^{(\nu)}(\theta_0) > 0$ and ν must be even. We remark also that we can assume without loss of generality that $\theta_0 = 0$. In fact the function $f(\theta + \theta_0) - f_{\min}$ has a zero at $\theta = 0$ and

$$A_n[f(\theta + \theta_0)] = V_n^* A_n[f(\theta)] V_n,$$

where $V_n = \mathrm{diag}(1, e^{-i\theta_0}, e^{-2i\theta_0}, \cdots, e^{-i(n-1)\theta_0})$, see Chan [4, Lemma 2].

Theorem 1 *Let $f \in \mathcal{C}^+$. Suppose that $f(\theta) - f_{\min}$ has a unique zero at $\theta = 0$ with order equal to 2μ. Let $C_n = A_n[b_\mu] + B_n + f_{\min} \cdot I$. Then $\kappa(C_n^{-1}(A_n[f] + B_n))$ is uniformly bounded for all $n > 0$.*

Proof: By assumption, there exists a neighborhood N of 0 such that f is continuous in N. Define

$$F(\theta) = \frac{f(\theta)}{(2 - 2\cos\theta)^\mu + f_{\min}}.$$

Clearly F is continuous and positive for $\theta \in N \setminus \{0\}$. Since

$$\lim_{\theta \to 0} F(\theta) = \begin{cases} 1 & f_{\min} > 0, \\ \dfrac{f^{(2\mu)}(0)}{(2\mu)!} & f_{\min} = 0, \end{cases}$$

is positive, F is a continuous positive function in N. Since f is piecewise continuous and positive almost everywhere in $[-\pi,\pi]\setminus N$, we see that F is a piecewise continuous function with a positive essential infimum in $[-\pi,\pi]$. Hence there exist constants $b_1, b_2 > 0$, such that $b_1 \le F(\theta) \le b_2$ almost everywhere in $[-\pi,\pi]$. Without loss of generality, we assume that $b_2 \ge 1 \ge b_1$. By using (1), we then have

$$b_1 \le \frac{u^* A_n[f] u}{u^*(A_n[b_\mu] + f_{\min}\cdot I_n)u} \le b_2$$

for any n-vector u. Recall that B_n is positive semidefinite and $C_n = A_n[b_\mu] + B_n + f_{\min}\cdot I_n$, we then have

$$b_1 \le \frac{u^*(A_n[f] + B_n)u}{u^* C_n u} \le b_2,$$

for any n-vector u. Hence $\kappa(C_n^{-1}(A_n[f]+B_n)) \le b_2/b_1$, which is independent of n. $\square$

We remark that the results can be readily generalized to the case where $f_{\min}$ is attained at finitely many points, cf. Chan [4, Theorem 3]. The band-width of C_n will be given by

$$2\ell + 1 = \max\{\sum_j \nu_j + 1, 2\omega + 1\},$$

where ν_j are the orders of the zeros of $f(\theta) - f_{\min}$ and the summation is over all such zeros.

Next we consider the computational cost and storage requirement of our method. The number of operations per iteration in the preconditioned conjugate gradient method depends mainly on the work of computing the matrix-vector multiplication $C_n^{-1}(A_n[f]+B_n)y$, see for instance Golub and van Loan [8]. In this case, the matrix-vector multiplication $B_n y$ requires only $(2\omega+1)n$ operations and the product $A_n[f]y$ can be done in $O(n\log n)$ operations by the Fast Fourier Transform. The system $C_n y = z$ can be solved by using any band matrix solver. The cost of factorizing C_n is about $\frac{1}{2}\ell^2 n$ operations and then each subsequent solve requires an extra $(2\ell+1)n$ operations. Hence the total operations per iteration is of order $O(n\log n)$ as ℓ and ω are independent of n. It is well-known that the number of iterations required to attain

a given tolerance ϵ is bounded by

$$\frac{1}{2}\sqrt{\kappa(C_n^{-1}(A_n+B_n))}\log(\frac{1}{\epsilon})+1.$$

Since the condition number is uniformly bounded in this case, the overall work required to attain the given tolerance is of $O(n\log n)$ operations.

