Componentwise error analysis for FFTs with applications to fast Helmholtz solvers

M. Arioli[a], H. Munthe-Kaas[b] and L. Valdettaro[c]

[a] *Istituto di Analisi Numerica, Consiglio Nazionale delle Ricerche, via Abbiategrasso 209, I-27100 Pavia, Italy*

[b] *Institute for Informatics, University of Bergen, N-5020 Bergen, Norway*

[c] *Dipartimento di Matematica, Politecnico di Milano, piazza Leonardo da Vinci, I-20133 Milano, Italy*

Received 25 March 1994; revised 15 June 1995
Communicated by G.H. Golub

We analyze the stability of the Cooley–Tukey algorithm for the Fast Fourier Transform of order $n = 2^k$ and of its inverse by using componentwise error analysis.

We prove that the components of the roundoff errors are linearly related to the result in exact arithmetic. We describe the structure of the error matrix and we give optimal bounds for the total error in infinity norm and in L_2 norm.

The theoretical upper bounds are based on a "worst case" analysis where all the rounding errors work in the same direction. We show by means of a statistical error analysis that in realistic cases the max-norm error grows asymptotically like the logarithm of the sequence length by machine precision.

Finally, we use the previous results for introducing tight upper bounds on the algorithmic error for some of the classical fast Helmholtz equation solvers based on the Faster Fourier Transform and for some algorithms used in the study of turbulence.

1. Introduction

Let F_n be the Fourier matrix of order n, i.e.

$$(F_n)_{j,k} = e^{2\pi ijk/n}; \qquad 0 \le j,k < n.$$

The one dimensional discrete Fourier transform is defined as the matrix vector product

$$\hat{x} = F_n \cdot x.$$

The one dimensional inverse transform is given as

$$\check{x} = \frac{1}{n} F_n^* \cdot x,$$

© J.C. Baltzer AG, Science Publishers

and the d-dimensional transform as

$$\hat{x} = F_{n_d} \otimes F_{n_{d-1}} \otimes \cdots \otimes F_{n_1} \cdot x,$$

where $\otimes$ denotes the matrix tensor product (see e.g. [3,10] for its definition and properties).

We henceforth assume that all n_j are powers of two. The Cooley–Tukey Fast Fourier Transform (FFT) [4,5] is based on the following factorization of F_n:

$$B_n \cdot F_n = D_k(I_{n/2} \otimes F_2)D_{k-1}(I_{n/4} \otimes F_2 \otimes I_2) \otimes D_2(I_2 \otimes F_2 \otimes I_{n/4})D_1(F_2 \otimes I_{n/2}), \quad (1)$$

where B_n is a permutation matrix described below, D_j are diagonal matrices and F_2 is the order 2 Fourier matrix, i.e.

$$F_2 = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}.$$

The diagonal entries of D_j are all of the form

$$(D_j)_{k,k} = e^{2\pi i \theta_k^j},$$

where θ_k^j are real. The matrices D_j are called *twiddle matrices* and $e^{2\pi i \theta_k^j}$ *twiddle factors*. The permutation matrix B_n performs a *bit reversal permutation* of a vector, i.e. it swaps two elements $x(j)$ and $x(k)$, whenever the binary representation of k equals the binary representation of j written backwards. Note that B_n is its own inverse.

In the multidimensional case, the Cooley–Tukey factorization can be written

$$(B_{n_d} \cdot F_{n_d}) \otimes \cdots \otimes (B_{n_1} \cdot F_{n_1})$$
$$= D_k(I_{n/2} \otimes F_2)D_{k-1}(I_{n/4} \otimes F_2 \otimes I_2) \cdots D_2(I_2 \otimes F_2 \otimes I_{n/4})D_1(F_2 \otimes I_{n/2}),$$

where $n = n_d \cdot n_{d-1} \cdots n_1$, and B_{n_i} are bit reversal matrices. The only differences between this equation and equation (1) are the values of the twiddle factors. (For details see [11].) Thus from a roundoff error analysis point of view, we can concentrate on the one dimensional case, and the results also apply to the multi-dimensional case. (If one wants to do a *very* careful analysis, one can use the fact that in the multi-dimensional case there are more twiddle factors equal to 1, and the constants in the error term can thus be improved slightly. This is, however, not of practical interest.)

