Selçuk J. Appl. Math.
Vol. 4, No. 2, pp. 71–88, 2003

Selçuk Journal of Applied Mathematics

Using ω-circulant matrices for the preconditioning of Toeplitz systems

Rainer Fischer and Thomas Huckle

Institute of Informatics, Technical University of Munich, Boltzmannstrasse 3, 95748 Garching, Gemany;
e-mail: `fischerr@in.tum.de`; e-mail: `huckle@in.tum.de`

Received: September 5, 2003

Summary. Toeplitz systems can be solved efficiently by using iterative methods such as the conjugate gradient algorithm. If a suitable preconditioner is used, the overall cost of the method is $O(n \log n)$ arithmetic operations. Circulant matrices are frequently employed for the preconditioning of Toeplitz systems. They can be chosen as preconditioners themselves, or they can be used for the computation of approximate inverses. In this article, we take the larger class of ω-circulant matrices instead of the well-known circulants to extend preconditioners of both types. This extension yields an additional free parameter ω which can be chosen in a way that speeds up convergence of the conjugate gradient method. The additional computational effort arising from the use of ω-circulant instead of circulant matrices is low.

Key words: circulant matrices, Toeplitz systems, preconditioning

2000 Mathematics Subject Classification: 65F10, 65F22

1. Introduction

Toeplitz matrices arise in a variety of applications, for example in the discretization process of partial differential equations. Since Toeplitz matrices are dense, but very structured matrices, this structure must be exploited by any solver, no matter whether it is direct or iterative. Until 1985 mostly direct Toeplitz solvers were developed [1], the best of these methods having a total cost of $O(n \log^2 n)$ operations.

Strang [5] was the first to develop a competitive iterative method for Hermitian positive definite Toeplitz matrices. He used the conjugate gradient algorithm, which requires only $O(n \log n)$ operations per iteration. If the number of iterations is low, this is, for large n, faster than the best direct methods. In most cases, fast convergence can only be achieved if a suitable preconditioner is used. Many efficient preconditioners for Toeplitz systems are either circulant matrices or they are constructed with the help of circulant matrices.

This paper is organized as follows. In Chapter 2 we review essential properties of Toeplitz matrices, circulant matrices, and the conjugate gradient method. Chapter 3 presents two classes of preconditioners for the conjugate gradient method which are based on circulant matrices: circulant preconditioners and approximate inverse preconditioners. In Chapter 4 we extend three of these preconditioners using ω-circulant matrices, and carry out extensive numerical tests to find out how the new preconditioners work in practice.

2. Toeplitz systems and circulant matrices

Definition 1. *An n-by-n matrix T_n is called Toeplitz if it is constant along its diagonals, i.e. if*

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

Its entries are given by $T_n^{(l,m)} = t_{l-m}$. In order to derive some essential properties of Toeplitz matrices we need to introduce the concept of a generating function.

Definition 2. *Let f be a 2π-periodic real-valued function defined on $[-\pi, \pi]$. The Fourier coefficients of f are given by*

$$t_k = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(\theta) e^{-ik\theta} d\theta \qquad (k \in \mathbb{Z}) .$$

We can now define the sequence of matrices $\{T_n(f)\}_n$, where T_n is the n-by-n Toeplitz matrix with entries $T_n^{(j,k)} = t_{j-k}$ $(0 \le j, k < n)$. f is called the generating function of the sequence $(T_n)_n$.

Since f is real-valued, the matrices T_n are Hermitian. If in addition f is even, the T_n are real symmetric and f can be represented by a cosine series. Grenander and Szegö [4] proved that all eigenvalues of a Toeplitz matrix are contained in the range of its generating function $[f_{min}, f_{max}]$ and that, for $\lim_{n\to\infty}$, the extreme eigenvalues tend to f_{min} and f_{max}.

The immediate consequence of this theorem is that a positive function f leads to a sequence of positive definite Toeplitz matrices $\{T_n(f)\}_n$. If, however, $f_{min} = 0$, the $T_n(f)$ are ill-conditioned for large n. In [9] it is shown that a zero of order 2ν in f lets the condition numbers of the $T_n(f)$ grow like $O(n^{2\nu})$.

Circulant matrices are a subclass of Toeplitz matrices, which plays an essential role in general Toeplitz matrix calculations.

Definition 3. *An n-by-n matrix C_n is called circulant if it is Toeplitz and, in addition, $c_{-k} = c_{n-k}$.*

The following theorem states that circulant matrices can be diagonalized efficiently. For a proof see [3].

Theorem 1. *A circulant matrix C_n has the decomposition $C_n = F_n^H \Lambda_n F_n$, where Λ_n is the diagonal matrix containing the eigenvalues of C_n, and F_n is the Fourier matrix, which is unitary.*

Theorem 1 implies that many computations involving circulant matrices can be done in $O(n \log n)$ operations with the Fast Fourier Transform (FFT).

Block-Toeplitz-Toeplitz-block (BTTB) matrices or two-level Toeplitz matrices are the two-dimensional analogues of Toeplitz matrices. A BTTB matrix is a block matrix with Toeplitz blocks, also having Toeplitz structure on the block level. The spectrum of the BTTB matrix is bounded by the range of the corresponding generating function, which, in this case, is a function in two variables. For large n the maximum and minimum eigenvalues of the matrix tend to the maximum and minimum values of the function. Block-circulant-circulant-block (BCCB) matrices are the two-dimensional analogues of circulant matrices. They are circulant within each block and on the block level. BCCB matrices are diagonalized efficiently by the two-dimensional FFT.

