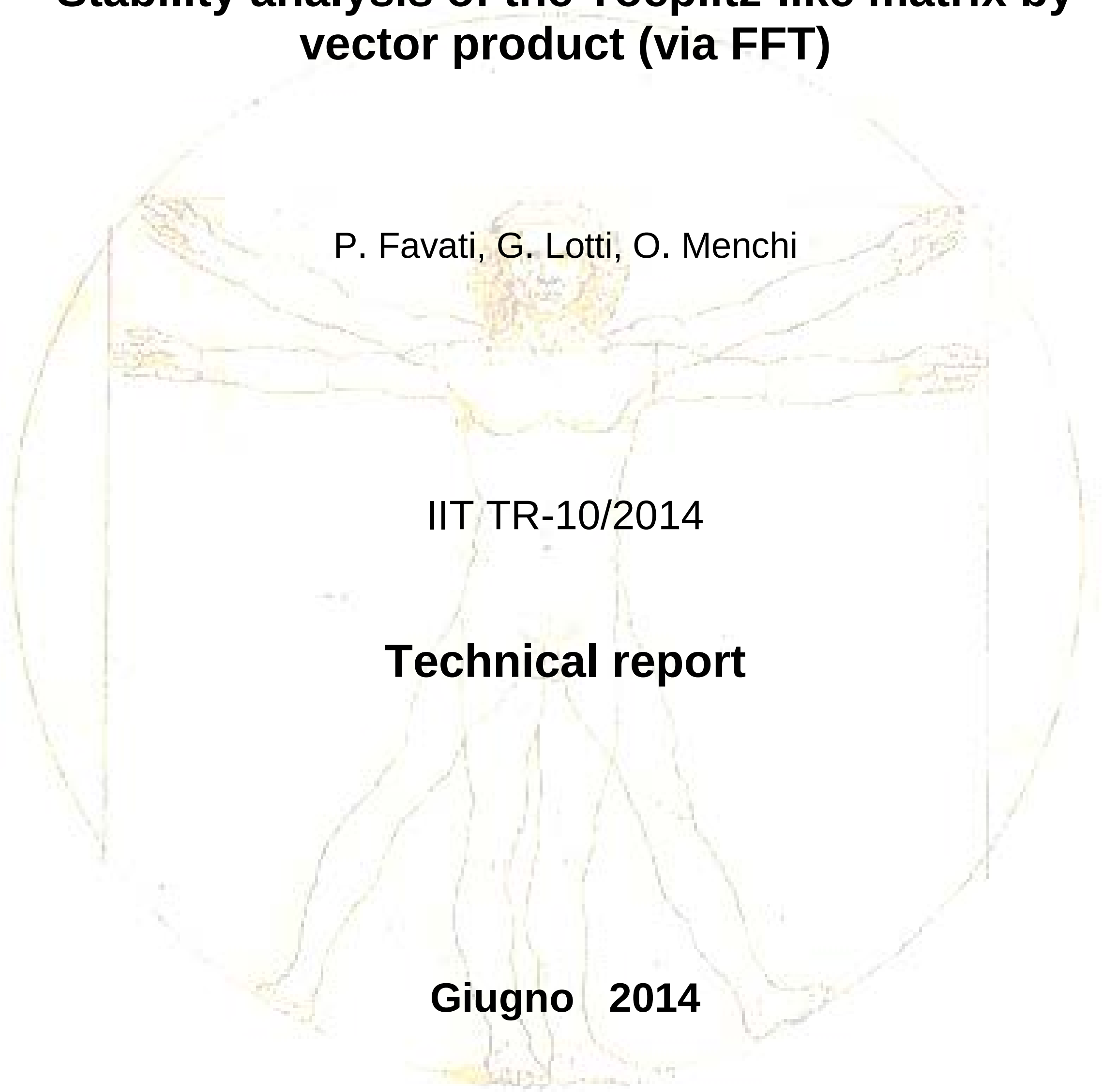
Consiglio Nazionale delle Ricerche



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

P. Favati, G. Lotti, O. Menchi

IIT TR-10/2014

Technical report

Giugno 2014

Istituto di Informatica e Telematica

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

P. Favati G. Lotti O. Menchi

Abstract

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

1 Introduction

The fast Fourier transform (FFT) was first discussed by Cooley and Tukey in 1965 [2], 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, for example in the digital signal processing, in solving partial differential equations, in the number theory, in the quick multiplication of large integers. This ubiquitous fortune is principally due to its 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 takes only $O(n \log_2 n)$ 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 (like CooleyTukey we consider here), enjoy excellent numerical stability (see [1, 4]). We are interested in investigating the stability of FFT used for multiplying a Toeplitz-like matrix by a vector. The paper is so organized: in Section 2 a brief description of Toeplitz-like matrices is given, then in Section 3 the function `prod` for computing the product is sketched. The analysis of the stability occupies Section 4, and finally in Section 5 the results of the numerical experiments are shown, confirming the relevance of the magnitude of the generators of the Toeplitz-like matrix.

2 Toeplitz-like matrices

The definition of Toeplitz-like structure is based on the concept of displacement rank [5, 6], which measures how close a matrix is to a Toeplitz matrix. Given an $n \times n$ matrix A, we consider the *displacement operator*

$$\nabla(A) = A - ZAZ^T, \tag{1}$$

where Z is the $n \times n$ *down-shift* matrix

$$Z = \begin{bmatrix} 0 & & & \\ 1 & 0 & & \\ & \ddots & \ddots & \\ & & 1 & 0 \end{bmatrix}.$$

