A NEW PRECONDITIONER FOR THE PARALLEL SOLUTION OF POSITIVE DEFINITE TOEPLITZ SYSTEMS

Dario Bini, Dipartimento di Matematica, Università di Pisa;
Fabio Di Benedetto, Dipartimento di Matematica, Università di Genova.

Abstract. We introduce a new preconditioner for solving a symmetric Toeplitz system of equations by the conjugate gradient method. This choice leads to an algorithm which is particularly suitable for parallel computations and, compared to the circulant preconditioner of [C3], has a better asymptotic convergence rate and a lower arithmetic cost per iteration.

1. Introduction. Let $A_n = (a_{i,j})$, $a_{i,j} = a_{|i-j|}$ be an $n \times n$ real symmetric Toeplitz matrix, that is a matrix having constant entries down each diagonal. The solution of Toeplitz linear systems has many applications in scientific and engineering problems. Effective sequential algorithms for the solution of the system $A_n\mathbf{x} = \mathbf{b}$ with $O(n \log^2 n)$ arithmetic operations have been devised in [BGY],[DH]. Despite their arithmetic efficiency, all these algorithms are intrinsecally sequential and no implementation in the PRAM model requiring less than $\Omega(n)$ parallel steps is known. In the PRAM model of parallel computation we assume that at each step each processor can perform a single arithmetic operation.

In [PR],[P], iterative methods for the parallel solution of Toeplitz systems have been considered. Such methods require $O(\log^2 n)$ parallel steps and $O(n^2)$ processors, and have a quadratic convergence.

Recently the preconditioned conjugate gradient method with circulant preconditioning has been proposed by Strang and Chan [S], [C1]. Each iteration of this algorithm can be executed in $O(\log n)$ parallel steps with only $O(n)$ processors, since solving circulant systems, as well as computing Toeplitz matrix-vector product, can be performed by means of FFT. Under some additional hypothesis on the matrix A_n, the convergence is proved with a superlinear rate. This makes preconditioned conjugate gradient methods particularly suitable for the effective parallel solution of Toeplitz systems.

In this paper we propose a new class of preconditioners. Instead of the class of circulant matrices as in [S],[C1], i.e., the algebra generated by the unit circulant matrix

$$Z = \begin{pmatrix} 0 & \dots & 0 & 1 \\ 1 & \dots & & 0 \\ & \ddots & & \vdots \\ 0 & & 1 & 0 \end{pmatrix},$$

we consider the class τ defined in [BC] as the algebra generated by

$$W = \begin{pmatrix} 0 & 1 & & \\ 1 & 0 & \ddots & \\ & \ddots & \ddots & 1 \\ & & 1 & 0 \end{pmatrix}.$$

Since τ systems can be solved by means of a sine transform [BC], the arithmetic cost of each iteration is reduced by a constant factor. Moreover, under the same assumptions on the matrix A_n of [C3], we can prove a better convergence rate.

Suppose that the matrices A_n, $n \geq 1$ are finite sections of a singly infinite symmetric matrix A_∞, generated by the real-valued function $f(z) = \sum_{j=-\infty}^{+\infty} a_j z^j$ defined on the unit circle in the complex plane. Moreover, assume that f belongs to the Wiener class, that is $\sum_{j=-\infty}^{+\infty} |a_j| < +\infty$; if the function f is positive, then all the matrices A_n are positive definite [GS]. In [CS] the preconditioner is chosen as the circulant matrix C_n copying the central diagonals of the Toeplitz matrix A_n; for instance, if $n = 2m$ then the first column of C_n contains the entries $a_0, a_1, \dots, a_m, a_{m-1}, \dots, a_1$.

