Università di Pisa
Dipartimento di Informatica

Technical Report: TR-02-04

The componentwise conditioning of the DFT

Enrico Bozzo Dario Fasino Ornella Menchi

April 10, 2002

Address: Corso Italia 40, 56125 Pisa, Italy. Tel: +39 050 2212700 Fax: +39 050 2212726

*Dipartimento di Informatica, Università di Pisa, Corso Italia 40, 56125 Pisa, Italy.
‡Dipartimento di Matematica e Informatica, Università di Udine, Viale delle Scienze 208, 33100 Udine, Italy.

The componentwise conditioning of the DFT

Enrico Bozzo* Dario Fasino‡ Ornella Menchi *

April 10, 2002

Abstract

Mixed and componentwise condition numbers are useful in order to understand stability properties of algorithms for solving structured linear systems. The DFT (discrete Fourier transform) is an essential building block of these algoritms. We obtain precise estimates of mixed and componentwise condition numbers of the DFT. To this end we explicitly compute certain unimodular vectors (a complex vector is said unimodular if the modulus of all its entries is equal to one) whose DFT is again unimodular.

1 Introduction

The *relative componentwise distance* between two points $x, \tilde{x} \in \mathbb{C}^n$ is defined as

$$\delta(x, \tilde{x}) = \min\{\varepsilon : |x_i - \tilde{x}_i| \leq \varepsilon |x_i|, 1 \leq i \leq n\}. \tag{1}$$

setting $\delta(x, \tilde{x}) = +\infty$ whenever the above minimum is not defined. This definition is of interest in numerical analysis, because it is the most appropriate way to measure errors induced by the finite precision representation of machine numbers. If an entry of x is zero then the corresponding entry in $\tilde{x}$ must vanish for $\delta(x, \tilde{x})$ be finite, that is, $\delta(x, \tilde{x})$ can be considered a measure of the largest relative error in the entries x with respect to $\tilde{x}$. Indeed, if $\epsilon = \delta(x, \tilde{x})$ is finite, there exists a diagonal matrix D such that $x - \tilde{x} = Dx$ and $\|D\|_\infty = \epsilon$. Furthermore, if Φ is an arbitrary continuous map from $\mathbb{C}^n$ to itself, and $\|\cdot\|$ is a vector norm, the *componentwise condition number* of Φ in x is defined as

$$c(\Phi, x) = \lim_{\varepsilon \to 0} \sup_{\delta(x,\tilde{x}) \leq \varepsilon} \frac{\delta(\Phi(x), \Phi(\tilde{x}))}{\delta(x, \tilde{x})}. \tag{2}$$

The last quantity is a measure of the entrywise sensitivity of Φ to entrywise relative perturbations in its variables, while the usual definition of conditioning,

$$\kappa(\Phi, x) = \lim_{\varepsilon \to 0} \sup_{\delta(x,\tilde{x}) \leq \varepsilon} \frac{\|\Phi(x) - \Phi(\tilde{x})\|}{\|\Phi(x)\|} \frac{\|x\|}{\|x - \tilde{x}\|}$$

is a measure of the sensitivity of Φ to normwise perturbations. Furthermore, the *mixed condition number* of Φ in x is defined as

$$m(\Phi, x) = \lim_{\varepsilon \to 0} \sup_{\delta(x,\tilde{x}) \le \varepsilon} \frac{\|\Phi(x) - \Phi(\tilde{x})\|}{\|\Phi(x)\|} \frac{1}{\delta(x, \tilde{x})}. \tag{3}$$

The above definitions are borrowed from [5, 6]. The main reason of studying componentwise condition numbers is the analysis of stability characteristics of algorithms for structured linear systems. Indeed, these refinements of the classical condition number were introduced and exploited in the analysis of fast algorithms for the solution of Cauchy and Vandermonde matrices, see [2, 5, 6, 7, 8] and references therein, in order to take in due consideration both accurate measures of data errors and the structure of the problem at hand.

In this paper we address the problem of quantifying the componentwise conditioning of the map associated to the discrete Fourier transform (DFT) of a complex vector. Moreover we study the mixed condition number of the DFT in the important case were the infinity norm is used in the definition. In order to stress this point the symbol m will be specialized as m_∞.

