Stability analysis of the matrix-vector product (via FFT) for a Toeplitz-like matrix

PAOLA FAVATI[1], ORNELLA MENCHI[2]
[1] IIT - CNR Via G. Moruzzi 1, 56124 Pisa, ITALY
[2] Dip. di Informatica, University of Pisa, Largo Pontecorvo 3, 56127 Pisa, ITALY

Abstract: -In this paper the numerical stability of the matrix-vector product, performed via FFT, is analyzed for the case of the Toeplitz-like matrices. The error appears to depend on the magnitude of the generators of the matrix. The numerical experimentation confirms the theoretical result.

Key-Words: -Stability analysis, Toeplitz-like matrix, Fast Fourier Transform

Received: May 4, 2021. Revised: January 11, 2022. Accepted: January 28, 2022. Published: February 25, 2022.

1 Introduction

Let us consider a Fredholm integral equation of the first kind

$$g(s) = \int K(s,t)x(t)\,dt, \tag{1}$$

where the square integrable kernel $K(s,t)$ and the right-hand side $g(s)$ are given functions and $x(t)$ is the unknown solution to be reconstructed. In many applications $g(s)$ consists of measured quantities. Problems modeled by equation (1) are frequent both in the one-dimensional context (for example in signal processing and in the computation of inverse transformations) and in the two-dimensional context (for example in the image deconvolution problem where $K(s,t)$ represents an imaging system, $x(t)$ and $g(s)$ represent a real object and its image, respectively).

By discretizing (1), a linear system

$$A\boldsymbol{x} = \boldsymbol{g} \tag{2}$$

is obtained, whose main features are the large dimension of A and the distribution of its singular values which often decay gradually to zero. Hence solving (2) is an ill-posed problem. When some structure can be considered for A, system (2) results to be solvable in practice also for large dimensions. In many applications $K(s,t)$ can be assumed to be invariant with respect to translations and with a bounded support, so that matrix A results to have *Toeplitz* structure and a limited bandwidth. Toeplitz systems arise frequently in linear algebra problems. Unfortunately, when a simple operation like multiplication or inversion or low rank modification is applied to a Toeplitz matrix, the Toeplitz structure is lost. For example, an important quantity as the Schur complement of a leading principal submatrix of a Toeplitz matrix has no longer a Toeplitz structure. For this reason we consider here a more general structure, the class of *Toeplitz-like* matrices, which is closed for the most common operations applied in the numerical algorithms. The Toeplitz-like structure is based on the concept of displacement rank [1, 2, 3, 4] and has been studied by many authors with applications in several fields (see for example [5] for an extensive bibliography). In the last decade several papers dealt with fast and superfast solver (see for example [6, 7, 8], application to the solution of fractional partial differential equations [9, 10, 11] and to the queueing problem [12]. Moreover a great interest was addressed to the study of Toeplitz-like operators in infinite dimensional spaces ([13, 14, 15])

The special low cost algorithms based on the *fast Fourier transform* (FFT), which have been devised to perform the matrix-vector products with Toeplitz matrices, can be applied also to Toeplitz-like matrices.

FFT was first discussed by Cooley and Tukey in 1965 [16], although Gauss had already described the critical factorization step as early as 1805. It is one of the most important numerical algorithms and has a wide range of applications. Its ubiquitous fortune is principally due to a low computational cost: computing the discrete Fourier transform of a sequence of length n according to the definition, takes $O(n^2)$ arithmetical operations, while using FFT it takes only $O(n \log n)$ operations.

When finite-precision floating-point arithmetic is used, FFT algorithms give results affected by error, but this error is typically quite small, in fact most FFT algorithms enjoy excellent numerical stability (see [17, 18]). We are interested in investigating the stability of the matrix-vector product based on FFT for the case of Toeplitz-like matrices. The paper is so organized: in Section 2 two simple programs describe the use of FFT in the computation of the matrix-vector product for triangular Toeplitz matrices, in Section 3 a brief description of Toeplitz-like matrices is given and the function which computes the matrix-vector product for Toeplitz-like matrices is sketched. The analysis of the stability occupies Section 4 giving an upper bound of the error which depends on the magnitude of the generators of the Toeplitz-like matrix, and finally in Section 5 the results of the numerical experiments are shown, confirming the theoretical findings.

Notation: uppercase letters are used for matrix names, bold lowercase letters are used for vector names, upper index T indicates the transpose of a matrix or of a vector, upper index $*$ indicates the conjugate transpose of a matrix. The special vectors $\mathbf{0}$ and $\boldsymbol{e}_i$ indicate the null vector and the i-th canonical vector of suitable length. Bold letter $\boldsymbol{i}$ denotes the imaginary unit. The magnitude of a vector $\boldsymbol{r}$ is measured by a norm. Specifically $\|\boldsymbol{r}\|_1 = \sum_i |r_i|$, $\|\boldsymbol{r}\|_2 = \sum_i |r_i|^2$ and $\|\boldsymbol{r}\|_\infty = \max_i |r_i|$. Finally, the symbol $\odot$ denotes the componentwise multiplication between two vectors of equal length.

2 Circulant and Toeplitz matrices

Circulant matrices and Toeplitz matrices (see [19], [20], [21]) arise frequently in the numerical treatment of problems of scientific areas such as physics, signal and image processing, probability, statistics and many others. Let us outline some of their properties.

A circulant matrix M of order k is a square $k \times k$ matrix in which the first row $[m_{1,1}, m_{1,2}, \ldots, m_{1,k}]^T$ is given and each subsequent row is rotated one element to the right relative to the preceding row. Formally, the (i,j)-th element of the i-th row with $i = 2, \ldots, k$ is $m_{i,j} = m_{i-1,j-1}$ for $j = 2, \ldots, k$ and $m_{i,1} = m_{i-1,k}$.