Our preconditioner is the τ matrix $T_n = A_n - H_n$, where $H_n = (h_{i,j})$ is a Hankel matrix whose antidiagonals are constant and equal to $a_2, \dots, a_{n-1}, 0, 0, 0, a_{n-1}, \dots, a_2$, i.e., $h_{i,j} = h_{i+j-1}$, $h_k = a_{n-|n-k|+1}$, $h_{n+1} = h_{n+2} = h_{n+3} = 0$. The circulant preconditioner C_n has the following properties (see [CS],[C3]):

1) for any $\epsilon > 0$, the spectrum of C_n lies on the interval $[f_{min} - \epsilon, f_{max} + \epsilon]$ for a sufficiently large n, where f_{min} and f_{max} are the (positive) extremal values of f on the unit circle; hence, each iteration requires the solution of a circulant system whose condition number is independent of n;
2) the spectrum of $C_n^{-1}A_n$ is clustered around 1, so that the conjugate gradient method converges to the solution;

3) each iteration requires about $\frac{9}{2}n\log n$ complex operations;
4) if the $(l+1)$-st derivative of f exists and is continuous on the unit circle, then the error on the solution is reduced after $2q$ iterations by a factor of $\frac{c^q}{(q-1)!^{2l}}$, where c depends only on f and l.

In the next section we will prove analogous properties for the τ preconditioner T_n:

1') for any $\epsilon > 0$, the spectrum of T_n lies on the interval $[f_{min} - \epsilon, f_{max} + \epsilon]$ for a sufficiently large n;
2') the spectrum of $T_n^{-1}A_n$ is clustered around 1;
3') each iteration requires about $\frac{15}{4}n\log n$ complex operations;
4') under the assumptions on f as in 4), the error is reduced after $2q$ iterations by an asymptotical factor of $\frac{c^q}{q!^{2l}}$, c being the same constant as in 4).

Comparing the theoretical bounds of 4) and 4'), we have that, after $2q$ iterations, the bound on the error obtained by the τ preconditioner is less than the bound of the circulant preconditioner by the factor of $\frac{1}{q^{2l}}$. For instance, for large values of n, the theoretical estimate of the error given for the new preconditioner after 8 iterations is about 16^l times less than the analogous estimate proved in [C3].

In the last section we will consider a different choice of the preconditioner, that is the τ matrix F_n which minimizes the Frobenius norm of the difference $F_n - A_n$. We will give the explicit expression of F_n and we will show that T_n and F_n yield the same asymptotical convergence rate.

2. Main results. We give an outline of the proofs of the properties displayed in the previous section.

Concerning 1'), we observe that T_n can be diagonalized as follows [BC]:

$$T_n = Q_n^T D_n Q_n, \; Q_n = \left(\sqrt{\frac{2}{n+1}}\sin\frac{\pi i j}{n+1}\right),$$
$$D_n = \mathrm{Diag}(\sigma_1, \ldots, \sigma_n).$$

From the relation $Q_n T_n = D_n Q_n$, we have

$$\sigma_i = \frac{\sum_{j=1}^{n} t_j \sin(j\alpha_i)}{\sin\alpha_i}$$

where $\alpha_i = \frac{\pi i}{n+1}$ and $(t_1, \ldots, t_n)^T$ is the first column of T_n. It is easy to see that

$$\sigma_i = a_0 + 2a_1 + 2\sum_{j=2}^{n-1} a_j \cos(j\alpha_i) = \mathrm{Re}\left(\sum_{\mathrm{j}=-\mathrm{n}+1}^{n-1} \mathrm{a_j e^{ij\alpha_i}}\right),$$

where $\alpha_i = \frac{\pi i}{n+1}$. Therefore, since the argument of the real part is a partial sum of the function f evaluated at the point $e^{ij\alpha_i}$ and f belongs to the Wiener class, we have $\sigma_i \in [f_{min} - \epsilon, f_{max} + \epsilon]$, for a sufficiently large n and $\epsilon > 0$.

Property 2') can be proved in the following way. Since $T_n^{-1}A_n = I_n + T_n^{-1}H_n$, it suffices to show that the eigenvalues of $T_n^{-1}H_n$ are clustered around 0. Let $\epsilon > 0$ be fixed; since f belongs to the Wiener class, we can choose N such that $\sum_{k=N+1}^{+\infty} |a_k| < \epsilon$.

The matrix H_n can be split as

$$\tilde{H}_n^{(N)} + E_n^{(N)}. \tag{1}$$

The first matrix agrees with H_n in the upper left and lower right $(N-1) \times (N-1)$ submatrices and vanishes in the other entries; its rank is not greater than $2(N-1)$. The 2-norm of the second matrix can be easily bounded by 2ϵ.