Let $\omega = e^{i2\pi/n}$ be a primitive root of unit, where i is the complex unit such that $i^2 = -1$ and $n \ge 2$ is an integer. Let F be the Fourier matrix of order n, having entries $F_{k,j} = n^{-1/2}\omega^{kj}$ with k, $j = 0, \ldots, n-1$. Furthermore, let $\mathrm{Diag}(v)$ denote the diagonal matrix whose diagonal entries are those of a given vector $v \in \mathbb{C}^n$. The following result was shown in [4], in the context of conditioning of Cauchy and Vandermonde matrices:

Theorem 1 *Let $\Phi : \mathbb{C}^n \mapsto \mathbb{C}^n$ be the map $\Phi(v) = Fv$. Then, $m(\Phi, v) \le \sqrt{n}$.*

Proof. It is known that

$$m_\infty(\Phi, v) = \|F\mathrm{Diag}(v)\|_\infty / \|Fv\|_\infty. \tag{4}$$

Indeed, the above equation holds when Φ is any linear map, see [6]. From Cauchy-Schwarz inequality we have

$$\|F\mathrm{Diag}(v)\|_\infty = \max_k \sum_j |F_{k,j} v_j| = \frac{1}{\sqrt{n}} \sum_j |v_j| \le \|v\|_2.$$

Since F is unitary, we have $\|Fv\|_2 = \|v\|_2$, hence

$$m_\infty(\Phi, v) \le \frac{\|v\|_2}{\|Fv\|_\infty} \le \sqrt{n} \frac{\|v\|_2}{\|Fv\|_2} = \sqrt{n},$$

and the proof is complete. ■

Numerical stability of some fast algorithms to compute the Discrete Fourier Transform were analyzed e.g., in [1]; one of the bounds for the algorithmic error derived in that paper has the following form: If $w = Fv$ and $\tilde{w}$ is the vector computed in finite precision arithmetic by such algorithms, then

$$\frac{\|w - \tilde{w}\|_\infty}{\|\tilde{w}\|_\infty} \le 10.7\, \varepsilon \sqrt{n} + O(\varepsilon^2),$$

where ε is the machine precision. If the bound in Theorem 1 is sharp, then we could conclude that the above upper bound for the algorithmic error in w is compatible with a small relative componentwise perturbation of the data vector v. Furthermore, in the paper [9] the following bound was derived for the relative error in the individual entries of the vector $\tilde{w}$ computed by a large class of fast algorithms:

$$\frac{|\tilde{w}_j - w_j|}{|\tilde{w}_j|} \leq C_n \frac{\|v\|_1}{\sqrt{n}|\tilde{w}_j|}\varepsilon + O(\varepsilon^2),$$

where C_n is a constant depending on the particular algorithm used to compute the twiddle factors of the DFT. Actually, the factor $\sqrt{n}$ does not appear in the above mentioned paper, due to a different scaling of the matrix F. We will show that the factor $\|v\|_1/(\sqrt{n}|\tilde{w}_j|)$ appearing in the above inequality is exactly the condition number for the computation of w_j.

How much sharp is the estimate in Theorem 1? And what can be said about the componentwise conditioning $c(\Phi, x)$ for the same map?

Let us denote with e the vector having all the entries equal to one. A vector a is said to be *unimodular* if $|a| = e$ where the modulus is to be intended entrywise. Moreover, let $\mathbb{T}^n$ be the set of all unimodular vectors of order n. In Section 2 we find certain special vectors $a \in \mathbb{T}^n$ such that $Fa \in \mathbb{T}^n$ and in Section 3 we will use this result in order to answer the preceding questions.

2 Unimodular vectors having unimodular DFT

The following properties of the matrix F are well known, see e.g., [3]:

(a) F is unitary, i.e., $F^*F = I$, moreover $F^* = \overline{F}$;

(b) $F^2 = K$ where K is a permutation matrix such that $K_{k,j} = 1$ if and only if $k = j = 0$ or $k + j = n$;

(c) if D is a diagonal matrix then $C = FDF^*$ is a circulant matrix and more precisely $C = F\,\mathrm{Diag}(\sqrt{n}F^*Ce_0)F^*$, where e_0 is the first canonical vector.