The most important feature of circulant matrices is that they are diagonalized by a discrete Fourier transform, as shown in the following lemma.

Lemma 1 *A circulant matrix M of size k is diagonalized by the Fourier matrix $\mathcal{F}_k$, whose elements are*

$$f_{i,j} = \frac{1}{\sqrt{k}}\, \omega^{(i-1)(j-1)}, \quad i,j = 1, \ldots, k,$$
$$\textit{with} \quad \omega = \exp(2\pi \boldsymbol{i}/k).$$

Denote by $\boldsymbol{m}$ the transpose of the first row of M, and by $\operatorname{diag}(\mathcal{F}_k\, \boldsymbol{m})$ the diagonal matrix whose i-th principal element is the i-th element of the vector $\mathcal{F}_k\, \boldsymbol{m}$. Then

$$M = \sqrt{k}\, \mathcal{F}_k \operatorname{diag}(\mathcal{F}_k\, \boldsymbol{m})\, \mathcal{F}_k^*,$$

and for any vector $\boldsymbol{v}$ it holds

$$M\boldsymbol{v} = \sqrt{k}\, \mathcal{F}_k \left(\mathcal{F}_k \boldsymbol{m} \odot \mathcal{F}_k^* \boldsymbol{v}\right). \tag{3}$$

For a proof of this Lemma, see [19]. □

The products by $\mathcal{F}_k$ and $\mathcal{F}_k^*$ can be efficiently computed by calling FFT with computational cost of order $O(k \log k)$ for $k \to \infty$.

A Toeplitz matrix T is a $k \times k$ matrix in which each descending diagonal from left to right is constant. Two k vectors $\boldsymbol{r}$ and $\boldsymbol{s}$, with $r_1 = s_1$, are given; $\boldsymbol{r}^T$ is assumed as first row of T and $\boldsymbol{s}$ is assumed as first column of T. The (i,j)-th element of T is $t_{i,j} = t_{1+j-i}$ for $j \geq i$ and $t_{i,j} = t_{1+i-j}$ for $i < j$. Circulant matrices are special cases of Toeplitz matrices with $s = [r_1, r_k, \ldots, r_2]^T$.

Since the case of triangular Toepliz matrices is of particular interest, we use the following notation. For a given vector $\boldsymbol{s}$, $L(\boldsymbol{s})$ denotes the lower triangular Toeplitz matrix whose first column is $\boldsymbol{s}$ and $U(\boldsymbol{r})$ denotes the upper triangular Toeplitz matrix whose first row is $\boldsymbol{r}^T$, as shown in Figure 1. The matrix-vector product of a Toeplitz triangular matrix can be computed by exploiting the properties of circulant matrices. To see how relation (3) is exploited, consider first the case of a lower triangular Toeplitz matrix of size n. Let $L = L(\boldsymbol{s})$, with $\boldsymbol{s} = [s_1, s_2, \ldots, s_n]^T$. Given a vector $\boldsymbol{v}$, let $\boldsymbol{y} = L\boldsymbol{v}$ be the vector to be computed. The vector $\boldsymbol{v}$ is embedded in a $2n$-vector $\widehat{\boldsymbol{v}}$ and the matrix L is embedded in a $2n \times 2n$ circulant matrix M, whose first row is the $2n$-vector $[s_1, \mathbf{0}^T, s_n, \ldots, s_2]$, with

$$L(\boldsymbol{s}) = \begin{bmatrix} s_1 & & & \\ s_2 & s_1 & & \\ \vdots & \vdots & \ddots & \\ s_k & s_{k-1} & \cdots & s_1 \end{bmatrix}, \quad U(\boldsymbol{r}) = \begin{bmatrix} r_1 & r_2 & \cdots & r_k \\ & r_1 & \cdots & r_{k-1} \\ & & \ddots & \vdots \\ & & & r_1 \end{bmatrix}$$

Figure 1: Lower and upper triangular Toeplitz matrices of size k

$$\widehat{\boldsymbol{v}} = \begin{bmatrix} \boldsymbol{v} \\ \mathbf{0} \end{bmatrix}, \quad M = \begin{bmatrix} L & \widehat{L} \\ \widehat{L} & L \end{bmatrix},$$

where $\widehat{L}$ turns out to be an upper triangular Toeplitz matrix. Since

$$\boldsymbol{w} = M\, \widehat{\boldsymbol{v}} = \begin{bmatrix} L\, \boldsymbol{v} \\ \widehat{L}\, \boldsymbol{v} \end{bmatrix},$$

the vector $\boldsymbol{y}$ is found in the first half of the vector $\boldsymbol{w}$. Then, using (3) the product $\boldsymbol{y}$ can be computed by the function `lowert` given in `Algorithm 1`.

`Algorithm 1`: product of a lower triangular Toeplitz matrix $L(\boldsymbol{s})$ by a vector $\boldsymbol{v}$

```
function lowert $(n, \boldsymbol{s}, \boldsymbol{v})$
    $\boldsymbol{m} = [s_1, \mathbf{0}^T, s_n, \ldots, s_2]^T$;   $\widehat{\boldsymbol{v}} = \begin{bmatrix} \boldsymbol{v} \\ \mathbf{0} \end{bmatrix}$;
    $\boldsymbol{z} = \mathcal{F}_{2n}\, \boldsymbol{m}$;   $\boldsymbol{p} = \mathcal{F}_{2n}^* \widehat{\boldsymbol{v}}$;
    $\boldsymbol{q} = \boldsymbol{z} \odot \boldsymbol{p}$;   $\boldsymbol{w} = \sqrt{2n}\, \mathcal{F}_{2n} \boldsymbol{q}$;
    return $\boldsymbol{w}(1:n)$;
```

A similar procedure applies to the upper triangular Toeplitz matrix $U = U(\boldsymbol{r})$ whose first row is $\boldsymbol{r}^T = [r_1, r_2, \ldots, r_n]$. Given a vector $\boldsymbol{v}$, let $\boldsymbol{y} = U\boldsymbol{v}$ be the vector to be computed. Proceeding with the embedding as before, we see that the circulant matrix M has first row $[r_1, \ldots, r_n, \mathbf{0}^T]$. So $\boldsymbol{y}$ can be computed by the function `uppert` given in `Algorithm 2`.