Toeplitz systems are efficiently solved with the conjugate gradient (cg) method. The cg method is a non-stationary iterative method for the solution of Hermitian positive definite matrix systems [10]. In [11] it is shown that fast convergence is reached if the eigenvalues of the matrix T_n are clustered around 1 for large n. Since this is not the case for most Toeplitz systems, a preconditioner P_n must be chosen in such a way that the clustering property holds for the preconditioned system $P_n^{-1}T_n$. Furthermore, all computations involving the preconditioner, e.g. the construction of P_n or the solution of a linear system $P_nh = r$ must be carried out in $O(n \log n)$ operations.

3. Preconditioning with circulant matrices

There are two fundamentally different principles for the construction of a preconditioner which is to be used in the preconditioned conjugate gradient (pcg) method. One way is to find an approximation P_n to the given Toeplitz matrix T_n, and then to solve the system $P_nh = r$ in each iteration. The other principle is to find an approximation M_n to T_n^{-1}, and then to compute h by a matrix-vector multiplication $h = M_nr$.

Since most calculations involving circulant matrices can be carried out in $O(n \log n)$ operations, this class of matrices is well-suited for the construction of preconditioners. Circulant matrices can be chosen as preconditioners themselves, representing the first principle of construction, or they can be used for the construction of approximations to T_n, representing the second principle.

3.1. Circulant preconditioners

The first circulant preconditioner for Toeplitz systems was given by Strang [5].

Definition 4. *Let T_n be an n-by-n Toeplitz matrix defined in (1). Then the diagonals s_j of Strang's preconditioner $S_n = [s_{k-l}]_{0\le k,l<n}$ are defined by*

$$s_j = \begin{cases} t_j, & 0 \le j \le \lfloor n/2 \rfloor, \\ t_{j-n}, & \lfloor n/2 \rfloor < j < n, \\ s_{n+j}, & 0 < -j < n. \end{cases}$$

T. Chan [2] developed the so called optimal circulant preconditioner $c_F(T_n)$.

Definition 5. *Let T_n be an n-by-n Toeplitz matrix. Then the diagonals c_j of T. Chan's preconditioner $c_F(T_n) = [s_{k-l}]_{0\le k,l<n}$ are defined by*

$$c_j = \begin{cases} \frac{(n-j)t_j + jt_{j-n}}{n}, & 0 \le j \le n-1, \\ c_{n+j}, & 0 < -j < n-1. \end{cases} \tag{2}$$

In [2] it is shown that $c_F(T_n)$ minimizes $\|C_n - T_n\|_F$ over all circulant matrices C_n, where $\|\cdot\|_F$ denotes the Frobenius norm.

The optimal preconditioner is extended to BTTB matrices T_{mn} by T. Chan and Olkin [14]. In this case the BCCB matrix C_{mn}^F minimizing $\|C_{mn} - T_{mn}\|_F$ over all BCCB matrices C_{mn} is used as a preconditioner. It is calculated in two steps. First, T. Chan's preconditioner is computed for each block of T_{mn}, and then (2) is applied to the resulting matrix on the block level.

3.2. Approximate inverse preconditioners

Hanke and Nagy [7] developed an approximate inverse preconditioner which is based on embedding the given Toeplitz matrix into a larger circulant matrix, which can be inverted in $O(n \log n)$ with the FFT. Let T_n be a banded n-by-n Hermitian positive definite Toeplitz matrix. Then T_n is embedded into the $(n+\beta)$-by-$(n+\beta)$ circulant matrix

$$C_{n+\beta} = \begin{pmatrix} T_n & T_{2,1}^H \\ T_{2,1} & T_{2,2} \end{pmatrix} . \tag{3}$$

$C_{n+\beta}$ can be diagonalized with the help of the Fourier matrix $F_{n+\beta}$. In the decomposition $C_{n+\beta} = F_{n+\beta}^H \Lambda_{n+\beta} F_{n+\beta}$, $\Lambda_{n+\beta}$ contains the eigenvalues of $C_{n+\beta}$. If $C_{n+\beta}$ is positive definite, and therefore all eigenvalues λ_j are positive, the inverse can be computed by

$$C_{n+\beta}^{-1} = \begin{pmatrix} M_n & M_{1,2} \\ M_{2,1} & M_{2,2} \end{pmatrix} = F_{n+\beta}^H \Lambda_{n+\beta}^{-1} F_{n+\beta}.$$

However, if $C_{n+\beta}$ has nonpositive eigenvalues, Hanke and Nagy use the matrix $\Lambda_{n+\beta}^-$ instead of $\Lambda_{n+\beta}^{-1}$, where $\Lambda_{n+\beta}^-$ is the diagonal matrix with entries

$$\lambda_j^- = \begin{cases} 1/\lambda_j, & \text{if } \lambda_j > 0; \\ 0, & \text{if } \lambda_j \le 0. \end{cases} \tag{4}$$

This leads to the following approximation for $C_{n+\beta}^{-1}$:

$$C_{n+\beta}^{-} = \begin{pmatrix} M_n & M_{1,2} \\ M_{2,1} & M_{2,2} \end{pmatrix} = F_{n+\beta}^H \Lambda_{n+\beta}^{-} F_{n+\beta}.$$

The leading n-by-n principal submatrix M_n of $C_{n+\beta}^{-1}$ or $C_{n+\beta}^{-}$ is used as an approximation for T_n. Hanke and Nagy [7] proved a clustering result for the preconditioned system, which will be extended in Section 4.2.

4. Extending the preconditioners with ω-circulant matrices

In the previous chapter we described some of the well-known preconditioners for the solution of Toeplitz systems with the cg method, which were either circulant themselves or constructed with the help of circulant matrices. In this paper we wish to design new preconditioners by using the larger class of ω-circulant matrices instead of the circulants. The following definition can be found for example in [1].