By Cauchy interlace theorem [W], the inequalities

$$\lambda_i(\tilde{H}_n^{(N)}) + \lambda_1(E_n^{(N)}) \le \lambda_i(H_n) \le \lambda_i(\tilde{H}_n^{(N)}) + \lambda_n(E_n^{(N)})$$

hold for $i = 1, \ldots, n$, where the eigenvalues are labelled in nondecreasing order. Since $\tilde{H}_n^{(N)}$ has rank not greater than $2(N-1)$, for at least $n - 2N + 2$ values of i there exists an eigenvalue of H_n lying on the interval $[\lambda_1(E_n^{(N)}), \lambda_n(E_n^{(N)})]$, which is included in $[-2\epsilon, 2\epsilon]$. Applying Courant-Fischer minimax characterization [W] to the symmetric matrix $T_n^{-1/2}H_nT_n^{-1/2}$, which is similar to $T_n^{-1}H_n$, we have for large n

$$|\lambda_i(T_n^{-1}H_n)| < \frac{|\lambda_i(H_n)|}{f_{min}/2};$$

hence, even the spectrum of $T_n^{-1}H_n$ is clustered around 0.

The clustering of the eigenvalues can be proved also by following the technique used in [CS], that is by relating the eigenvalue problem at the dimension n to a limiting singly infinite problem. For this purpose, the change of variable $\nu = 1 - \frac{1}{\lambda}$ brings the initial problem $A_n\mathbf{x} = \lambda T_n\mathbf{x}$ into the form $H_n\mathbf{x} = \nu A_n\mathbf{x}$. If n is even, by a suitable change of basis we can split the last problem of size n into two subproblems of size $\frac{n}{2}$:

$$(K + SJ)\mathbf{x} = \nu_+(U + RJ)\mathbf{x}$$

and

$$(K - SJ)\mathbf{x} = \nu_-(U - RJ)\mathbf{x}.$$

The matrices K, R, S, U derive from the partitionings $A_n = \begin{pmatrix} U & R \\ R^T & U \end{pmatrix}$ and $H_n = \begin{pmatrix} K & S \\ S^T & JKJ \end{pmatrix}$, while J is the "reflection matrix" $\begin{pmatrix} 0 & & 1 \\ & \iddots & \\ 1 & & 0 \end{pmatrix}$.

Proving analogous results as Lemma 1 and Theorem 3 of [CS], it can be shown that each of these subproblems tends to the singly infinite problem $K_\infty\mathbf{y}_\infty = \nu_\infty T_\infty \mathbf{y}_\infty$, where K_∞ is a Hankel matrix with entries $a_2, a_3, \ldots$ and T_∞ is a symmetric Toeplitz matrix with entries $a_0, a_1, \ldots$.

Since K_∞ is a compact operator, [CS] show that the limits ν_∞ are clustered around 0, and so are the eigenvalues ν for the size n. We point out that this argument implies that every limiting eigenvalue, which the eigenvalues of both subproblems tend to, must have at least multiplicity 2: we will use this information later.

Concerning 3) and 3'), at each iteration the computational cost is dominated by three real discrete Fourier transforms for C_n, by three real sine transforms for T_n and by a Toeplitz matrix-vector product for both; since the cost of sine transform is twice less than Fourier transform, a simple operation count gives the result mentioned in the first section.

In order to prove 4'), we note that the assumptions on f imply that $|a_j| \le \frac{\hat{c}}{|j|^{l+1}}$ for all j, where $\hat{c}$ is the L^1-norm on the unit circle of the $(l+1)$-st derivative of f [K]. For every N, we consider the splitting (1) ; by using the above bound for a_j as in [C3], for all $k \ge 1$ we obtain

$$\sum_{j=k+1}^{n-1} |a_j| \le \sum_{j=k+1}^{n-1} \frac{\hat{c}}{j^{l+1}} \le \int_k^{+\infty} \frac{\hat{c}}{x^{l+1}}\, dx \le \frac{\hat{c}}{k^l},$$