The matrix A is said to be *Toeplitz-like* if the quantity $r_{\text{disp}}(A) = \operatorname{rank} \nabla(A)$ (called *displacement rank*) is small with respect to n (more formally $r_{\text{disp}}(A) = O(1)$ for $n \to \infty$). The set of Toeplitz-like matrices, unlike the set of Toeplitz matrices, is closed under the operations of multiplication and inversion. Let $r_{\text{disp}}(A) = \rho$, then

$$\nabla(A) = C\, D^T, \tag{2}$$

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

$$\nabla(A) = \sum_{i=1}^{\rho} \boldsymbol{c}_i \boldsymbol{d}_i^T. \tag{3}$$

with $\boldsymbol{d}_i = \boldsymbol{c}_i$ or $\boldsymbol{d}_i = -\boldsymbol{c}_i$ in the symmetric case.

For a Toeplitz matrix A of elements $a_{i,j}$ with $a_{1,1} \neq 0$ we have

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

i.e. $\boldsymbol{c}_1$ is the first column of A and $\boldsymbol{d}_2^T$ is the first row of A without the first component. In general, the displacement rank of a Toeplitz matrix is $\rho = 2$, except in some special case where $\rho = 1$.

The decomposition (2) of $\nabla(A)$, and consequently the representation of A by means of the generators, is not unique. An important representation is the *orthogonal* one [7], obtained by computing the SVD decomposition $\nabla(A) = UWV^T$, where W is the $\rho \times \rho$ diagonal matrix of the nonzero singular values $w_1 \geq \ldots \geq w_\rho > 0$ and U and V are $n \times \rho$ matrices with orthogonal columns. Now (2) is given by

$$\nabla(A) = C\, D^T, \quad \text{where} \quad C = UW^{1/2}, \quad D = VW^{1/2}.$$

The generators enable us to represent a Toeplitz-like matrix as the sum of products of lower and upper triangular Toeplitz factors. Denoting by $L(\boldsymbol{s})$ the lower triangular Toeplitz matrix whose first column is $\boldsymbol{s}$ and by $L^T(\boldsymbol{s})$ the upper triangular Toeplitz matrix whose first row is $\boldsymbol{s}$, then

$$A = \sum_{i=1}^{\rho} L(\boldsymbol{c}_i)\, L^T(\boldsymbol{d}_i). \tag{4}$$

From (1) it follows that

$$\|\nabla(A)\|_2 \le \|A\|_2 + \|ZAZ^T\|_2 \le \left(1 + \|Z\|_2^2\right) \|A\|_2 \le 2\,\|A\|_2.$$