Definition 6. *Let $\omega = e^{i\theta}$ with $\theta \in [-\pi, \pi]$. An n-by-n matrix W_n is said to be ω-circulant if it has the spectral decomposition*

$$W_n = \Omega_n F_n^H \Lambda_n F_n \Omega_n^H = \Omega_n C_n \Omega_n^H . \tag{5}$$

F_n is the Fourier matrix, Λ_n is diagonal containing the eigenvalues of W_n, $\Omega_n = diag(1, \omega^{1/n}, \ldots, \omega^{(n-1)/n})$, and C_n denotes the circulant matrix from Theorem 1.

If we choose $\theta = 0$ in Definition 6, $\omega = 1$ and W_n is circulant. Although the class of ω-circulant matrices is slightly more general than the class of circulant matrices, most calculations involving ω-circulants such as matrix-vector products or the solution of linear systems can also be carried out in $O(n \log n)$ operations. This is due to the fact that diagonalization of an ω-circulant matrix requires, in addition to the FFT, only one matrix-vector multiplication involving the diagonal matrix Ω_n.

Since the additional computational effort arising from the use of ω-circulant matrices is low, we try to extend the preconditioners described in Chapter 3 by using ω-circulant matrices instead of circulants. Then, the choice of θ yields an extra degree of freedom which can be used to improve the performance of the preconditioner. In the first part of this chapter we choose θ in order to minimize a norm, whereas in the subsequent sections θ improves the rank of the circulant extension matrix.

4.1. Extending the optimal circulant preconditioner

In the first part of this section we develop an ω-circulant extension of the preconditioner of T. Chan, whereas in the second part we extend its two-dimensional analogue.

4.1.1. Extending the preconditioner of T. Chan Following the idea of Huckle [6] we seek to minimize

$$\|C_n(\omega) - T_n\|_F$$

over all ω-circulant matrices $C_n(\omega)$. Since $C_n(\omega)$ has the decomposition $C_n(\omega) = \Omega_n C_n \Omega_n^H$ with a circulant matrix C_n, and since multiplication by a unitary matrix does not change the Frobenius norm, the minimization problem becomes

$$\min_{C_n \; circulant} \|\Omega_n C_n \Omega_n^H - T_n\|_F \;=\; \min_{C_n \; circulant} \|C_n - T_n(\omega)\|_F \tag{6}$$

with $T_n(\omega) := \Omega_n^H T_n \Omega_n$. From (6) the strategy for computing the optimal ω-circulant preconditioner $c_F^{\omega}(T_n)$ becomes clear. After choosing the optimal ω and calculating $T_n(\omega)$, we compute the optimal circulant preconditioner $c_F(T_n(\omega))$ for the Toeplitz matrix $T_n(\omega)$, minimizing the Frobenius norm over all circulant matrices. Finally, $c_F^{\omega}(T_n)$ is determined by $c_F^{\omega}(T_n) = \Omega_n\, c_F(T_n(\omega))\Omega_n^H$. The only remaining question is, how can the optimal ω be found? From $c_F(T_n(\omega))$, $T_n(\omega)$, and (6) we can derive a formula for the optimal ω. Since

$$\|c_F^{\omega}(T_n) - T_n\|_F = \|c_F(T_n(\omega)) - T_n(\omega)\|_F,$$

ω is the solution of the minimization problem

$$\min_{\omega} \|c_F(T_n(\omega)) - T_n(\omega)\|_F. \tag{7}$$

After computing

$$\begin{aligned} \|c_F(T_n(\omega)) - T_n(\omega)\|_F^2 &= \tfrac{1}{n}\sum_{j=1}^{n-1}(n-j)j|t_{-j}|^2 \\ &+\tfrac{1}{n}\sum_{j=1}^{n-1}(n-j)j|t_j|^2 - \tfrac{2}{n}\, Re(\omega \sum_{j=1}^{n-1}(n-j)j\,\overline{t_{-j}}\, t_{n-j}), \end{aligned} \tag{8}$$

(7) is solved as a one-dimensional real minimization problem in the argument θ of $\omega = e^{i\theta}$. The result is

$$\theta \;=\; -\arg\left(\sum_{j=1}^{n-1}(n-j)j\,\overline{t_j}\, t_{j-n}\right) + 2k\pi \;(k \in \mathbb{Z})\,.$$

The clustering property for the optimal ω-circulant preconditioner can be proved in the same way as for the optimal circulant preconditioner. Carrying over the results of Chan and Yeung [8], [12] leads to the following result.

Theorem 2. *Let f be a 2π-periodic continuous positive function with the associated sequence of Toeplitz matrices $\{T_n\}_n$. Moreover, let $c_F^\omega(T_n)$ be the optimal ω-circulant preconditioner for T_n. Then, the spectra of $c_F^\omega(T_n)^{-1}T_n$ are clustered around 1 for large n.*

To find out whether the optimal ω-circulant preconditioner is a real improvement, we start with the following observation. For the matrices $T_n = tridiag(-1, 2, -1)$ of the discrete one-dimensional Laplacian $\|c_F(T_n(\omega)) - T_n(\omega)\|_F$ is independent of θ. This observation is just a special case of the following result on banded Toeplitz matrices, which follows directly from (8).

Theorem 3. *Let T_n be a banded Toeplitz matrix with bandwidth $\beta < n/2$. Then*

$$\|R_n\|_F^2 := \|c_F^\omega(T_n) - T_n\|_F^2 = \|c_F(T_n(\omega)) - T_n(\omega)\|_F^2$$

is independent of ω, and therefore the same as $\|c_F(T_n) - T_n\|_F^2$ for the optimal circulant preconditioner of T. Chan.

For non-banded Toeplitz matrices T_n, a suitable choice of ω leads to an improvement of $\|R_n\|_F^2$, which in many cases yields far better results of the pcg method. For example, if a Toeplitz matrix is closely related to a skew-circulant matrix, the use of a skew-circulant preconditioner not only minimizes the Frobenius norm, but also leads to faster convergence. This can be illustrated by the following example. It shows how $\|c_F^\omega(T_n) - T_n\|_F$ changes when we move from a circulant to a skew-circulant matrix T_n. It is well known that each Toeplitz matrix can be written as the sum of a circulant and a skew-circulant matrix.