By virtue of the last property, given a vector v and denoting with $\circ$ the entrywise (Hadamard) vector product, it holds that

$$Cv = F((\sqrt{n}F^*Ce_0) \circ (F^*v)) \tag{5}$$

which is a discrete form of the convolution theorem. Indeed, the product Cv is the (circular) convolution between the vectors v and Ce_0, and the above equation states that the inverse Fourier transform of Cv equals the componentwise product of the Fourier transform of c and of Ce_0, in perfect analogy with the continuous case. Let $a \in \mathbb{T}^n$. The relation $Fa \in \mathbb{T}^n$ holds if and only if $Fa \circ \overline{Fa} = e$. By writing the preceding relation in the form $F^*FFa \circ \overline{F}\,\overline{a} = e$ and using property (b) above, we obtain $F^*Ka \circ F^*\overline{a} = e$. The multiplication by $\sqrt{n}F$ yields

$$F(\sqrt{n}F^*Ka \circ F^*\overline{a}) = \sqrt{n}Fe = ne_0.$$

Let C be the circulant matrix whose first column is Ka, $C = F\mathrm{Diag}(\sqrt{n}F^*\overline{a})F^*$. By using (5) we see that the above equation can be restated as $C\overline{a} = ne_0$, i.e.,

$$\begin{pmatrix} a_0 & a_1 & \cdots & a_{n-1} \\ a_{n-1} & a_0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & a_1 \\ a_1 & \cdots & a_{n-1} & a_0 \end{pmatrix} \begin{pmatrix} \overline{a}_0 \\ \overline{a}_1 \\ \vdots \\ \overline{a}_{n-1} \end{pmatrix} = \begin{pmatrix} n \\ 0 \\ \vdots \\ 0 \end{pmatrix}. \tag{6}$$

We remark that the matrix $n^{-1/2}C$ is unitary, since its spectral factorization is $n^{-1/2}C = F\mathrm{Diag}(Fa)F^*$ and $Fa \in \mathbb{T}^n$ by hypothesis. Hence the problem of finding a unimodular vector a such that Fa is also unimodular could also be restated as follows: Find a circulant matrix of order n which is unitary and has all the entries of the same modulus (that must be $n^{-1/2}$). The first equation in (6) is satisfied whenever $a \in \mathbb{T}^n$. The k-th equation can be rewritten as $\sigma_k = 0$ where

$$\sigma_k = \sum_{j=0}^{n-1-k} a_j \overline{a}_{k+j} + \sum_{j=0}^{k-1} \overline{a}_j a_{n-k+j}, \qquad k = 1, \ldots, n-1.$$

It is useful to extend by periodicity the vector a by setting $a_{n+j} = a_j$ for every integer j, so that

$$\sigma_k = \sum_{j=0}^{n-1} a_j \overline{a}_{k+j}. \tag{7}$$

Without loss of generality we set $a_j = \omega^{\varphi(j)}$ where φ is a real function; the periodicity assumption $a_{n+j} = a_j$ implies that

$$\forall j \in \mathbb{Z}\ \exists h \in \mathbb{Z} : \quad \varphi(n+j) - \varphi(j) = nh. \tag{8}$$

Now (7) takes the form

$$\sigma_k = \sum_{j=0}^{n-1} \omega^{\varphi(j) - \varphi(k+j)},$$

and our task is to find a suitable φ satisfying the condition (8), such that

$$\sigma_1 = \ldots = \sigma_{n-1} = 0. \tag{9}$$

Let us try to guess a valid form restricting the search to the class of polynomials where equation (8) is clearly satisfied. First of all, let us observe that a valid φ remains valid if we add to it an arbitrary constant. One sees almost immediately that linear polynomials do not fulfil (9). The choice $\varphi(j) = c\,j^2$ leads to

$$\sigma_k = \sum_{j=0}^{n-1} \omega^{-ck^2 - 2ckj} = \omega^{-ck^2} \sum_{j=0}^{n-1} \omega^{-2ckj}.$$

Hence, let c be any integer such that $gcd(c, n) = 1$. If n is odd, then $\omega^{-2ck} \neq 1$ for every $k = 0, \ldots, n-1$ and

$$\sigma_k = \omega^{-ck^2} \frac{1 - \omega^{-2ckn}}{1 - \omega^{-2ck}} = 0.$$

However if n is even and $k = n/2$, then $\omega^{-2ck} = \omega^{-cn} = 1$ and $\sigma_{n/2}$ is no longer equal to zero. Nevertheless, if n is even we can set $\varphi(j) = \frac{1}{2}cj^2$ and this again implies (9). In summary, we proved the following result:

Theorem 2 *Given an integer c such that $gcd(c, n) = 1$ and a real number α, the vector a with entries*

$$a_j = \begin{cases} \omega^{cj^2+\alpha}, & \text{if } n \text{ is odd,} \\ \omega^{cj^2/2+\alpha}, & \text{if } n \text{ is even,} \end{cases}$$

satisfies system (6), i.e., a is unimodular and has unimodular DFT.