In [3] the stability of a method for solving a linear system, whose Toeplitz-like matrix A is not explicitly given but only represented through its generators, is recognized depending on how large the generators, given in (2), are with respect to the magnitude of A. So, when stability is analyzed, we suggest to consider the function

$$\psi_2(C,D) = \sum_{i=1}^{\rho} \|\boldsymbol{c}_i\,\boldsymbol{d}_i^T\|_2 = \sum_{i=1}^{\rho} \|\boldsymbol{c}_i\|_2\,\|\boldsymbol{d}_i\|_2, \tag{5}$$

which verifies $\|\nabla(A)\|_2 \le \psi_2(C,D)$. In general it is not possible to give a bound of the ratio $\psi_2(C,D)/\|A\|_2$ depending only on n and ρ. However, if the decomposition (2) is orthogonal, then

$$\begin{aligned} &\|\nabla(A)\|_2 = w_1 = \|C\|_2\,\|D\|_2, \\ &\psi_2(C,D) = w_1 + \ldots + w_\rho \le \rho\|\nabla(A)\|_2 \le 2\rho\|A\|_2. \end{aligned} \tag{6}$$

3 Product of Toeplitz-like matrices

To compute the product of a Toeplitz-like matrix A by a vector $\boldsymbol{v}$ we exploit formula (4), so the product can be computed by multiplying first upper and then lower triangular Toeplitz matrices by vectors. If n is large, it is worthwhile to compute these products at low cost using FFT.

We consider first the upper triangular case. Given two vectors $\boldsymbol{v}^T = [v_1, \ldots, v_n]$ and $\boldsymbol{r}^T = [r_1, \ldots, r_n]$, let $L^T(\boldsymbol{r})$ be the $n \times n$ upper triangular Toeplitz matrix whose first row is $\boldsymbol{r}^T$ and let the vector $\boldsymbol{u} = L^T(\boldsymbol{r})\boldsymbol{v}$ to be computed. We note that if the last p components of $\boldsymbol{v}$ are zero, also $\boldsymbol{u}$ has the last p components equal to zero and the size of the computation can be reduced by deleting the last p columns of $L^T(\boldsymbol{r})$. For this reason we assume without loss of generality that $v_n \neq 0$.

The vectors $\boldsymbol{v}$ and $\boldsymbol{r}$ are embedded in vectors $\widehat{\boldsymbol{v}}$ and $\widehat{\boldsymbol{r}}$ of double size and the matrix $L^T(\boldsymbol{r})$ is embedded in a circulant matrix M whose first row is $\widehat{\boldsymbol{r}}^T$

$$\widehat{\boldsymbol{v}} = \begin{bmatrix} \boldsymbol{v} \\ \boldsymbol{0}_n \end{bmatrix}, \quad \widehat{\boldsymbol{r}} = \begin{bmatrix} \boldsymbol{r} \\ \boldsymbol{0}_n \end{bmatrix}, \quad M = \begin{bmatrix} L^T(\boldsymbol{r}) & L(\boldsymbol{r}') \\ L(\boldsymbol{r}') & L^T(\boldsymbol{r}) \end{bmatrix},$$

where $\boldsymbol{0}_n$ is the zero vector of length n and $L(\boldsymbol{r}')$ the $n \times n$ lower triangular Toeplitz matrix whose first column is $\boldsymbol{r}' = [0, r_n, \ldots, r_2]^T$. The vector $\boldsymbol{u}$ is found in the first half of the vector

$$\boldsymbol{m} = M\,\widehat{\boldsymbol{v}} = \begin{bmatrix} L^T(\boldsymbol{r})\boldsymbol{v} \\ L(\boldsymbol{r}')\boldsymbol{v} \end{bmatrix}.$$

A circulant matrix of order $2n$ is diagonalized by the Fourier matrix $\mathcal{F}$, whose elements are

$$f_{ij} = \frac{1}{\sqrt{2n}}\,\omega^{ij}, \quad i,j = 0,\dots,2n-1, \quad \text{with} \quad \omega = \exp(\pi \boldsymbol{i}/n).$$