`Algorithm 2`: product of an upper triangular Toeplitz matrix $U(\boldsymbol{r})$ by a vector $\boldsymbol{v}$

```
function uppert $(n, \boldsymbol{r}, \boldsymbol{v})$
$\boldsymbol{m} = \begin{bmatrix} \boldsymbol{r} \\ \mathbf{0} \end{bmatrix}$;   $\widehat{\boldsymbol{v}} = \begin{bmatrix} \boldsymbol{v} \\ \mathbf{0} \end{bmatrix}$;
$\boldsymbol{z} = \mathcal{F}_{2n}\, \boldsymbol{m}$;   $\boldsymbol{p} = \mathcal{F}_{2n}^* \widehat{\boldsymbol{v}}$;
$\boldsymbol{q} = \boldsymbol{z} \odot \boldsymbol{p}$;   $\boldsymbol{w} = \sqrt{2n}\, \mathcal{F}_{2n} \boldsymbol{q}$;
return $\boldsymbol{w}(1:n)$;
```

Since the Toeplitz structure is not maintained when simple operations like multiplication or inversion are applied, some generalizations have been proposed to deal with this aspect. Among them, we consider here a Toeplitz-like structure.

3 Toeplitz-like matrices

The definition of Toeplitz-like structure is based on the concept of displacement rank [1, 2, 3, 4], which depends on a particularly chosen displacement operator and measures how close a matrix is to a Toeplitz matrix. Given an $n \times n$ matrix A, we consider here the *displacement operator*

$$\nabla(A) = A - ZAZ^T, \tag{4}$$

where Z is the $n \times n$ *down-shift* matrix, i.e. the binary matrix with ones only on the subdiagonal and zeros elsewhere.

The matrix A is said to be *Toeplitz-like* if the quantity $r_{\text{disp}}(A) = \text{rank}\,\nabla(A)$ (called *displacement rank*) is small with respect to n (more formally $r_{\text{disp}}(A) = O(1)$ for $n \to \infty$). Let $\rho = r_{\text{disp}}(A)$, then

$$\nabla(A) = C\, D^T, \tag{5}$$

for suitable $n \times \rho$ matrices C and D, called *generators* of A. Denoting by $\boldsymbol{c}_i$ and $\boldsymbol{d}_i$, $i = 1, \ldots, \rho$, the columns of C and D respectively, then

$$\nabla(A) = \sum_{i=1}^{\rho} \boldsymbol{c}_i\, \boldsymbol{d}_i^T. \tag{6}$$

In this sense, we can say that A is represented through the generators $\boldsymbol{c}_i$, $\boldsymbol{d}_i$. In particular, a Toeplitz matrix A of elements $a_{i,j}$ with $a_{1,1} \neq 0$, is so represented

$$\nabla(A) = \boldsymbol{c}_1\, \boldsymbol{e}_1^T + \boldsymbol{e}_1\, \boldsymbol{d}_2^T,$$

$$\text{with} \quad \boldsymbol{c}_1 = A\, \boldsymbol{e}_1, \quad \boldsymbol{d}_2 = A^T \boldsymbol{e}_1 - a_{11} \boldsymbol{e}_1,$$

i. e. $\boldsymbol{c}_1$ is the first column of A and $\boldsymbol{d}_2^T$ is the first row of A with the first component set to zero. This shows that the displacement rank of a Toeplitz matrix is $\rho = 2$, except in the case of a triangular matrix where $\rho = 1$.

The set of Toeplitz-like matrices, unlike the set of Toeplitz matrices, is closed for the most common operations applied in the numerical algorithms. This does not mean that the displacement rank is maintained during the computation. For example, the matrix obtained by multiplying two matrices having $r_{\text{disp}} = \rho$ has the same r_{disp}, while the displacement rank of the inverse of a matrix which has $r_{\text{disp}} = \rho$ may rise up to $\rho + 2$. The leading principal submatrix of a matrix which has $r_{\text{disp}} = \rho$ and its Schur complement have still $r_{\text{disp}} = \rho$ (see [2, 7]).

The generators enable us to express a Toeplitz-like matrix as the sum of products of lower and upper triangular Toeplitz factors. In fact, from (4) we have

$$A = \nabla(A) + ZAZ^T = \nabla(A) + Z\nabla(A)Z^T$$
$$+Z^2 A (Z^2)^T = \ldots = \sum_{s=0}^{n-1} Z^s \nabla(A) (Z^s)^T.$$

Since $\displaystyle\sum_{s=0}^{n-1} \left(Z^s \boldsymbol{c}_i\right) \left(Z^s \boldsymbol{d}_i\right)^T = L(\boldsymbol{c}_i)\, U(\boldsymbol{d}_i),$ from (6) it follows that

$$A = \sum_{i=1}^{\rho} L(\boldsymbol{c}_i)\, U(\boldsymbol{d}_i),$$

and, given an n-vector $\boldsymbol{v}$, we have

$$A\boldsymbol{v} = \sum_{i=1}^{\rho} L(\boldsymbol{c}_i)\, U(\boldsymbol{d}_i)\, \boldsymbol{v}. \tag{7}$$

Using (7) we can compute $A\boldsymbol{v}$ by calling alternatively the matrix-vector products of upper and lower triangular Toeplitz matrices, as outlined in Algorithm 3. A saving of the cost can be achieved by skipping the last FFT call of `lowert` and exploiting the linearity of $\mathcal{F}_{2n}$ in the final sum.