so that the 2-norm of $E_n^{(N)}$ is not greater than $\hat{c}(\frac{1}{N^l} + \frac{1}{(n-N+1)^l})$. Asymptotically, this bound approaches $\frac{\hat{c}}{N^l}$. The difference $E_n^{(N)} - E_n^{(N+1)}$ is a symmetric matrix of rank at most 4; we can express it as $\frac{1}{2}(w_N^+ w_N^{+T} + \tilde{w}_N^+ \tilde{w}_N^{+T} - w_N^- w_N^{-T} - \tilde{w}_N^- \tilde{w}_N^{-T})$, for a suitable choice of the vectors $w_N^\pm, \tilde{w}_N^\pm$. It is easy to prove by induction that

$$H_n = E_n^{(1)} = E_n^{(N)} + V_N^+ - V_N^-,$$

where the matrices

$$V_N^\pm = \frac{1}{2} \sum_{j=1}^{N-1} (w_j^\pm w_j^{\pm T} + \tilde{w}_j^\pm \tilde{w}_j^{\pm T})$$

are positive semidefinite and have rank not greater than $2N-2$.

Now we have to study the spectrum of the matrix $T_n^{-1} H_n$ which is similar to $T_n^{-\frac{1}{2}} H_n T_n^{-\frac{1}{2}}$; this matrix can be expressed as $\tilde{E}_n^{(N)} + \tilde{V}_N^+ - \tilde{V}_N^-$, where $\tilde{V}_n^\pm = T_n^{-\frac{1}{2}} V_n^\pm T_n^{-\frac{1}{2}}$ has the same properties of $V_N^\pm$ and the 2-norm of $\tilde{E}_n^{(N)} = T_n^{-\frac{1}{2}} E_n^{(N)} T_n^{-\frac{1}{2}}$ can be asymptotically bounded by the quantity $\frac{\tilde{c}}{N^l}$, $\tilde{c} = \frac{\hat{c}}{f_{min}}$, by using Courant-Fischer minimax characterization.

If the eigenvalues are labelled in nondecreasing order, Cauchy interlace theorem can be used to show that, for every i,

$$\lambda_i(T_n^{-1} H_n) \le \lambda_i(\tilde{V}_N^+) + \lambda_n(\tilde{E}_n^{(N)}) \le \lambda_i(\tilde{V}_N^+) + \frac{\tilde{c}}{N^l}$$

since $\lambda_n(-\tilde{V}_n^-)$ is nonpositive, and

$$\begin{aligned} \lambda_i(T_n^{-1} H_n) &\ge -\lambda_{n-i+1}(\tilde{V}_N^-) + \lambda_1(\tilde{E}_n^{(N)}) \\ &\ge -\lambda_{n-i+1}(\tilde{V}_N^-) - \frac{\tilde{c}}{N^l}, \end{aligned}$$

since $\lambda_1(\tilde{V}_n^+)$ is nonnegative. Since $\tilde{V}_N^\pm$ has low rank, for at most the first and last $2N-2$ values of i the corresponding eigenvalues of $T_n^{-1} H_n$ lie outside the interval $[-\frac{\tilde{c}}{N^l}, +\frac{\tilde{c}}{N^l}]$. If we label the eigenvalues of $T_n^{-1} H_n$ as $\mu_0^- \le \mu_1^- \le \ldots \le \mu_1^+ \le \mu_0^+$, we get for all N the inequality

$$|\mu_{2N}^\pm| < \frac{\tilde{c}}{(N+1)^l}. \tag{2}$$

We recall from [GV] that the error ϵ_q of the conjugate gradient method, after q iterations, is reduced by a factor which is not greater than the maximum value $|P_q(\lambda)|$, reached by an arbitrary polynomial P_q of degree q and constant term equal to 1, over the spectrum of $T_n^{-1} H_n$. We will make a suitable choice of P_q in order to estimate this factor under our assumptions.