Since $M = \sqrt{2n}\,\mathcal{F}\,\mathrm{diag}(\mathcal{F}\,\widehat{\boldsymbol{r}})\,\mathcal{F}^*$, it holds

$$\boldsymbol{m} = \sqrt{2n}\,\mathcal{F}\,\mathrm{diag}(\mathcal{F}\,\widehat{\boldsymbol{r}})\,\mathcal{F}^*\,\widehat{\boldsymbol{v}}.$$

Then the product $\boldsymbol{u}$ can be computed by the function

```
function $\boldsymbol{u} =$ uppert $(n, \boldsymbol{r}, \boldsymbol{v})$
$\boldsymbol{s} = \mathcal{F}\,\widehat{\boldsymbol{r}}$;   $\boldsymbol{t} = \mathcal{F}^*\widehat{\boldsymbol{v}}$;   $\boldsymbol{q} = \boldsymbol{s} \odot \boldsymbol{t}$;   $\boldsymbol{m} = \sqrt{2n}\,\mathcal{F}\boldsymbol{q}$;   $\boldsymbol{u} = \mathtt{take}(\boldsymbol{m})$;
```

where $\odot$ indicates the element-wise product of two vectors of the same size, and `take` is a function that takes the first half of a vector. The multiplications by $\mathcal{F}$ and $\mathcal{F}^*$ can be efficiently computed by calling FFT.

Moreover, since $M^T = \sqrt{2n}\,\mathcal{F}\,\mathrm{diag}\big(\overline{\mathcal{F}\,\widehat{\boldsymbol{r}}}\big)\,\mathcal{F}^*$, where the overline indicates the conjugate, we can compute $\boldsymbol{y} = L(\boldsymbol{r})\boldsymbol{v}$ as follows

```
function $\boldsymbol{y} =$ lowert $(n, \boldsymbol{r}, \boldsymbol{v})$
$\boldsymbol{s} = \mathcal{F}\,\widehat{\boldsymbol{r}}$;   $\boldsymbol{t} = \mathcal{F}^*\widehat{\boldsymbol{v}}$;   $\boldsymbol{q} = \overline{\boldsymbol{s}} \odot \boldsymbol{t}$;   $\boldsymbol{m} = \sqrt{2n}\,\mathcal{F}\boldsymbol{q}$;   $\boldsymbol{y} = \mathtt{take}(\boldsymbol{m})$;
```

Let now A be a Toeplitz-like matrix with generators C and D and displacement rank ρ. Then by (4) it is

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

and we can compute $A\boldsymbol{v}$ by the following function, where $\{A\}$ denotes the set $\{n, \rho, C, D\}$.

```
function $\boldsymbol{u} =$ prod $(\{A\}, \boldsymbol{v})$
%      $A$ has displacement rank $\rho$ and generators $C$ and $D$
%      $C$ and $D$ have columns $\boldsymbol{c}_i$ and $\boldsymbol{d}_i$
for $i = 1 : \rho$
   $\boldsymbol{h}_i =$ uppert $(n, \boldsymbol{d}_i, \boldsymbol{v})$;   $\boldsymbol{g}_i =$ lowert $(n, \boldsymbol{c}_i, \boldsymbol{h}_i)$;
end
$\boldsymbol{u} = \sum_{i=1}^{\rho} \boldsymbol{g}_i$;
```

A saving of the cost can be achieved by skipping the last FFT call of `lowert` and exploiting the linearity of $\mathcal{F}$ in the final sum. A simplified version of `prod` can be used to reconstruct a single column of A or the whole A according to (4).

The function $\texttt{prod}(\{A\}, V)$ to multiply A by the matrix V can be easily derived from the function $\texttt{prod}(\{A\}, \boldsymbol{v})$ given above. Obviously, when a same matrix A has to be multiplied by several different vectors, the transformation of the generators is performed only once.