The roundoff error analysis of the FFT has been studied from several different points of view and for an extended bibliographical survey we refer to chapter 2 of [3] and to the more recent chapter 1 of [10]. We refer to [10] for the general framework on FFT.

The motivation of our study comes from the large discrepancy between the very small errors observed in numerical experiments and the much larger upper bounds of the errors that would be expected from simple arguments. Indeed $y = F_N x$ is obtained in $\log_2 N$ steps of the form $x^{(k)} \leftarrow Hx^{(k-1)}$. Because the $\|F_2\|_\infty$ is 2, at each of the step the local error grows like 2 times machine precision times the $\|\bar{x}^{(k-1)}\|_\infty$, where $\bar{x}^{(k-1)}$ is the computed value at step k.

Simple computations [3, theorem 2.5.2, p. 59] show that the local factor 2 implies that the final bound on the error grows with a factor $n \log_2 n$. If we want to have a tighter bound of the error we must delay the use of the norms and compute first a componentwise expression of the global error.

In section 2 we present the results of the componentwise error analysis applied to the Cooley–Tukey algorithm for the FFT and its inverse, assuming that the computations are done using a rounding arithmetic. From the results proved in [13,3], we assume that the twiddle factors are computed by direct call to the library functions so that they are the rounded values of the exact results. Other techniques such as Forward and Logarithmic Recursion, Repeated Multiplications, Subvector Scaling and Recursive Bisection give less accurate algorithms [13,3]. We give new upper bounds for the roundoff error in the output signal measured in various different norms. In particular, our upper bound on the max norm of the error is a substantial improvement of the one we know in the literature (see [5,3,10]). We want to underline here the importance of measuring the error in max-norm for general applications, including the ones in turbulence theory which we are studying in section 4.

In section 3 we describe statistical properties of the algorithmic error under some hypotheses on the statistical distribution of the rounding errors. The results of that section, describing the expected value of the maximum roundoff error relative to the infinity norm of the output signal for any input signal, produce a more sound and general statistical analysis compared with those in the papers by Weinstein, by Kaneko and Liu and by Oppenheim and Weinstein [16,8,12]. In [16] the error analysis is done assuming a white noise input signal. The results are given in terms of the mean squared output error rather than on the expectation value of the maximum error. In [12] the results of [16] are reviewed and some experimental investigations are carried out on very simple sinusoidal signals. Finally in [8] a bound is given on the total mean square error relative to the L_2-norm of the output signal. We conclude section 3 by verifying, with numerical experiments, the statistical error bounds that we have derived. The experiments have been run on several computers with different arithmetics, such as Alliant-FX80, CRAY2, Risc workstations.

In section 4 we analyze the consequences of the previous result on the simulations of developed turbulence using spectral methods.

Finally, in section 5 we apply our results to the roundoff error analysis of fast solvers for the Helmholtz equation.

As a short hand notation we introduce

$$H_j = D_j(I_{2^{j-1}} \otimes F_2 \otimes I_{2^{k-j}}) \tag{2}$$

and write equation (1) as:

$$B_{2^k} \cdot F_{2^k} = \prod_{j=1}^{k} H_j \equiv H_k \cdots H_1. \tag{3}$$

After a symmetric permutation, H_j can be written as a block diagonal matrix with 2 by 2 blocks.

It is useful to know that by truncating this product, one can obtain the transforms of order 2^l with $l \leq k$:

$$I_{2^{k-l}} \otimes (B_{2^l} \cdot F_{2^l}) = \prod_{j=k-l+1}^{k} H_j. \tag{4}$$

In the following, if u and v are vectors of entries u_i and v_i and Q and P are matrices of entries q_{ij} and p_{ij}, $|u|$ is the vector of entries $|u_i|$, $|Q|$ is the matrix of entries $|q_{ij}|$. $u \leq v$ means $u_i \leq v_i$ for all i, and $Q \leq P$ means $q_{ij} \leq p_{ij}$ for all i and j.