Example 1. Let A_n be the symmetric positive definite Toeplitz matrix given by $a_k = \frac{1}{k+1}$ $(0 \le k < n)$. Then A_n has the decomposition $A_n = C_n + S_n$ with the circulant matrix C_n and the skew-circulant matrix S_n, where $c_0 = s_0 = a_0/2$, $c_k = a_k + a_{k-n}$, and $s_k = a_k - a_{k-n}$. With C_n and S_n we can define the Toeplitz matrices

$$T_n = p \cdot C_n + (2-p) \cdot S_n$$

with the parameter $p \in [0, 2]$. For $p = 0$, T_n is skew-circulant, whereas for $p = 2$, it is circulant. The closer p is to 0, the closer T_n is related to a skew-circulant matrix. For larger p, T_n becomes more and more circulant. Figure 1 depicts $\|c_F^\omega(T_n) - T_n\|_F$ for different values of p showing that for matrices which are dominated by the circulant component the Frobenius norm has its minimum at 0, whereas skew-circulant dominance leads to a minimum at π.

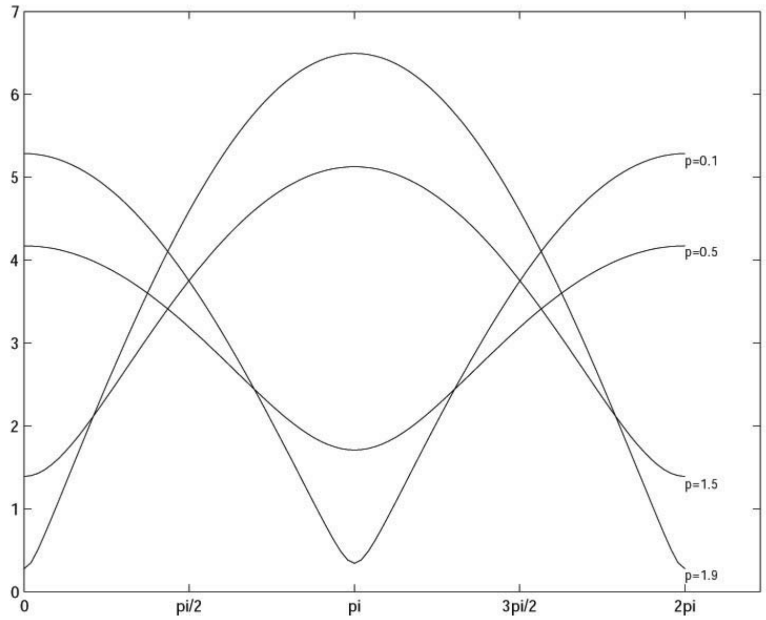

Fig. 1. $\|c_F^\omega(T_n) - T_n\|_F$ depending on θ for the matrices T_n from Example 1 with $p = 0.1, 0.5, 1.5, 1.9$ and $n = 1000$

Not only the Frobenius norm is improved by a suitable choice of θ, but also the performance of the pcg method. The table 1 summarizes the numerical results.

	p=0.1		p=0.5		p=1.5		p=1.9	
n	θ=0	θ=π	θ=0	θ=π	θ=0	θ=π	θ=0	θ=π
5000	9	5	8	7	6	9	5	9
10000	9	5	8	7	6	9	5	9
15000	9	5	9	7	6	9	5	10
20000	9	5	9	7	6	9	5	10

Table 1

4.1.2. Extending the preconditioner of Chan and Olkin Now we wish to carry over the results of the previous paragraph to the block case. For the preconditioning of BTTB systems T_{mn} we extend the preconditioner of T. Chan and Olkin by allowing two free parameters α and ω. On the first level of approximation, each block of T_{mn} is substituted by an α-circulant matrix instead of a circulant. In each block, the first element of the j-th row is obtained by multiplying the last element of the $(j-1)$-th row by α. On the second level of approximation, i.e. on the block level, we replace the circulant structure by an ω-circulant structure. This means the first block of the second block row is obtained by multiplying the last block in the first block row by ω. The goal is to minimize

$$\|C_{mn}(\alpha,\omega) - T_{mn}\|_F \tag{9}$$

over all block ω-circulant matrices with α-circulant blocks. This is done in a similar way as it was done for Toeplitz matrices. $C_{mn}(\alpha,\omega)$ has the decomposition

$$C_{mn}(\alpha,\omega) \; = \; \Omega_{mn} C_{mn} \Omega_{mn}^H \quad ,$$

where C_{mn} is a BCCB matrix, and $\Omega_{mn} = \Omega_m \otimes \Gamma_n$ with

$$\Omega_m = diag(1, \omega^{\frac{1}{m}}, \ldots, \omega^{\frac{m-1}{m}}), \; \Gamma_n = diag(1, \alpha^{\frac{1}{n}}, \ldots, \alpha^{\frac{n-1}{n}}) \; .$$

The two free parameters are defined as $\omega = e^{i\Phi}$ and $\alpha = e^{i\Psi}$. The matrix Ω_{mn} is a diagonal matrix of the form

$$\Omega_{mn} = diag(1, \alpha^{\frac{1}{n}}, \ldots, \alpha^{\frac{n-1}{n}}, \omega^{\frac{1}{m}} 1, \omega^{\frac{1}{m}} \alpha^{\frac{1}{n}}, \ldots,$$

$$\omega^{\frac{1}{m}} \alpha^{\frac{n-1}{n}}, \ldots, \omega^{\frac{m-1}{m}} 1, \omega^{\frac{m-1}{m}} \alpha^{\frac{1}{n}}, \ldots, \omega^{\frac{m-1}{m}} \alpha^{\frac{n-1}{n}}) \; .$$