3 Mixed and componentwise condition numbers of the DFT

We can now obtain sharp estimates for mixed and componentwise condition numbers of the DFT. Their sharpness is a consequence of the existence of unimodular vectors having unimodular DFT, shown in the preceding section. Owing to property (a), the following results hold also for the Inverse DFT, defined by the matrix F^*.

Corollary 1 *For the mixed condition number of the DFT it holds that*

$$\max_{v \neq 0} m_\infty(\Phi, v) = \sqrt{n},$$

hence the bound in Theorem 1 is optimal. Moreover,

$$\min_{v \neq 0} m_\infty(\Phi, v) = 1.$$

Proof. Let $v \in \mathbb{T}^n$ be a vector such that $Fv \in \mathbb{T}^n$. Then,

$$\|F\mathrm{Diag}(v)\|_\infty = \frac{1}{\sqrt{n}} \sum_j |v_j| = \sqrt{n}.$$

Furthermore, $\|Fv\|_\infty = 1$ and using (4) we obtain the first claim. Furthermore,

$$\|Fv\|_\infty = \|F\mathrm{Diag}(v)e\|_\infty \leq \|F\mathrm{Diag}(v)\|_\infty \|e\|_\infty = \|F\mathrm{Diag}(v)\|_\infty,$$

thus from (4) we have $m_\infty(\Phi, v) \geq 1$. Since equality is attained when v is a canonical basis vector, we complete the proof. ■

Corollary 2 *For the componentwise condition number of the DFT it holds that*

$$c(\Phi, v) = \frac{\|v\|_1}{\sqrt{n}\min_j |w_j|} \geq \sqrt{n}.$$

The lower bound is attainable.

PROOF. For any nonzero vector $v = (v_1, \ldots, v_n)^T \in \mathbb{C}^n$, let $w = (w_1, \ldots, w_n)^T = Fv$. Then we have

$$c(\Phi, v) = \|\mathrm{Diag}(w)^{-1} F \mathrm{Diag}(v)\|_\infty,$$

see [6]. In the preceding equation, we set $c(\Phi, v) = +\infty$ whenever w has some zero entries. Without loosing in generality, we can assume $\|v\|_2 = 1$, hence $\|w\|_2 = 1$. Using simple inequalities, we find

$$\begin{aligned} c(\Phi, v) &= \frac{1}{\sqrt{n}} \max_{1 \leq k \leq n} \frac{1}{|w_k|} \sum_{j=1}^{n} |v_j| \\ &\geq \max_{1 \leq i \leq n} \frac{1}{|w_i|} \\ &\geq \sqrt{n}. \end{aligned}$$

Finally, let $v \in \mathbb{T}^n$ be a vector such that $w = Fv \in \mathbb{T}^n$. Then,

$$c(\Phi, v) = \|\mathrm{Diag}(w)^{-1} F \mathrm{Diag}(v)\|_\infty = \|F\|_\infty = \sqrt{n},$$

and we have the claim. ■

References

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

[2] S. G. Bartels and D. J. Higham; The structured sensitivity of Vandermonde-like systems. *Numer. Math.* 62 (1992), 17–33.

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

[4] D. Fasino, Structured conditioning of Cauchy and Vandermonde matrices, submitted to *Lin. Algebra Appl.*, 2002.

[5] I. Gohberg and I. Koltracht; On the inversion of Cauchy matrices, in: *Signal processing, scattering and operator theory, and numerical methods* (Amsterdam, 1989), 381–392, Progr. Systems Control Theory, 5, Birkhäuser Boston, Boston, MA, 1990.

[6] I. Gohberg and I. Koltracht; Mixed, componentwise, and structured condition numbers. *SIAM J. Matrix Anal. Appl.* 14 (1993), 688–704.

[7] N. J. Higham; Error analysis of the Björck-Pereyra algorithms for solving Vandermonde systems. *Numer. Math.* 50 (1987), 613–632.

[8] D. J. Higham and N. J. Higham; Backward error and condition of structured linear systems. *SIAM J. Matrix Anal. Appl.* 13 (1992), 162–175.

[9] P. Yalamov; *Normwise and componentwise stability of the Fast Fourier Transform*, Preprint N41, University of Rousse, August 1998.