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:

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

$$\boldsymbol{x} = \widetilde{\boldsymbol{x}} - \widetilde{\boldsymbol{x}} \odot \boldsymbol{\epsilon} + O(\epsilon^2). \tag{8}$$

- 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

$$fl\left(\widetilde{\boldsymbol{x}} \odot \widetilde{\boldsymbol{y}}\right) = \boldsymbol{x} \odot \boldsymbol{y} + \boldsymbol{\omega} + O(\epsilon^2), \tag{9}$$

where

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

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

$$fl\Big(\sum_{i=1}^{\rho} \alpha_i\, \boldsymbol{x}_i\Big) = \sum_{i=1}^{\rho} \alpha_i \Big(\boldsymbol{x}_i + \boldsymbol{x}_i \odot \boldsymbol{\chi}_i\Big) + O(\epsilon^2). \tag{10}$$

The following stability result applies to FFT [1]:

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

$$\widetilde{\boldsymbol{y}} = \boldsymbol{y} + \Phi\,\boldsymbol{y} + O(\epsilon^2), \quad \text{with} \quad \|\Phi\|_2 \le 10.7\,\epsilon\,\log_2(2n). \tag{11}$$

An analogous bound holds for $\mathcal{F}^*$, with Φ replaced by a matrix Φ^*, which satisfies the same bound.

The first theorem 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 1 *Given the n-vectors $\boldsymbol{r}$ and $\boldsymbol{v}$, let $\boldsymbol{u} = L^T(\boldsymbol{r})\,\boldsymbol{v}$ and let*

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

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

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

where

$$\|H(\boldsymbol{r})\|_2 \le \epsilon\,\gamma'\,\|\boldsymbol{r}\|_2, \quad \gamma' = 42\sqrt{n}\,\log_2(2n). \tag{12}$$

Proof. Applying algorithm `uppert` we get

$$\widetilde{\boldsymbol{s}} = fl\,\big(\mathcal{F}\widehat{\boldsymbol{r}}\big), \quad \widetilde{\boldsymbol{t}} = fl\,\big(\mathcal{F}^*\widehat{\boldsymbol{v}}\big), \quad \widetilde{\boldsymbol{g}} = fl\,\big(\widetilde{\boldsymbol{s}} \odot \widetilde{\boldsymbol{t}}\big), \quad \widetilde{\boldsymbol{m}} = fl\,\big(\sqrt{2n}\,\mathcal{F}\widetilde{\boldsymbol{g}}\big).$$

The vector $\widetilde{\boldsymbol{u}}$ is found in the first half of $\widetilde{\boldsymbol{m}}$. Using (9) and (11) we have

$$\widetilde{\boldsymbol{s}} = \boldsymbol{s} + \varPhi\,\boldsymbol{s}, \quad \widetilde{\boldsymbol{t}} = \boldsymbol{t} + \varPhi^*\,\boldsymbol{t}, \quad \widetilde{\boldsymbol{g}} = \boldsymbol{s} \odot \boldsymbol{t} + \boldsymbol{\omega}, \quad \widetilde{\boldsymbol{m}} = \sqrt{2n}\,\mathcal{F}\widetilde{\boldsymbol{g}} + \varPhi\,\boldsymbol{m},$$

where

$$\boldsymbol{\omega} = \big(\varPhi\,\boldsymbol{s}\big) \odot \boldsymbol{t} + \boldsymbol{s} \odot \big(\varPhi^*\,\boldsymbol{t}\big) + \boldsymbol{\epsilon} \odot \boldsymbol{s} \odot \boldsymbol{t} = \varOmega\,\boldsymbol{t},$$

with

$$\varOmega = \operatorname{diag}\big(\varPhi\,\boldsymbol{s}\big) + \operatorname{diag}\big(\boldsymbol{s}\big)\,\varPhi^* + \operatorname{diag}\big(\boldsymbol{\epsilon} \odot \boldsymbol{s}\big).$$

Then

$$\widetilde{\boldsymbol{m}} = \sqrt{2n}\,\mathcal{F}\boldsymbol{g} + \sqrt{2n}\,\mathcal{F}\boldsymbol{\omega} + \varPhi\,\boldsymbol{m} = \boldsymbol{m} + \sqrt{2n}\,\mathcal{F}\,\varOmega\,\boldsymbol{t} + \varPhi\,\boldsymbol{m}.$$