If the matrix A is known to be Toeplitz-like, but it is only given explicitly, any factorization of $\nabla(A)$ can be employed to detect ρ and to construct the generators. For example, we can compute the Gaussian factorization $\nabla(A) = \mathcal{L}\mathcal{U}$, where $\mathcal{L}$ and $\mathcal{U}$ are lower and upper triangular matrices, employing a diagonal pivoting strategy and stopping at the first null pivot.

It follows that while $\rho = \text{rank}\,\nabla(A)$ is uniquely determined, the decomposition (5) of $\nabla(A)$, and consequently the representation of A through the generators, is not unique. An important representation is the

Algorithm 3: product of a Toeplitz-like matrix by a vector

```
function prod $(n, \rho, C, D, \boldsymbol{v})$;
for $i = 1$ to $\rho$;
    $\boldsymbol{h}_i =$ uppert $(n, \boldsymbol{d}_i, \boldsymbol{v})$;
    $\boldsymbol{g}_i =$ lowert $(n, \boldsymbol{c}_i, \boldsymbol{h}_i)$;
end for;
return $\sum_{i=1}^{\rho} \boldsymbol{g}_i$;
```

orthogonal one [22], obtained by computing the SVD decomposition $\nabla(A) = U\Sigma V^T$, where ρ is the number of strictly positive singular values σ_i in Σ and the matrices U and V have orthogonal columns.

Denoting by $\widehat{\Sigma}$ the $n \times \rho$ matrix having the i-th principal element equal to $\sqrt{\sigma_i}$ for $i = 1, \ldots, \rho$, and zero elsewere, the $n \times \rho$ matrices

$$C_{\text{ort}} = U\widehat{\Sigma}, \quad D_{\text{ort}} = V\widehat{\Sigma}, \tag{8}$$

have orthogonal columns and can be assumed as the generators C and D in (5).

The following relations hold between the magnitudes of A and $\nabla(A)$

$$\|\nabla(A)\|_2 \le 2\,\|A\|_2, \text{ and } \|A\|_2 \le n\,\|\nabla(A)\|_2. \tag{9}$$

To measure the magnitude of A when it is represented through the generators, we consider the function

$$\psi(C, D) = \sum_{i=1}^{\rho} \|\boldsymbol{c}_i\, \boldsymbol{d}_i^T\|_2 = \sum_{i=1}^{\rho} \|\boldsymbol{c}_i\|_2\, \|\boldsymbol{d}_i\|_2, \tag{10}$$

which obviously verifies $\|\nabla(A)\|_2 \le \psi(C, D)$.

It is known that the stability of a method solving a linear system depends on the growth of the matrix factors computed by the method. If the computations are performed on Toeplitz-like matrices represented through the generators, we expect the stability to depend on how large the generators become [23]. In the general case, no upper bound of $\psi(C, D)$ in terms of $\|A\|_2$ can be given. However, if the decomposition (5) is orthogonal, then from (8)

$$\|\nabla(A)\|_2 = \sigma_1,\ \boldsymbol{c}_i = \sqrt{\sigma_i}\,\boldsymbol{u}_i,\ \boldsymbol{d}_i = \sqrt{\sigma_i}\,\boldsymbol{v}_i,$$

where $\boldsymbol{u}_i$ is the i-th column of U and $\boldsymbol{v}_i$ is the i-th column of V. Then

$$\|\boldsymbol{c}_i\boldsymbol{d}_i^T\|_2 = \sigma_i\|\boldsymbol{u}_i\|_2\,\|\boldsymbol{v}_i\|_2 = \sigma_i,$$

hence from (9)

$$\psi(C, D) = \sum_{i=1}^{\rho} \sigma_i \le \rho\|\nabla(A)\|_2 \le 2\rho\|A\|_2. \tag{11}$$

4 Stability of the function prod

For the stability analysis we assume that the computations are carried out in a floating point arithmetic with unit roundoff ϵ. The computed value of a variable (scalar, vector or matrix) v will be denoted by $\widetilde{v}$ or by "$fl(v)$". We assume also that the quantities which appear in the bounds are not so large to invalidate a first order error analysis. For simplicity the term "$+O(\epsilon^2)$", which appears in the thesis of the theorems, is omitted in the proofs. Consequently, any expression of the form $x\,\widetilde{y}$, where $x = O(\epsilon)$ and $\widetilde{y} - y = O(\epsilon)$, is replaced by $x\,y$.

The following bounds are used [18]:

- For any vector $\boldsymbol{s}$ it is $\|L(\boldsymbol{s})\|_1 = \|L(\boldsymbol{s})\|_\infty = \|\boldsymbol{s}\|_1$, then

$$\begin{aligned}\|L(\boldsymbol{s})\|_2 &\le \sqrt{\|L(\boldsymbol{s})\|_1\|L(\boldsymbol{s})\|_\infty}\\ &= \|\boldsymbol{s}\|_1 \le \sqrt{n}\|\boldsymbol{s}\|_2,\end{aligned} \tag{12}$$

and analogously $\|U(\boldsymbol{r})\|_2 \le \sqrt{n}\,\|\boldsymbol{r}\|_2$ for any vector $\boldsymbol{r}$.