Finally, by inspecting the pattern of the matrices, it follows that

$$|H_j||H_k| = |H_j H_k|, \qquad j \neq k. \tag{5}$$

2. Error analysis of the FFT

Let $fl(\cdot)$ denote the result of a floating point computation. We assume that the arithmetic of the computer satisfies the following for real arithmetic:

$$fl(a \square b) = (a \square b)(1 + \delta); \qquad |\delta| < \epsilon \tag{6}$$

where ϵ is the machine precision and $\square$ is one of $+$, $-$, $*$, $/$. To a great extent modern computers have arithmetic that satisfies assumption (6) with the exception of CRAY computers. For complex arithmetic (6) leads to the following error bounds for the basic operations:

Lemma 2.1

Complex floating point arithmetic is bounded by:

$$fl(z * w) = (z * w)(1 + \mu); \qquad |\mu| < (1 + \sqrt{2}) \cdot \epsilon,$$
$$fl(z + w) = (z + w)(1 + \sigma); \qquad |\sigma| < \epsilon,$$

where ϵ is the machine precision.

For proof of the lemma, see [3].

This leads to the following bound for multiplication with the elementary matrices H_j in equation (2):

Lemma 2.2

$$fl(H_j x) = (I + \Delta_j) H_j x,$$

where Δ_j is a diagonal matrix satisfying:

$$|\Delta_j| < (3 + \sqrt{2}) \cdot \epsilon I + \mathcal{O}(\epsilon^2).$$

Proof
After a symmetric permutation, H_j can be written as a block diagonal matrix with blocks of the form:

$$\begin{pmatrix} \omega_1 & 0 \\ 0 & \omega_2 \end{pmatrix} \cdot \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}.$$

We assume that the twiddle factors ω_i are computed exactly and then rounded, i.e.

$$fl(\omega) = \omega(1+\rho), \qquad \text{where } |\rho| < \epsilon.$$

Using lemma 2.1 we obtain this bound. □

Lemma 2.2 yields the following result for the componentwise errors in the FFT:

Theorem 2.3
The error in the radix 2 Cooley–Tukey FFT of size $n = 2^k$ is bounded by:

$$F_n x - fl(F_n x) = GF_n x + \mathcal{O}(\epsilon^2),$$

where

$$G = B_n \cdot \left(\sum_{j=1}^{k} H_k H_{k-1} \cdots H_{j+1} \Delta_j H_{j+1}^{-1} \cdots H_{k-1}^{-1} H_k^{-1} \right) \cdot B_n$$

and B_n is the bit-reversal permutation.

Proof
From equation (3) and lemma 2.2 we get:

$$\begin{aligned} B_n \cdot fl(F_n x) &= (I + \Delta_k) H_k \cdots (I + \Delta_1) H_1 x \\ &= H_k \cdots H_1 x + \sum_{j=1}^{k} H_k H_{k-1} \cdots H_{j+1} \Delta_j H_j H_{j-1} \cdots H_1 x + \mathcal{O}(\epsilon^2) \\ &= B_n F_n x + \sum_{j=1}^{k} H_k \cdots H_{j+1} \Delta_j H_{j+1}^{-1} \cdots H_k^{-1} B_n F_n x + \mathcal{O}(\epsilon^2). \end{aligned}$$

Since B_n is its own inverse, we obtain the result. □

Interestingly, the error matrix G has a rather regular structure, being essentially the sum of block diagonal matrices with Toeplitz circulant diagonal blocks:

Theorem 2.4
Let

$$C_j = H_k H_{k-1} \cdots H_{j+1} \Delta_j H_{j+1}^{-1} \cdots H_{k-1}^{-1} H_k^{-1},$$

let $m = 2^{k-j}$, and let B_m be the bit reversal permutation of order m. Then

$$(I_{n/m} \otimes B_m) C_j (I_{n/m} \otimes B_m)$$

is a block diagonal matrix with m by m Toeplitz circulant matrices on the diagonal.