With this notation, (9) can be rewritten as

$$\min_{C_{mn} \in BCCB} \|\Omega_{mn} C_{mn} \Omega_{mn}^H - T_{mn}\|_F = \min_{C_{mn} \in BCCB} \|C_{mn} - T_{mn}(\alpha,\omega)\|_F .$$

with $T_{mn}(\alpha,\omega) := \Omega_{mn}^H T_{mn} \Omega_{mn}$. This leads to the same strategy for computing the optimal block ω-circulant matrix with α-circulant blocks $C_{mn}^{\alpha,\omega}$ as in the one-dimensional case. The Frobenius norm of $R_{mn} := T_{mn}(\alpha,\omega) - C_{mn}^{(2)}$ with the optimal BCCB approximation $C_{mn}^{(2)}$ for T_{mn} has the form

$$\|R_{mn}\|_F^2 = c_0 + c_1\alpha + \overline{c_1\alpha} + c_2\omega + \overline{c_2\omega} + c_3\alpha\omega$$

$$+\overline{c_3\alpha\omega} + c_4\overline{\alpha}\omega + \overline{c_4}\alpha\overline{\omega} \tag{10}$$

$$= c_0 + 2Re(c_1\alpha) + 2Re(c_2\omega) + 2Re(c_3\alpha\omega) + 2Re(c_4\overline{\alpha}\omega),$$

where the parameters $c_0, \ldots, c_4$, which are independent of α and ω, can be computed in $O(mn)$.

We can now derive a similar result for BTTB matrices which are banded either within each block or on the block level.

Theorem 4. *Let T_{mn} be a BTTB matrix with blocks of size n-by-n, and $R_{mn} = T_{mn}(\alpha, \omega) - C_{mn}^{(2)}$. Moreover, let β be the maximum bandwidth over all blocks T_j, and γ the bandwidth on the block level, i.e. the smallest positive integer j such that T_j is different from the zero matrix only for $j \leq \gamma$.*

1. *If $\beta < \frac{n}{2}$, i.e. if each block of T_{mn} is banded, $\|R_{mn}\|_F$ does not depend on α.*
2. *If $\gamma < \frac{m}{2}$, i.e. if T_{mn} is banded on the block level, $\|R_{mn}\|_F$ does not depend on ω.*
3. *If $\beta < \frac{n}{2}$ and $\gamma < \frac{m}{2}$, $\|R_{mn}\|_F$ does neither depend on α nor on ω. For any choice of Φ and Ψ, the Frobenius norm is the same as if the preconditioner of T. Chan and Olkin is used.*

For non-banded matrices (10) can be used to compute optimal parameters α and ω. The first important subclass of BTTB matrices which we want to examine are real matrices with symmetric blocks which are also symmetric on the block level. In this case we can deduce that $c_3 = c_4$. Then with $\omega = e^{i\Phi}$ and $\alpha = e^{i\Psi}$, (10) becomes

$$\|R_{mn}\|_F^2 = c_0 + 2c_1 \cos\Psi + 2c_2 \cos\Phi + 4c_3 \cos\Phi \cos\Psi. \tag{11}$$

The first partial derivatives of (11) are

$$\begin{aligned} \frac{\partial \|R_{mn}\|_F^2}{\partial \Phi} &= -\sin\Phi(2c_2 + 4c_3 \cos\Psi), \\ \frac{\partial \|R_{mn}\|_F^2}{\partial \Psi} &= -\sin\Psi(2c_1 + 4c_3 \cos\Phi). \end{aligned} \tag{12}$$

The following candidates for a minimum lead to real α and ω:

$$(\Phi, \Psi) = (0,0), (0,\pi), (\pi,0), (\pi,\pi) \quad .$$

Since in all four cases the Hessian matrix is diagonal, one can directly read off whether there is a minimum, a maximum, or none of those.

The advantages of our new preconditioner shall be demonstrated in the following example, in which the preconditioner is applied to BTTB matrices which are close to being circulant or skew-circulant on the block level and close to being circulant or skew-circulant within the blocks. The example is based on the fact that a BTTB matrix A_{mn} can be written as the sum of four matrices

$$A_{mn} = CC + SC + CS + SS \quad . \tag{13}$$

In this decomposition CC is a BCCB matrix, CS is circulant on the block level and has skew-circulant blocks, SC has circulant blocks, but is skew-circulant on the block level, and SS is skew-circulant on both levels.

Example 2. Let A_{mn} be the BTTB matrix defined by $a_0^{(0)} = 2$ and

$$a_l^{(k)} = \frac{1}{k+l+2} \quad \text{for} \quad (k,l) \neq (0,0) \ ,$$

which has the decomposition (13). In order to test the preconditioner we weight the terms of the sum (13) and define the matrices

$$T_{mn} = p_1 \cdot CC + p_2 \cdot SC + p_3 \cdot CS + p_4 \cdot SS \quad ,$$

where the parameters p_j satisfy $p_j \geq 0$ and

$$p_1 + p_2 + p_3 + p_4 = 4 \quad .$$

If p_1 is large compared to the other p_j, CC is the dominant component in T_{mn}, and $\|R_{mn}\|_F^2$ is minimal for $(\Phi, \Psi) = (0,0)$. For large p_2, SC is dominant and $\|R_{mn}\|_F^2$ has its minimum at $(\Phi, \Psi) = (\pi, 0)$. For large p_3 or p_4, the minimum is found at $(\Phi, \Psi) = (0, \pi)$ or $(\Phi, \Psi) = (\pi, \pi)$, respectively. This optimal choice of the parameters Φ and Ψ not only minimizes the Frobenius norm, but also improves the behavior of the pcg method. The table 2 shows the numerical results for $m = 80$ and $n = 120$.