- Given a vector $\boldsymbol{x}$, a vector $\boldsymbol{\epsilon}$ whose components are bounded in modulus by ϵ exists such that

$$\boldsymbol{x} = \widetilde{\boldsymbol{x}} - \widetilde{\boldsymbol{x}} \odot \boldsymbol{\epsilon} + O(\epsilon^2). \tag{13}$$

- Given two vectors $\boldsymbol{x}$ and $\boldsymbol{y}$, with $\widetilde{\boldsymbol{x}} = \boldsymbol{x} + \boldsymbol{\delta}_x$ and $\widetilde{\boldsymbol{y}} = \boldsymbol{y} + \boldsymbol{\delta}_y$, then

$$fl\left(\widetilde{\boldsymbol{x}} \odot \widetilde{\boldsymbol{y}}\right) = \boldsymbol{x} \odot \boldsymbol{y} + \boldsymbol{\theta} + O(\epsilon^2), \tag{14}$$

where

$$\boldsymbol{\theta} = \boldsymbol{x} \odot \boldsymbol{\delta}_y + \boldsymbol{\delta}_x \odot \boldsymbol{y} + \boldsymbol{\epsilon} \odot \boldsymbol{x} \odot \boldsymbol{y},$$

and $\boldsymbol{\epsilon}$ is a vector whose components are bounded in modulus by ϵ.

- Given ρ scalars α_i and ρ vectors $\boldsymbol{x}_i$, $i = 1, \ldots, \rho$, then ρ vectors $\boldsymbol{\chi}_i$, $i = 1, \ldots, \rho$, with entries bounded in modulus by $\rho\,\epsilon$, exist such that

$$\begin{aligned}&fl\left(\sum_{i=1}^{\rho} \alpha_i\, \boldsymbol{x}_i\right)\\ &= \sum_{i=1}^{\rho} \alpha_i\Big(\boldsymbol{x}_i + \boldsymbol{x}_i \odot \boldsymbol{\chi}_i\Big) + O(\epsilon^2).\end{aligned} \tag{15}$$

The following stability result applies to FFT [17]:

- Given a $2n$-vector $\boldsymbol{x}$, let $\boldsymbol{y} = \mathcal{F}_{2n}\boldsymbol{x}$ and $\widetilde{\boldsymbol{y}} = fl\big(\mathcal{F}_{2n}\boldsymbol{x}\big)$, then a $2n \times 2n$ matrix Φ exists such that

$$\widetilde{\boldsymbol{y}} = \boldsymbol{y} + \Phi\,\boldsymbol{y} + O(\epsilon^2), \tag{16}$$

with $\|\Phi\|_2 \le 10.7\epsilon\log_2(2n)$. An analogous bound holds for $\mathcal{F}_{2n}^*$, with Φ replaced by a matrix Φ^*, which satisfies the same bound.

Theorem 2 shows how the computed product of a triangular Toeplitz matrix by a vector can be regarded as the exact product of a slightly perturbed matrix by the vector.

Theorem 2 *Given two n-vectors $\boldsymbol{r}$ and $\boldsymbol{v}$, let $U(\boldsymbol{r})$ be the $n \times n$ upper triangular Toeplitz matrix whose first row is $\boldsymbol{r}^T$,*

$$\boldsymbol{u} = U(\boldsymbol{r})\,\boldsymbol{v} \quad \textit{and} \quad \widetilde{\boldsymbol{u}} = fl\big(\texttt{uppert}(n, \boldsymbol{r}, \boldsymbol{v})\big).$$

Then a matrix $H(\boldsymbol{r})$ exists such that

$$\widetilde{\boldsymbol{u}} = \boldsymbol{u} + H(\boldsymbol{r})\,\boldsymbol{v} + O(\epsilon^2),$$

where

$$\|H(\boldsymbol{r})\|_2 \le \epsilon\,\gamma'\,\|\boldsymbol{r}\|_2, \tag{17}$$

with $\gamma' = 42.5\sqrt{n}\,\log_2(2n)$.

Proof. Applying algorithm `uppert` we get

$$\widetilde{\boldsymbol{z}} = fl\,\big(\mathcal{F}_{2n}\,\boldsymbol{m}\big),\ \widetilde{\boldsymbol{p}} = fl\,\big(\mathcal{F}^*_{2n}\widehat{\boldsymbol{v}}\big),$$
$$\widetilde{\boldsymbol{q}} = fl\,\big(\widetilde{\boldsymbol{z}} \odot \widetilde{\boldsymbol{p}}\big),\ \widetilde{\boldsymbol{w}} = fl\,\big(\sqrt{2n}\,\mathcal{F}_{2n}\widetilde{\boldsymbol{q}}\big).$$

Using (14) and (16) we have

$$\widetilde{\boldsymbol{z}} = \boldsymbol{z} + \Phi\,\boldsymbol{z}, \quad \widetilde{\boldsymbol{p}} = \boldsymbol{p} + \Phi^*\,\boldsymbol{p},$$
$$\widetilde{\boldsymbol{q}} = \boldsymbol{z} \odot \boldsymbol{p} + \boldsymbol{\theta}, \quad \widetilde{\boldsymbol{w}} = \sqrt{2n}\,\mathcal{F}_{2n}\widetilde{\boldsymbol{q}} + \Phi\,\boldsymbol{w},$$

where

$$\boldsymbol{\theta} = \big(\Phi\,\boldsymbol{z}\big) \odot \boldsymbol{p} + \boldsymbol{z} \odot \big(\Phi^*\,\boldsymbol{p}\big) + \boldsymbol{\epsilon} \odot \boldsymbol{z} \odot \boldsymbol{p} = \Lambda\,\boldsymbol{p},$$

with

$$\Lambda = \text{diag}\,\big(\Phi\,\boldsymbol{z}\big) + \text{diag}\,\big(\boldsymbol{z}\big)\,\Phi^* + \text{diag}\,\big(\boldsymbol{\epsilon} \odot \boldsymbol{z}\big).$$