Denoting by E the upper half of the identity matrix of order $2n$, we have

$$\widetilde{\boldsymbol{u}} = \boldsymbol{u} + \boldsymbol{z},$$

where

$$\boldsymbol{z} = H(\boldsymbol{r})\,\boldsymbol{v}, \quad \text{with} \quad H(\boldsymbol{r}) = \sqrt{2n}\,E\,\mathcal{F}\,\varOmega\,\mathcal{F}^*\,E^T + E\,\varPhi \begin{bmatrix} L^T(\boldsymbol{r}) \\ L(\boldsymbol{r}') \end{bmatrix}.$$

Using (11) we get

$$\begin{aligned}
\|H(\boldsymbol{r})\|_2 &\le \sqrt{2n}\,\|\varOmega\|_2 + \|\varPhi\|_2\,\|\boldsymbol{r}\|_1 \\
&\le \sqrt{2n}\,\big(\|\varPhi\|_2 + \|\varPhi^*\|_2 + \epsilon\big)\,\|\boldsymbol{r}\|_2 + \sqrt{n}\|\varPhi\|_2\,\|\boldsymbol{r}\|_2 \\
&\le \epsilon\,\sqrt{n}\big(10.7(2\sqrt{2}+1)\log_2(2n) + 1\big)\,\|\boldsymbol{r}\|_2 \\
&\le 42\,\epsilon\,\sqrt{n}\,\log_2(2n)\,\|\boldsymbol{r}\|_2. \qquad \square
\end{aligned}$$

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

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

Theorem 2 *Given an $n \times n$ Toeplitz-like matrix A with $\nabla(A) = C\,D^T$ and an n-vector $\boldsymbol{v}$, let $\boldsymbol{u} = A\boldsymbol{v}$ and let $\widetilde{\boldsymbol{u}} = fl\big(\texttt{prod}(\{A\}, \boldsymbol{v})\big)$ be the computed product. Then a matrix Θ_A exists such that*

$$\widetilde{\boldsymbol{u}} = \boldsymbol{u} + \Theta_A\,\boldsymbol{v} + O(\epsilon^2) \quad \textit{with} \quad \|\Theta_A\|_2 \le \epsilon\,\gamma''\,\psi_2(C, D). \tag{13}$$

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

Proof. From (7) we have

$$\boldsymbol{u} = \sum_{i=1}^{\rho} \boldsymbol{g}_i, \quad \text{where} \quad \boldsymbol{g}_i = L(\boldsymbol{c}_i)\boldsymbol{h}_i, \quad \boldsymbol{h}_i = L^T(\boldsymbol{d}_i)\boldsymbol{v}, \quad \text{for} \quad i = 1, \ldots, \rho.$$

The following quantities are effectively computed

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

By Theorem 1 we have

$$\widetilde{\boldsymbol{h}}_i = \boldsymbol{h}_i + H(\boldsymbol{d}_i)\,\boldsymbol{v}, \quad \text{where} \quad \|H(\boldsymbol{d}_i)\|_2 \le \epsilon\,\gamma'\,\|\boldsymbol{d}_i\|_2,$$

and

$$\widetilde{\boldsymbol{g}}_i = L(\boldsymbol{c}_i)\,\widetilde{\boldsymbol{h}}_i + H(\boldsymbol{c}_i)\,\boldsymbol{h}_i, \quad \text{where} \quad \|H(\boldsymbol{c}_i)\|_2 \le \epsilon\,\gamma'\,\|\boldsymbol{c}_i\|_2,$$

then

$$\widetilde{\boldsymbol{g}}_i = \boldsymbol{g}_i + \boldsymbol{\delta}_i\,\boldsymbol{v}, \quad \text{where} \quad \boldsymbol{\delta}_i = L(\boldsymbol{c}_i)\,H(\boldsymbol{d}_i) + H(\boldsymbol{c}_i)\,L^T(\boldsymbol{d}_i).$$