	$(0,0)$	$(0,\pi)$	$(\pi,0)$	(π,π)
p_1=3.7 , p_2=p_3=p_4=0.1	4	12	13	16
p_1=2.5 , p_2=p_3=p_4=0.5	5	10	13	12
p_2=3.7 , p_1=p_3=p_4=0.1	11	19	5	10
p_2=2.5 , p_1=p_3=p_4=0.5	9	14	8	9
p_3=3.7 , p_1=p_2=p_4=0.1	10	5	20	12
p_3=2.5 , p_1=p_2=p_4=0.5	9	8	17	11
p_4=3.7 , p_1=p_2=p_3=0.1	15	13	12	5
p_4=2.5 , p_1=p_2=p_3=0.5	10	11	11	8

Table 2

Even if a real BTTB matrix T_{mn} is not symmetric on both levels, it can be shown that $(0,0), (0,\pi), (\pi,0), (\pi,\pi)$ are candidates for minima. Although the Hessian matrix is not diagonal for such matrices T_{mn}, it can be used to determine the minimum.

Finally, let us look at complex BTTB matrices which are Hermitian on both levels. In this case, (10) can be simplified further. We obtain that $c_3 = \overline{c_4}$ and that c_2 is real. With $c_1 = r_1 e^{i\theta_1}$, $c_2 = r_2$, and $c_3 = r_3 e^{i\theta_3}$ the following equations are the analogues of (11) and (12) in the Hermitian case.

$$\|R_{mn}\|_F^2 = c_0 + 2r_1 \cos(\theta_1 + \Psi) + 2r_2 \cos\Phi + 4r_3 \cos\Phi \cos(\theta_3 + \Psi),$$

$$\frac{\partial \|R_{mn}\|_F^2}{\partial \Phi} = -\sin\Phi(2r_2 + 4r_3 \cos(\theta_3 + \Psi)),$$

$$\frac{\partial \|R_{mn}\|_F^2}{\partial \Psi} = -2r_1 \sin(\theta_1 + \Psi) - 4r_3 \cos\Phi \sin(\theta_3 + \Psi).$$

Thus, possible candidates for a minimum need to have $\Phi = 0$ or $\Phi = \pi$ and, in addition, satisfy

$$-2r_1 \sin(\theta_1 + \Psi) \pm 4r_3 \sin(\theta_3 + \Psi) = 0 \quad ,$$

respectively. This leads to the following pairs of parameters:

$$(\Phi, \Psi) = (0, \arctan(\tfrac{4r_3 \sin(\theta_3 - \theta_1)}{-2r_1 - 4r_3 \cos(\theta_3 - \theta_1)}) - \theta_1), \quad (\pi, \arctan(\tfrac{-4r_3 \sin(\theta_3 - \theta_1)}{-2r_1 + 4r_3 \cos(\theta_3 - \theta_1)}) - \theta_1) \; .$$

4.2. Extending the preconditioner of Hanke and Nagy

The approximate inverse preconditioner of Hanke and Nagy is computed by embedding T_n into a circulant matrix $C_{n+\beta}$ and by exploiting the fast invertability of $C_{n+\beta}$. Again, we try to find a new preconditioner by using ω-circulant matrices, this time for the embedding of T_n into the ω-circulant matrix $C_{n+\beta}(\omega)$. In analogy to (3) we embed T_n into

$$C_{n+\beta}(\omega) = \begin{pmatrix} T_n & T_{2,1}^H \\ T_{2,1} & T_{2,2} \end{pmatrix}.$$

To make $C_{n+\beta}(\omega)$ an ω-circulant matrix, we define

$$T_{2,1} = \begin{pmatrix} \omega t_\beta & & & 0 & \cdots & 0 & \overline{t_\beta} & \cdots & \overline{t_1} \\ \vdots & \ddots & & \vdots & \ddots & \vdots & & \ddots & \vdots \\ \omega t_1 & \cdots & \omega t_\beta & 0 & \cdots & 0 & & & \overline{t_\beta} \end{pmatrix} \quad ,$$

where $\omega = e^{i\theta}$ with $\theta \in [\pi, \pi]$. As we have seen in (5), the diagonal matrix $\Lambda_{n+\beta}$ containing the eigenvalues of $C_{n+\beta}(\omega)$ is computed as follows:

$$\Lambda_{n+\beta} = F_{n+\beta} \Omega_{n+\beta}^H C_{n+\beta}(\omega) \Omega_{n+\beta} F_{n+\beta}^H = F_{n+\beta} C_{n+\beta}^{circ}(\omega) F_{n+\beta}^H \quad .$$

In this equation, $\Omega_{n+\beta}$ is the diagonal matrix

$$\Omega_{n+\beta} = diag(1, \omega^{1/(n+\beta)}, \ldots, \omega^{(n+\beta-1)/(n+\beta)})$$

and $F_{n+\beta}$ the Fourier matrix with entries $F_{k,j}^{(n+\beta)} = \frac{1}{\sqrt{n+\beta}} e^{\frac{2\pi ijk}{n+\beta}}$. Once the eigenvalues are obtained, the inverse of the ω-circulant matrix $C_{n+\beta}(\omega)$ must be computed. If all eigenvalues are positive, this is done via

$$(14) \quad C_{n+\beta}(\omega)^{-1} = \begin{pmatrix} M_n & M_{1,2} \\ M_{2,1} & M_{2,2} \end{pmatrix} = \Omega_{n+\beta} F_{n+\beta}^H \Lambda_{n+\beta}^{-1} F_{n+\beta} \Omega_{n+\beta}^H \quad .$$

However, if $\Lambda_{n+\beta}$ contains nonpositive eigenvalues, $\Lambda_{n+\beta}^{-1}$ is replaced by $\Lambda_{n+\beta}^{-}$ as it was done in (4). The result is

$$(15) \quad C_{n+\beta}(\omega)^{-} = \begin{pmatrix} M_n & M_{1,2} \\ M_{2,1} & M_{2,2} \end{pmatrix} = \Omega_{n+\beta} F_{n+\beta}^H \Lambda_{n+\beta}^{-} F_{n+\beta} \Omega_{n+\beta}^H \quad .$$

To show that $M_n T_n$ has the clustering property we extend the result of Hanke and Nagy to the ω-circulant case.