Then

$$\begin{aligned}\widetilde{\boldsymbol{w}} &= \sqrt{2n}\mathcal{F}_{2n}\boldsymbol{q} + \sqrt{2n}\mathcal{F}_{2n}\boldsymbol{\theta} + \Phi\boldsymbol{w} \\ &= \boldsymbol{w} + \sqrt{2n}\mathcal{F}_{2n}\Lambda\boldsymbol{p} + \Phi\boldsymbol{w}.\end{aligned}$$

The vector $\widetilde{\boldsymbol{u}}$ is found in the first half of $\widetilde{\boldsymbol{w}}$. Denoting by E the first half of the identity matrix of order $2n$, we have

$$\widetilde{\boldsymbol{u}} = E^T\widetilde{\boldsymbol{w}},\ \boldsymbol{p} = \mathcal{F}^*_{2n}\,E\boldsymbol{v} \text{ and } \widetilde{\boldsymbol{u}} = \boldsymbol{u} + H(\boldsymbol{r})\boldsymbol{v}$$
$$\text{with } H(\boldsymbol{r}) = \sqrt{2n}E^T\mathcal{F}_{2n}\Lambda\mathcal{F}^*_{2n}E + E^T\Phi M,$$

where M is the circulant matrix whose first row is $[\boldsymbol{r}^T, \boldsymbol{0}^T]$. Using (12) and (16) we get

$$\begin{aligned}\|H(\boldsymbol{r})\|_2 &\le \sqrt{2n}\,\|\Lambda\|_2 + \sqrt{n}\,\|\Phi\|_2\,\|\boldsymbol{r}\|_2 \\ &\le \Big(\sqrt{2n}\,\big(\|\Phi\|_2 + \|\Phi^*\|_2 + \epsilon\big) + \sqrt{n}\,\|\Phi\|_2\Big)\,\|\boldsymbol{r}\|_2 \\ &\le 42.5\,\epsilon\,\sqrt{n}\,\log_2(2n)\,\|\boldsymbol{r}\|_2. \qquad \square\end{aligned}$$

An analogous result holds for the matrix-vector product of a lower triangular Toeplitz computed by applying algorithm `lowert`.

Theorem 3 shows how the matrix-vector product of a Toeplitz-like matrix, computed by the function `prod` of Section 3, can be regarded as the exact product of a slightly perturbed Toeplitz-like matrix by the vector.

Theorem 3 *Given an $n \times n$ Toeplitz-like matrix A with $\nabla(A) = C\,D^T$ and an n-vector $\boldsymbol{v}$, let*

$$\boldsymbol{u} = A\boldsymbol{v} \quad \textit{and} \quad \widetilde{\boldsymbol{u}} = fl\big(\texttt{prod}(n, \rho, C, D, \boldsymbol{v})\big).$$

Then a matrix Θ exists such that

$$\widetilde{\boldsymbol{u}} = \boldsymbol{u} + \Theta\,\boldsymbol{v} + O(\epsilon^2) \textit{ with } \|\Theta\|_2 \le \epsilon\,\gamma''\,\psi(C, D),$$

where $\gamma'' = \mathrm{c}\,n\,\log_2(2n)$, c not depending on n.

Proof. From (7) we have

$$\boldsymbol{u} = \sum_{i=1}^{\rho} \boldsymbol{g}_i,$$

where for $i = 1, \ldots, \rho$,

$$\boldsymbol{g}_i = L(\boldsymbol{c}_i)\boldsymbol{h}_i \quad \text{and} \quad \boldsymbol{h}_i = U(\boldsymbol{d}_i)\boldsymbol{v}.$$

The following quantities are effectively computed

$$\begin{aligned}\widetilde{\boldsymbol{h}}_i &= fl\big(\texttt{uppert}\,(n, \boldsymbol{d}_i, \boldsymbol{v})\big), \\ \widetilde{\boldsymbol{g}}_i &= fl\big(\texttt{lowert}\,(n, \boldsymbol{c}_i, \widetilde{\boldsymbol{h}}_i\big), \\ \widetilde{\boldsymbol{u}} &= fl\Big(\sum_{i=1}^{\rho} \widetilde{\boldsymbol{g}}_i\Big).\end{aligned}$$

By Theorem 2 we have

$$\widetilde{\boldsymbol{h}}_i = \boldsymbol{h}_i + H(\boldsymbol{d}_i)\,\boldsymbol{v},$$

where $\|H(\boldsymbol{d}_i)\|_2 \le \epsilon\,\gamma'\,\|\boldsymbol{d}_i\|_2$,

$$\widetilde{\boldsymbol{g}}_i = L(\boldsymbol{c}_i)\,\widetilde{\boldsymbol{h}}_i + H(\boldsymbol{c}_i)\,\boldsymbol{h}_i,$$

where $\|H(\boldsymbol{c}_i)\|_2 \le \epsilon\,\gamma'\,\|\boldsymbol{c}_i\|_2$, then

$$\widetilde{\boldsymbol{g}}_i = \boldsymbol{g}_i + \boldsymbol{\delta}_i\,\boldsymbol{v},$$

where $\boldsymbol{\delta}_i = L(\boldsymbol{c}_i)\,H(\boldsymbol{d}_i) + H(\boldsymbol{c}_i)\,U(\boldsymbol{d}_i)$. Using (12) and (17) we have

$$\|L(\boldsymbol{c}_i)\|_2 \le \sqrt{n}\,\|\boldsymbol{c}_i\|_2$$

and

$$\begin{aligned}\|\boldsymbol{\delta}_i\|_2 &\le \|L(\boldsymbol{c}_i)\|_2\,\|H(\boldsymbol{d}_i)\|_2 + \|H(\boldsymbol{c}_i)\|_2\,\|U(\boldsymbol{d}_i)\|_2 \\ &\le 2\epsilon\,\sqrt{n}\,\gamma'\,\|\boldsymbol{c}_i\|_2\,\|\boldsymbol{d}_i\|_2.\end{aligned}$$