For $k = 0, \ldots, q-1$ let $p_k(x)$ be the quadratic polynomial, of constant term 1, that vanishes at the eigenvalues $\lambda_{2k}^\pm$ (where $\lambda_k^\pm = 1 + \mu_k^\pm$); using (2), a simple count as in [C3] shows that the maximum value of $|p_k(\lambda)|$ on the interval $[\lambda_{2k}^-, \lambda_{2k}^+]$ is bounded by $\frac{c}{(k+1)^{2l}}$, where c is the same constant of [C3], depending on f and l.

As we have seen in the proof of 2'), the eigenvalues of $T_n^{-1} H_n$ are double at the limit, so, asymptotically, we have that p_{2k} vanishes even at $\lambda_{2k+1}^\pm$; hence, the product $P_{2q} = p_0 p_1 \cdots p_{q-1}$, of degree $2q$, vanishes at the first and the last $2q$ eigenvalues. Its maximum value on the whole spectrum is bounded by the quantity

$$c \cdot \frac{c}{2^{2l}} \cdot \cdots \cdot \frac{c}{q^{2l}} = \frac{c^q}{(q!)^{2l}};$$

this proves the asymptotical superlinear rate of convergence shown at the point 4').

3. Another τ preconditioner.

There exist other possible choices of the preconditioner in the τ class, whose numerical behaviour may be the same as T_n, or perhaps better. For example, in analogy with [C1], we discuss the τ matrix F_n such that

$$||F_n - A_n||_F = \min_{B_n \in \tau} ||B_n - A_n||_F,$$

where $||\cdot||_F$ is the Frobenius matrix norm.

Taking as unknowns the entries $\phi_1, \ldots, \phi_n$ of the first column of F_n, writing down the gradient of $||F_n - A_n||_F^2$ and solving the related system gives us the following solution:

$$\phi_1 = a_0 - \frac{n-2}{n+1} a_2, \; \phi_2 = a_1 - \frac{n-3}{n+1} a_3;$$

$$\phi_i = \frac{n-i+3}{n+1}a_{i-1} - \frac{n-i-1}{n+1}a_{i+1},\ i = 3,\dots,n-2;$$

$$\phi_{n-1} = \frac{4}{n+1}a_{n-2},\ \phi_n = \frac{3}{n+1}a_{n-1}.$$

The study of the spectrum of $F_n^{-1}A_n$ is more difficult than that of $T_n^{-1}A_n$; hence, to compare the two rates of convergence we will follow the same argument of [C2], by showing that

$$\lim_{n\to\infty} \rho(T_n - F_n) = 0. \tag{3}$$

In fact, such relation implies the following corollaries:

i) Since the spectra of T_n and F_n are asymptotically equal, the property 1') of section 1 holds for F_n too; in particular, even F_n is positive definite.

ii) Since the spectrum of $F_n^{-1}A_n$ approaches that of $T_n^{-1}A_n$, for large values of n the rate of convergence of the conjugate gradient method is the same for both the preconditioners.

In order to prove (3), note that the matrix $\Delta_n = F_n - T_n$ is symmetric and it still belongs to the τ class. By recalling the proof of 1'), we can express the i-th eigenvalue σ_i of Δ_n as $\frac{\sum_{j=1}^n d_j \sin(j\alpha_i)}{\sin\alpha_i}$, where $\alpha_i = \frac{\pi i}{n+1}$ and $(d_1,\dots,d_n)^T$ is the first column of Δ_n. A simple count shows that

$$|\sigma_i| \le 2\sum_{j=2}^{n-1} \frac{j}{n+1}|a_j| + \frac{2|\text{cotg}\alpha_i|}{n+1}\sum_{j=2}^{n-1}|a_j|\,|\sin(j\alpha_i)|. \tag{4}$$

Since f belongs to the Wiener class, for all ϵ there exists M such that $\sum_{j=M+1}^{n-1}|a_j| < \frac{\epsilon}{2}$. To show that the first sum of (4) tends to 0 as n grows, it suffices to write it as

$$\sum_{j=2}^{M}\frac{j}{n+1}|a_j| + \sum_{j=M+1}^{n-1}\frac{j}{n+1}|a_j| \le \frac{M}{n+1}\sum_{j=2}^{M}|a_j| + \sum_{j=M+1}^{n-1}|a_j|;$$

this is less than ϵ if $n > \frac{2M}{\epsilon}\sum_{j=2}^{M}|a_j|$.