Theorem 5. *Let T_n be an Hermitian positive definite Toeplitz matrix with bandwidth $\beta < n/2$, which is embedded into the $(n+\beta)$-by-$(n+\beta)$ ω-circulant matrix $C_{n+\beta}(\omega)$ with $\omega = e^{i\theta}$ and $\theta \in [-\pi, \pi]$.*

1. *If $C_{n+\beta}(\omega)$ is positive definite, and M_n given as in (14), then M_n is positive definite, and $M_n T_n = I_n + R_n$, where $rank(R_n) \le \beta$.*
2. *If $C_{n+\beta}(\omega)$ has ν nonpositive eigenvalues, and M_n is defined as in (15), then M_n is positive definite, and $M_n T_n = I_n + R_n$, where $rank(R_n) \le \beta + \nu \le 2\beta$.*

This time we do not choose the parameter ω to minimize a norm. In order to find criteria for a suitable choice we consider the eigenvalues of $C_{n+\beta}(\omega)$. With the FFT and with some simplifications we compute the following expression for the elements of $\Lambda_{n+\beta}$:

$$\begin{pmatrix} \lambda_0 \\ \lambda_1 \\ \vdots \\ \lambda_{n+\beta-1} \end{pmatrix} = \begin{pmatrix} t_0 + 2\sum_{j=1}^{\beta} r_j \cos\left(\frac{j\theta}{n+\beta} + \varphi_j\right) \\ t_0 + 2\sum_{j=1}^{\beta} r_j \cos\left(\frac{-2\pi j}{n+\beta} + \frac{j\theta}{n+\beta} + \varphi_j\right) \\ \vdots \\ t_0 + 2\sum_{j=1}^{\beta} r_j \cos\left(\frac{-2\pi(n+\beta-1)j}{n+\beta} + \frac{j\theta}{n+\beta} + \varphi_j\right) \end{pmatrix}$$

with $t_j = r_j e^{i\varphi_j}$ and $\omega = e^{i\theta}$. Theorem 5 gives an estimate for the rank of the matrix R_n and therefore, implicitly, for the number of iterations the pcg method needs to converge. With each nonpositive eigenvalue, this estimate deteriorates. Thus, we try to choose θ such that as many eigenvalues as possible are positive.

Example 3. Again, we start with the matrices

$$T_n = tridiag(-1, 2, -1).$$

The ω-circulant extension matrix $C_{n+\beta}$ with first row

$$(2, -1, 0, \ldots, 0, -\omega^{-1})$$

has the eigenvalues

$$\lambda_j = 2 - 2\cos\left(\frac{\theta - 2j\pi}{n+1}\right) \quad (0 \leq j \leq n) \ ,$$

which are all nonnegative. For the original preconditioner of Hanke and Nagy, which is obtained for $\theta = 0$, C_{n+1} has the zero eigenvalue λ_0. For all other choices of θ all eigenvalues are positive, the minimum eigenvalue taking its maximum for $\theta = \pi$. The table 3 shows that the theoretical improvement corresponds to the numerical results.

n	$\theta = 0$	$\theta = \pi$
10000	6	2
15000	6	2
20000	9	2
25000	9	2

Table 3

We wish to extend this result to all weakly diagonally dominant matrices, i.e. to all matrices satisfying $t_0 \geq 2\sum_{j=1}^{n} |t_j| = 2\sum_{j=1}^{n} r_j$. If $t_0 > 2\sum_{j=1}^{n} r_j$, the corresponding generating function is strictly positive. In this case, the preconditioner of Hanke and Nagy converges very fast, and cannot be further improved by a different choice of ω. However, if $t_0 = 2\sum_{j=1}^{n} r_j$, the problem of zero eigenvalues arises. Let us especially consider the case where either all non-diagonal elements are positive or where they are all negative. In the following theorem we prove for these matrices that for the Hanke/Nagy preconditioner $\lambda_0 = 0$, and in addition, that $C_{n+\beta}(1)$ has k zero eigenvectors if only the k-th, $2k$-th, $3k$-th upper and lower diagonals of T_n are nonzero, and all other entries are zero.

Theorem 6. *Let T_n be a real symmetric Toeplitz matrix, $C_{n+\beta}(1)$ its circulant extension, and k a positive integer with $k|(n+\beta)$. Let $t_0 > 0$, $t_{p\cdot k} \leq 0$ for $p > 1$, and $t_r = 0$ for all other r. In addition to this, let T_n satisfy $t_0 = 2\sum_{j=1}^{n} r_j$. Then the following k eigenvalues of $C_{n+\beta}(1)$ are zero:*

$$\lambda_{\frac{s(n+\beta)}{k}}, \qquad s = 0, \ldots, k-1.$$

Proof. The matrix $C_{n+\beta}(1)$ has the eigenvalues

$$\text{(16)} \qquad \lambda_l = t_0 - 2\sum_{j=1}^{\beta} r_j \cos\left(\frac{-2\pi lj}{n+\beta}\right), \qquad l = 0, \ldots, n+\beta-1.$$

Since $t_j \neq 0$ only if $j = p \cdot k$, (16) becomes

$$\text{(17)} \qquad \lambda_l = t_0 - 2\sum_{p=1}^{\beta/k} r_{p\cdot k} \cos\left(\frac{-2\pi lpk}{n+\beta}\right), \qquad l = 0, \ldots, n+\beta-1.$$

From (17) we can conclude that for $s = 0, \ldots, k-1$ the eigenvalues $\lambda_{\frac{s(n+\beta)}{k}}$ are zero. The theorem is proved.