Summing for $i = 1, \dots, \rho$ and applying (15) we get

$$\begin{aligned}\widetilde{\boldsymbol{u}} &= fl\Big(\sum_{i=1}^{\rho} \widetilde{\boldsymbol{g}}_i\Big)\\ &= \sum_{i=1}^{\rho} \boldsymbol{g}_i + \sum_{i=1}^{\rho} \Big(\boldsymbol{\delta}_i\,\boldsymbol{v} + \boldsymbol{g}_i \odot \boldsymbol{\chi}_i\Big) = \boldsymbol{u} + \Theta\,\boldsymbol{v},\end{aligned}$$

where the entries of $\boldsymbol{\chi}_i$ are bounded in modulus by $\epsilon\,\rho$ and

$$\Theta = \sum_{i=1}^{\rho} \big(\delta_i + \mathrm{diag}(\boldsymbol{\chi}_i)\, L(\boldsymbol{c}_i) U(\boldsymbol{d}_i)\big).$$

Then from (10)

$$\begin{aligned}\|\Theta\|_2 &\le \epsilon\,\big(2\sqrt{n}\,\gamma' + \rho\, n\big) \sum_{i=1}^{\rho} \|\boldsymbol{c}_i\|_2\, \|\boldsymbol{d}_i\|_2\\ &\le \epsilon\,\gamma''\psi(C, D), \text{ where } \gamma'' = 2\sqrt{n}\,\gamma' + \rho\, n.\end{aligned}$$

Taking into account the expression of γ' in (17), the thesis follows. □

If the decomposition of $\nabla(A)$ is orthogonal, then from (11) it follows

$$\|\Theta\|_2 \le \epsilon\,\gamma''\rho\,\|\nabla(A)\|_2 \le 2\,\epsilon\,\gamma''\rho\,\|A\|_2, \tag{18}$$

suggesting the stability of the algorithm `prod`.

5 Numerical experiments

The experiments, which have been conducted on an Intel Core Duo @ 3 GHz, 2GB RAM, using double precision arithmetic, have been carried out on Toeplitz-like matrices of growing size n and different displacement rank ρ.

Two sets of numerical experiments are performed, in order to validate the upper bound given in Theorem 3 by investigating the behaviour of the error produced in the computation of `prod` $(n, \rho, C, D, \boldsymbol{v})$.

(i) The matrices for the first set of experiments have been generated for different values of the displacement rank and growing values of n in the range $[2^3, 2^9]$. For each size n and fixed values of ρ, ten triples $\{C,\ D,\ \boldsymbol{v}\}$, with entries uniformly distributed in $[-10, 10]$ and $\boldsymbol{v} \neq \boldsymbol{0}$, are randomly generated. In Figure 2 the arithmetic mean μ_n of the errors $\|\widetilde{\boldsymbol{u}} - \boldsymbol{u}\|_2/\|\boldsymbol{v}\|_2$ is plotted versus n for the case $\rho = 5$ (no significant differences occurring for other values of ρ), together with the upper bound $\tau_n = \epsilon\,\gamma''\psi(C, D)$ of Theorem 3, with $\gamma'' = 85\, n\, \log_2(2n) + 5\, n$. As expected, μ_n is largely overestimated by τ_n.

(ii) For the second set of experiments we fix $n = 2^9$ and $\rho = 5$ and generate matrices C and D as in the previous case, except for the fact that different pairs of generators corresponding to the same matrix A are obtained by allowing the columns of C and D to depend on a parameter β. In this way, very different values of the function $\psi(C_\beta, D_\beta)$ occur, which increase with β. Table 1 shows that also the relative errors $e_\beta = \|\widetilde{\boldsymbol{u}} - \boldsymbol{u}\|_2/\|\boldsymbol{v}\|_2$ increase with β. However by using the orthogonal representation C_{ort}, D_{ort} of A, the function $\psi(C_{\text{ort}}, D_{\text{ort}})$ can be bounded by $\|A\|_2$ (in this example $\|A\|_2 = 355$) and consequently the relative error is reduced as suggested by (18) (see last row of Table 1).

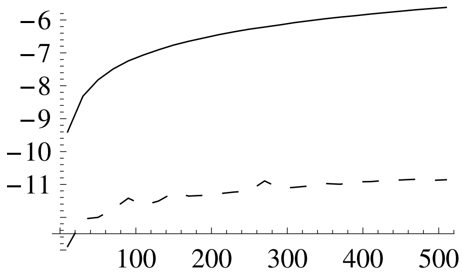

Figure 2: Log plot of μ_n(dashed line) and τ_n (solid line) as functions of n.

β	$\psi(C_\beta, D_\beta)$	e_β
10^1	260	$5.5\ 10^{-11}$
10^2	391	$6.5\ 10^{-11}$
10^3	$1.7\ 10^3$	$6.1\ 10^{-10}$
10^4	$1.5\ 10^4$	$6.7\ 10^{-8}$
10^5	$1.5\ 10^5$	$7.4\ 10^{-6}$
10^6	$1.5\ 10^6$	$5.6\ 10^{-5}$
10^7	$1.5\ 10^7$	$9.1\ 10^{-2}$
10^8	$1.5\ 10^8$	$5.4\ 10^0$
ort	247	$1.2\ 10^{-10}$

Table 1: Values of $\psi(C_\beta, D_\beta)$ and of e_β varying β. Last row shows the corresponding values for the orthogonal representation.