As for the storage, we need five n-vectors in the conjugate gradient method. The diagonals of A_n and the bands of B_n require extra $(\omega+2)$ n-vectors, and finally, we need an n-by-$(\ell+1)$ matrix to hold the factors of the preconditoner C_n. Thus the overall storage requirement is about $(8+\ell+\omega)n$, which is significantly less than the $O(n^2)$ storage required by Gaussian elimination method.

4 Numerical Results and Concluding Remarks

To test the convergence rate of the preconditioner, we considered two different band matrices. The first one is the diagonal matrix

$$D_n = f_{\max}\cdot \mathrm{diag}[0,\frac{1}{n},\frac{2}{n},\cdots,\frac{n-1}{n}]$$

whose eigenvalues are distributed uniformly in the interval $[0, f_{\max}]$. The second one is a symmetric tridiagonal matrix given by

$$B_n^{(\alpha)} = (n+1)^\alpha\cdot\frac{2\pi}{n+1}\begin{bmatrix} 2 & -\frac{3}{2} & & & \\ -\frac{3}{2} & 4 & -\frac{5}{2} & & \\ & -\frac{5}{2} & 6 & \ddots & \\ & & \ddots & \ddots & -\frac{2n-1}{2} \\ & & & -\frac{2n-1}{2} & 2n \end{bmatrix}.$$

Notice that $B_n^{(2)}$ is the discretization matrix of the operator

$$\frac{d}{d\theta}\left\{(\theta+\pi)\frac{d}{d\theta}\right\}$$

in $[-\pi,\pi]$. Clearly, the matrices $B_n^{(\alpha)}$ are irreducibly diagonally dominant, hence they are positive definite.

In our tests, the vector of all ones is the right hand side vector, the zero vector is the initial guess and the stopping citerion is $||r_q||_2/||r_0||_2 \leq 10^{-7}$, where r_q is the residual vector after q iterations. The computation are done with 8-byte arithmetic on a Vax 6420. Three different generating functions were tested. They are θ^4, $\cosh\theta$ and

$$J(\theta) \equiv \begin{cases} \theta^2 & |\theta| \leq \pi/2, \\ 1 & |\theta| > \pi/2. \end{cases}$$

The corresponding band-widths of C_n are 5, 3 and 3 respectively.

For comparison, we also solved the problems with two other preconditioners. The first one is the T. Chan circulant preconditioner T_n corresponding to the matrix $A_n[f]+B_n$. The second preconditioner E_n, which has the same band-width as C_n, is obtained by just copying the diagonals of $A_n[f]+B_n$. We note that some of the E_n may be indefinite. In contrast, C_n and T_n are always positive definite. Tables 1 to 4 show the number of iterations required for convergence (** means more than 1000 iterations). We see that as n increases, the number of iterations stays almost the same when C_n is used as the preconditioner while it increases if others are used.

f	θ^4				$\cosh\theta$				$J(\theta)$			
n	No	C_n	E_n	T_n	No	C_n	E_n	T_n	No	C_n	E_n	T_n
16	16	9	8	16	15	8	8	15	14	12	9	14
32	26	11	9	23	21	9	9	18	18	14	12	16
64	36	12	11	31	25	9	10	21	23	14	15	19
128	50	14	16	40	29	10	11	23	30	15	18	24
256	68	15	21	53	32	10	12	25	39	15	23	30
512	91	15	32	70	34	10	14	27	50	15	28	38
1024	122	16	64	91	36	10	16	27	63	15	35	47

Table 1: Number of Iterations for $B_n = D_n$.