To conclude this section we consider the example Hanke and Nagy [7] gave to demonstrate the capabilities of their preconditioner.

Example 4. Let T_n be the real symmetric Toeplitz matrix with $t_0 = 1$, $t_1 = -0.25$, $t_6 = -0.25$, and $t_j = 0$ for all other j. The following table displays the number of iterations the pcg method needs to converge for $\theta = 0$ and for $\theta = \pi$.

n	$\theta = 0$	$\theta = \pi$
10000	10	7
15000	11	7
20000	11	7
25000	12	7

Table 4

4.3. Extending the preconditioner of Strang

In this final section we wish to develop an ω-circulant version $S_n(\omega)$ of Strang's preconditioner. Again, we are not interested in minimizing a norm, but rather in avoiding a singular preconditioning matrix. For banded matrices T_n we can carry over the results of the previous section, because Strang's preconditioner for T_n is equivalent with the circulant matrix C_n Hanke and Nagy define for the embedding of $T_{n-\beta}$, if the matrices T_n and $T_{n-\beta}$ have the same generating function. From this observation we can conclude that our extended preconditioner with a choice of ω different from 1 behaves exactly the same as the preconditioner of Strang as long as the generating function is strictly positive. If the generating function has zeros, however, convergence of the cg method depends crucially on a suitable choice of ω. In this case, the main goal is to make the preconditioning matrix $S_n(\omega)$ regular, i.e. to avoid zero eingenvalues. This can be done according to the same criteria as for the construction of the extended Hanke/Nagy preconditioner. For real, weakly diagonally dominant matrices which have positive entries only in the main diagonal this means avoiding the choice $\theta = 0$. To conclude this section we revisit Example 4. The preconditioner of Strang completely fails for the matrix $tridiag(-1, 2, -1)$, whereas for all other choices of θ the pcg method converges extremely fast. The numerical results are shown in the table 5.

n	$\theta = 0$	$\theta = \frac{\pi}{2}$	$\theta = \pi$	$\theta = -\frac{\pi}{2}$
10000	–	3	3	3
15000	–	3	3	3
20000	–	3	3	3

Table 5

Our improved version of Strang's preconditioner has the same convergence properties as the improved circulant preconditioner suggested by Tyrtyshnikov [13]. Whereas Tyrtyshnikov avoids singular circulant preconditioners by replacing the zero eigenvalues by a small positive number δ, we achieve the same result with a suitable choice of ω.

5. Conclusions

In this paper we have presented preconditioners for Toeplitz systems which are either ω-circulant or constructed with ω-circulant matrices. The extension of T. Chan's preconditioner, which minimizes the Frobenius norm over all ω-circulant matrices, works for all ω. It improves the convergence of the pcg method in many examples, especially in those containing Toeplitz matrices which are closely related to skew-circulant matrices. We have subsequently carried over these results to the two-dimensional case. Block-ω-circulant matrices with α-circulant blocks extend the preconditioner of T. Chan and Olkin for BTTB matrices. For matrices which are almost skew-circulant on both levels it is a significant improvement.

The extension of the approximate inverse preconditioner, on the other hand, is also a real improvement compared to the preconditioner of Hanke and Nagy. If it is possible to reduce the number of negative or zero eigenvalues of the ω-circulant extension matrix, the pcg method converges considerably faster. Similar results are obtained for the extension of Strang's preconditioner.

References

1. Chan, R. and Ng, M. (1996): Conjugate Gradient Methods for Toeplitz Systems, SIAM Review, **38**,427–482.
2. Chan, T. (1988): An Optimal Circulant Preconditioner for Toeplitz Systems, SIAM J. Sci. Stat. Comp., **9**, 766–771.
3. Davis, P. (1979): *Circulant Matrices*, John Wiley and Sons, New York.
4. Grenander, U. and Szegö, G. (1984): *Toeplitz Forms and Their Applications*, Chelsea Publishing, New York, 2-nd edition.
5. Strang, G. (1986):A Proposal for Toeplitz Matrix Calculations, Stud. Appl. Math., **74**, 171–176.
6. Huckle, T. (1994):Iterative Methods for Toeplitz-like Matrices, SCCM-94-05, Computer Science Dept., Stanford Univ.
7. Hanke, M. and Nagy, J. (1994): Toeplitz Approximate Inverse Preconditioner for Banded Toeplitz Matrices, Numerical Algorithms, **7**, 183–199.
8. Chan, R. (1989): Circulant Preconditioners for Hermitian Toeplitz Systems, SIAM J. Matrix Anal. Appl., **10**, 542–550.
9. Serra, S. (1998): On the Extreme Eigenvalues of Hermitian (Block) Toeplitz Matrices, Linear Algebra Appl., **270**, 109–129.
10. Golub, G. and Ortega, J.M. (1993): *Scientific Computing, An Introduction with Parallel Computing*, Academic Press.
11. Kailath, T. and Sayed, A. H. (1999): *Fast Reliable Algorithms for Matrices with Structure*, SIAM.
12. Chan, R. and Yeung, M. (1992): Jackson's Theorem and Circulant Preconditioned Toeplitz Systems, J. Approx. Theory, **70**, 191–205.
13. Tyrtyshnikov, E. E. (1995): Circulant Preconditioners with Unbounded Inverses, Linear Algebra Appl., **216**,1–24.
14. Chan, T. and Olkin J. (1994): Circulant Preconditioners for Toeplitz-block Matrices, Numerical Algorithms, **6**, 89–101.