Since $\|L(\boldsymbol{c}_i)\|_1 = \|L(\boldsymbol{c}_i)\|_\infty = \|\boldsymbol{c}_i\|_1$, we have

$$\|L(\boldsymbol{c}_i)\|_2 \le \sqrt{\|L(\boldsymbol{c}_i)\|_1\,\|L(\boldsymbol{c}_i)\|_\infty} = \|\boldsymbol{c}_i\|_1 \le \sqrt{n}\,\|\boldsymbol{c}_i\|_2$$

and

$$\|\boldsymbol{\delta}_i\|_2 \le \|L(\boldsymbol{c}_i)\|_2\,\|H(\boldsymbol{d}_i)\|_2 + \|H(\boldsymbol{c}_i)\|_2\,\|L^T(\boldsymbol{d}_i)\|_2 \le 2\epsilon\,\sqrt{n}\,\gamma'\,\|\boldsymbol{c}_i\|_2\,\|\boldsymbol{d}_i\|_2.$$

Summing for $i = 1, \ldots, \rho$ and applying (10) we get

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

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

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

Then, since $\rho = O(1)$ for $n \to \infty$,

$$\|\Theta_A\| \le \epsilon\,\big(2\,\sqrt{n}\,\gamma' + \rho\,n\big) \sum_{i=1}^{\rho} \|\boldsymbol{c}_i\|_2\,\|\boldsymbol{d}_i\|_2 \le \epsilon\,\gamma''\psi_2(C, D). \qquad \square$$

If the decomposition of $\nabla(A)$ is orthogonal, then from (6) it follows

$$\|\Theta_A\| \le \epsilon\,\gamma''\rho\,\|\nabla(A)\|_2 \le 2\,\epsilon\,\gamma''\rho\,\|A\|_2.$$

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 2 by investigating the behaviour of the relative error produced in the computation of $\texttt{prod}\ (\{n, \rho, C, D\}, \boldsymbol{v})$.

(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 $[-10, 10]$ are randomly generated. The arithmetic mean μ_n of the relative errors $\|\widetilde{\boldsymbol{u}} - \boldsymbol{u}\|/\|\boldsymbol{v}\|$ is plotted versus n in Figure 1 (for the case $\rho = 5$, no significant differences occur for other values of ρ), together with the upper bound $\tau_n = \epsilon\,\gamma''\psi_2(C, D)$ of Theorem 2.

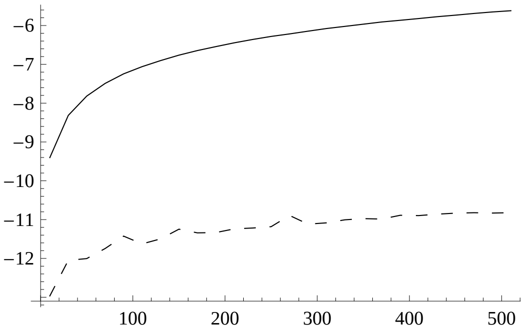


Figure 1: Log plot of μ_n(dashed line) and τ_n (solid line) as functions of n.

(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 $\psi_2(C_\beta, D_\beta)$ occur, which increase with β. Table 1 shows that also the relative errors $e_\beta = \|\widetilde{\boldsymbol{u}} - \boldsymbol{u}\|/\|\boldsymbol{v}\|$ increase with β. However by using the orthogonal representation C_{ort}, D_{ort} of A, the function $\psi_2(C_{ort}, D_{ort})$ can be bounded by $\|A\|_2$ (in this example $\|A\|_2 = 355$) and consequently the relative error is reduced according to the bound of Theorem 2 (see last row of Table 1).

6 Conclusions

The numerical stability of the Toeplitz-like matrix by vector product, performed via FFT, has been analyzed. The analysis has pointed out that the error greatly

β	$\psi_2(C_\beta, D_\beta)$	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 $\psi_2(C_\beta, D_\beta)$ and of e_β varying β. Last row shows the corresponding values for the orthogonal representation.

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 becomes too large with respect to the magnitude of the associated matrix.

References

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

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

[3] 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.

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

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

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

[7] 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.