In order to bound the second term of (4), we recall that $|\text{cotg}\alpha| < \frac{1}{\alpha}$ if $0 < \alpha \le \frac{\pi}{2}$ and $|\text{cotg}\alpha| < \frac{1}{\pi-\alpha}$ if $\frac{\pi}{2} < \alpha \le \pi$.

Hence, if $1 \le i \le \frac{n+1}{2}$ then $0 < \alpha_i \le \frac{\pi}{2}$ and the second term of (4) is less than

$$\frac{2}{(n+1)\alpha_i}\sum_{j=2}^{n-1}|a_j|\, j\alpha_i = 2\sum_{j=2}^{n-1}\frac{j}{n+1}|a_j| < \epsilon$$

for large n, as we have seen above.

If $\frac{n+1}{2} < i \le n$, then $\frac{\pi}{2} < \alpha_i < \pi$, so that the second term of (4) is less than

$$\frac{2}{(n+1)\beta_i}\sum_{j=2}^{n-1}|a_j|\,|\sin(j\pi - j\beta_i)| = \frac{2}{(n+1)\beta_i}\sum_{j=2}^{n-1}|a_j|\,|\sin(j\beta_i)|$$

with $\beta_i = \pi - \alpha_i$; now the proof can be carried out as in the previous case.

We have also proved that any eigenvalue of Δ_n tends to 0 as n increases, and this completes the proof of (3).

References.

[BA] R.Bitmead, B.Anderson, "Asymptotically fast solution of Toeplitz and related systems of equations", *Linear Algebra Appl.*, **34** (1980), 103-116.

[BC] D.Bini, M.Capovani, "Spectral and computational properties of band symmetric Toeplitz matrices", *Linear Algebra Appl.*, **52** (1983), 99-126.

[BGY] R.Brent, F.Gustavson, D.Yun, "Fast solution of Toeplitz systems of equations and computations of Padé approximants", *J. Algorithms*, **1** (1980), 259-295.

[C2] R.Chan, "The spectrum of a family of circulant preconditioned Toeplitz systems", *SIAM J. Numer. Anal.*, **26** (1989), 503-506.

[C3] R.Chan, "Circulant preconditioners for Hermitian Toeplitz systems", *SIAM J. Matrix Anal. Appl.*, **4** (1989), 542-550.

[CS] R.Chan, G.Strang, "Toeplitz equations by conjugate gradients with circulant preconditioner", *SIAM J. Sci. Stat. Comp.*, **10** (1989), 104-119.

[C1] T.Chan, "An optimal circulant preconditioner for Toeplitz systems", *SIAM J. Sci. Stat. Comp.*, **9** (1988), 766-771.

[DH] F.De Hoog, "On the solution of Toeplitz systems", *Linear Algebra Appl.*, **88/89** (1987), 123-138.

[GV] G.H.Golub, C.Van Loan, *Matrix Computations*, The Johns Hopkins University Press, Baltimore, MD, 1983.

[GS] U.Grenander, G.Szego, *Toeplitz Forms and Their Applications*, Second edition, Chelsea, New York, 1984.

[K] Y.Katznelson, *An Introduction to Harmonic Analysis*, Second edition, Dover, New York, 1976.

[P] V.Pan, "Fast and efficient parallel inversion of Toeplitz and block Toeplitz matrices", TR 88-8, Dept. of Computer Sci. SUNY Albany, N.Y., 1988.

[PR] V.Pan, J.Reif, "Displacement structure and the processor efficiency of fast parallel Toeplitz computations", *Proc. of 28-th annual IEEE Symposium FOCS* (1987), 173-184.

[S] G.Strang, "A proposal for Toeplitz matrix calculations", *Stud. Appl. Math.*, **74** (1986), 171-176.

[T] W.Trench, "An algorithm for the inversion of finite Toeplitz matrices", *SIAM J. Appl. Math.*, **12** (1964), 515-522.

[W] J.Wilkinson, *The Algebraic Eigenvalue Problem*, Clarendon Press, Oxford, 1965.