6 Conclusions

The numerical stability of the matrix-vector product for Toeplitz-like matrices, performed via FFT, has been analyzed. The analysis has pointed out that the error greatly depends on the magnitude of the generators of the matrix. The numerical experimentation confirms this result, suggesting that the magnitude of the generators should be monitored, and the generators should be replaced by orthogonal ones when they become too large with respect to the magnitude of the associated matrix. For example, when $\psi(C, D)$ becomes larger than $2\rho\|A\|_2$. The present study is part of a research which analyzes some superfast algorithms, like the one proposed in [7], for the solution of

Toeplitz-like systems from the stability point of view. From a theoretical error analysis, it turns out that instability and computational complexity balance, since in general larger errors tend to be produced by faster algorithms. For the near future we are planning to continue our study of this interesting field, by detecting the parameters which rule the stability behavior of iterative superfast algorithms and focusing on the conditioning properties connected with the magnitude of the generators of the matrices involved.

References:

[1] T. Kailath, A. Viera and M. Morf, "Inverses of Toeplitz operators, innovations and orthogonal polynomials", *SIAM Rev.*, 20, pp. 106-119, 1978.

[2] T. Kailath, S. Y. Kung and M. Morf, "Displacement ranks of matrices and linear equations", *J. Math. Anal. Appl.*, 68, pp. 395-407, 1979.

[3] G. Heinig and K. Rost, *Algebraic methods for Toeplitz-like matrices and operators*, Akademie-Verlag, Berlin, 1984.

[4] T. Kailath and A. H. Sayed, "Displacement structure: theory and applications", *SIAM Rev.*, 37, pp. 297-386, 1995.

[5] D. A. Bini, "Matrix structures and applications", Centre international de rencontres mathematiques, U.M.S. 822 C.N.R.S./S.M.F., 4, pp. 1-45, 2014.

[6] A. Arició and G. Rodriguez, " A fast solver for linear systems with displacement structure", *Numer. Algorithms*, 55, pp. 529-556, 2010.

[7] P. Favati, G. Lotti and O. Menchi, "A Divide and Conquer Algorithm for the Superfast solution of Toeplitz-like Systems", *SIAM. J. Matrix Anal. Appl.*, 33, pp. 1039-1056, 2012.

[8] Y. Xi, J. Xia, S.Cauley andV.Balakrishnan, "Superfast and stable structured solvers for Toeplitz least squares via randomized sampling", *SIAM. J. Matrix Anal. Appl.*, 35, pp. 44-72, 2014.

[9] R. Ke, M. K. Ng and H. W. Sun, "A fast direct method for block triangular Toeplitz-like with tridiagonal block systems from time-fractional partial differential equations", *Journal of Computational Physics*, 303, pp. 203-211, 2015.

[10] X. Lin, M. K. Ng and H. W. Sun, "A Splitting Preconditioner for Toeplitz-Like Linear Systems Arising from Fractional Diffusion Equations", *SIAM. J. Matrix Anal. Appl.*, 38, pp. 1580-1614, 2017.

[11] N. Akhoundi, "Toeplitz-like preconditioner for linear systems from spatial fractional diffusion equations", *Iranian Journal of Numerical Analysis and Optimization*, 11, pp. 95-106, 2021.

[12] D. Bini and B. Meini, "On Cyclic Reduction Applied to a Class of Toeplitz-Like Matrices Arising in Queueing Problems", in: W. J. Stewart (eds) Computations with Markov Chains. Springer, Boston, MA , 1995.

[13] A. Böttcher, C. Garoni and S. Serra-Capizzano, "Exploration of Toeplitz-like matrices with unbounded symbols is not a purely academic journey", *Sb. Math.*, 208, pp.1602-1627, 2017.

[14] A. Bostan, C. P. Jeannerod, C. Mouilleron and E. Schost, "On matrices with displacement structure: generalized operators and faster algorithms", *SIAM. J. Matrix Anal. Appl.*, 38, pp. 733-775, 2017.

[15] G. J. Groenewald, S. ter Horst, J. Jaftha and A. C. M. Ran, "A Toeplitz-Like Operator with Rational Matrix Symbol Having Poles on the Unit Circle: Fredholm Properties", *Complex Analysis and Operator Theory* 15, pp. 1-29, 2021.

[16] J. W. Cooley and O. W. Tukey, "An Algorithm for the Machine Calculation of Complex Fourier Series", *Math. Comput.*, 19, pp. 297-301, 1965.

[17] M. Arioli, H. Munthe-Kaas and L. Valdettaro, "Componentwise error analysis for FFT's with applications to fast Helmholtz solvers", *Numer. Algorithms*, 12, pp. 65-88, 1996.

[18] N. J. Higham, *Accuracy and Stability of Numerical Algorithms*, SIAM, Philadelphia, PA, 1996.

[19] P. J. Davis, *Circulant Matrices*, Wiley, New York, 1979.

[20] G. H. Golub and C. F. Van Loan, "Circulant Systems", par. 4.7.7 in *Matrix Computations* (3rd ed.), Johns Hopkins, 1996.

[21] R. M. Gray, "Toeplitz and Circulant Matrices: A Review", *Foundations and Trends in Communications and Information Theory*, 2, pp. 155-239, 2006.

[22] V. Y. Pan, Y. Rami and X. Wang, "Structured matrices and Newton's iteration: unified approach", *Linear Algebra and its Applications*, 343-344, pp. 233-265, 2002.

[23] P. Favati, G. Lotti and O. Menchi, "Stability of the Levinson algorithm for Toeplitz-like systems", *SIAM Journal on Matrix Analysis and Applications*, 31, pp. 2531-2552, 2010.

Contribution of individual authors to the creation of a scientific article (ghostwriting policy)

All the authors have contributed substantially and in equal measure to all the phases of the work reported.

Creative Commons Attribution License 4.0 (Attribution 4.0 International , CC BY 4.0)