f	θ^4				$\cosh\theta$				$J(\theta)$			
n	No	C_n	E_n	T_n	No	C_n	E_n	T_n	No	C_n	E_n	T_n
16	17	12	15	16	16	7	9	14	16	9	9	16
32	42	15	35	23	31	8	13	16	35	10	14	27
64	107	17	98	32	38	9	18	17	75	12	23	35
128	268	19	372	45	42	9	30	17	162	14	40	42
256	652	21	**	63	43	9	48	17	329	16	75	51
512	**	22	**	90	43	10	82	17	670	17	146	61
1024	**	23	**	127	43	10	146	16	**	18	293	74

Table 2: Number of Iterations for $B_n = B_n^{(0)}$

f	θ^4				$\cosh\theta$				$J(\theta)$			
n	No	C_n	E_n	T_n	No	C_n	E_n	T_n	No	C_n	E_n	T_n
16	16	8	8	16	16	5	6	16	16	5	5	16
32	37	8	10	31	36	5	7	31	36	5	6	32
64	82	8	13	47	82	5	8	46	83	5	8	51
128	188	8	18	70	184	5	9	65	189	5	9	77
256	415	8	23	104	408	5	12	91	418	5	11	112
512	893	8	31	152	826	5	13	130	896	5	14	164
1024	**	8	40	220	**	5	16	183	**	5	17	238

Table 3: Number of Iterations for $B_n = B_n^{(1)}$

f	θ^4				$\cosh\theta$				$J(\theta)$			
n	No	C_n	E_n	T_n	No	C_n	E_n	T_n	No	C_n	E_n	T_n
16	16	4	5	17	16	3	4	16	16	3	3	16
32	37	4	5	32	37	3	4	32	37	3	4	32
64	83	4	5	52	83	3	4	52	83	3	4	52
128	189	3	5	77	190	3	4	77	190	3	4	77
256	418	3	5	113	417	3	4	114	418	3	4	113
512	897	3	5	166	897	2	4	165	898	2	4	166
1024	**	3	5	241	**	2	4	240	**	2	4	241

Table 4: Number of Iterations for $B_n = B_n^{(2)}$

We conclude that our algorithm solves the system $(A_n + B_n)x = b$ in $O(n \log n)$ operations for a certain class of Toeplitz matrices A_n. The cost is significantly less than the $O(n^3)$ cost required by Gaussian elimination method. We note that the spectra of $C_n^{-1}(A_n[f] + B_n)$ in general will not be clustered around 1 although they are uniformly bounded. We finally remark that our results in this paper extend those obtained in Chan [4]. More precisely, in [4], we proved that $\kappa(C_n^{-1}A_n[f])$ is uniformly bounded whenever f is 2π-periodic continuous. However, using Theorem 1 with B_n equal to a zero matrix, we see that the same conclusion holds whenever f is 2π-periodic piecewise continuous.

Acknowledgements: We would like to thank Professor Israel Koltrach of University of Connecticut for his valuable suggestions.

References

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

[2] G. Carayannis, N. Kalouptsidis and D. Manolakis, *Fast Recursive Algorithms for a Class of Linear Equations*, IEEE Trans. Acoust. Speech Signal Process., V30 (1982), pp. 227-239.

[3] R. Chan and M. Yeung, *Circulant Preconditioners for Toeplitz Matrices with Positive Continuous Generating Functions*, Math. Comp., to appear.

[4] R. Chan, *Toeplitz Preconditioners for Toeplitz Systems with Nonnegative Generating Functions*, IMA J. Numer. Anal., V11 (1991), pp. 333-345.

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

[6] L. Delves and J. Mohamed, *Computational Methods for Integral Equations,* Cambridge University Press, Cambridge, 1985.

[7] Grenander and Szegö, *Toeplitz Form and Its Applications*, 2nd Ed., Chelsea Pub. Co., New York, 1984.

[8] G. Golub and C. van Loan, *Matrix Computations*, 2nd Ed., The Johns Hopkins University Press, Baltimore, 1989.

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

[10] W. Trench, *An Algorithm for the Inversion of Finite Toeplitz Matrices*, SIAM J. Appl. Math., V12 (1964), pp. 515-522.