Before we prove this theorem, we recall the following well known relationship between Toeplitz circulant matrices and discrete Fourier transforms:

Lemma 2.5
Let

$$\Delta = \operatorname{diag}(\delta_0, \delta_1, \ldots, \delta_{n-1}).$$

Then

$$F_n \Delta F_n^{-1} = \operatorname{circ}(\check{\delta}_0, \check{\delta}_1, \ldots, \check{\delta}_{n-1}),$$

where the numbers $\check{\delta}_j$ are the inverse Fourier transform of the numbers δ_k, i.e.

$$\check{\delta}_j = \frac{1}{n} \sum_{k=0}^{n-1} \delta_k e^{-2\pi i j k/n}.$$

Proof of theorem 2.4
If we block the diagonal matrix Δ_j in blocks Δ_j^l of size m, we find from equation (4) that the blocks of

$$(I_{n/m} \otimes B_m) C_j (I_{n/m} \otimes B_m)$$

are given as $F_m \Delta_j^l F_m^{-1}$ which are Toeplitz by the above lemma. □

We now have the necessary results for finding the infinity norm relative error of the computed $\hat{x}$:

Theorem 2.6
The infinity norm of the error matrix G is bounded by

$$\|G\|_\infty < 10.7\sqrt{n}\epsilon + \mathcal{O}(\epsilon^2).$$

Thus the relative error of the computed $\hat{x}$ is bounded by

$$\rho_\infty = \|fl(F_n x) - \hat{x}\|_\infty / \|\hat{x}\|_\infty < 10.7\sqrt{n}\epsilon + \mathcal{O}(\epsilon^2).$$

Proof
From the Plancherel identity $\|x\|_2 = (1/\sqrt{n})\|\hat{x}\|_2$ we get:

$$\|C_j\|_\infty = \sum_{l=0}^{m-1} |\check{\delta}_l| \le \sqrt{m} \left(\sum_{l=0}^{m-1} |\check{\delta}_l|^2 \right)^{1/2} = \left(\sum_{l=0}^{m-1} |\delta_l|^2 \right)^{1/2}$$
$$\le \sqrt{m}(3 + \sqrt{2})\epsilon + \mathcal{O}(\epsilon^2),$$

where the numbers $\check{\delta}_j$ are the inverse Fourier transform of the numbers δ_k. Thus

$$\|G\|_\infty \le \sum_{j=1}^{k} \|C_j\|_\infty \le (3+\sqrt{2})\epsilon \cdot (1+\sqrt{2}+\sqrt{4}+\cdots+\sqrt{n/2}) + \mathcal{O}(\epsilon^2)$$

$$= \frac{(3+\sqrt{2})(\sqrt{n}-1)}{\sqrt{2}-1}\epsilon + \mathcal{O}(\epsilon^2) < 10.7\sqrt{n}\epsilon + \mathcal{O}(\epsilon^2). \qquad \square$$

Theorem 2.7
The 2-norm of the error matrix G is bounded by

$$\|G\|_2 < 10.7\epsilon \log_2 n + \mathcal{O}(\epsilon^2).$$

Thus the relative error of the computed $\hat{x}$ is bounded by

$$\|fl(F_n x) - \hat{x}\|_2 / \|\hat{x}\|_2 < 10.7\epsilon \log_2 n + \mathcal{O}(\epsilon^2).$$

Proof
The proof follows from the orthogonal properties of F_n and theorem 2.3. $\square$

Finally, the following theorem gives the componentwise error bounds.

Theorem 2.8

$$|(fl(F_n x) - \hat{x})_j| < 10.7\epsilon \log_2 n \|F_n x\|_1, \qquad j = 1, \ldots, n, \tag{7}$$

$$|(fl(F_n x) - \hat{x})_j| < 10.7\epsilon \log_2 n \|x\|_1, \qquad j = 1, \ldots, n, \tag{8}$$

$$|(fl(F_n x) - \hat{x})_j| < 10.7\sqrt{n}\epsilon \|F_n x\|_\infty, \qquad j = 1, \ldots, n. \tag{9}$$

Proof
Formula (7) follows from theorem 2.7 and from $\|F_n x\|_2 \le \|F_n x\|_1$.

Formula (9) is the componentwise version of theorem 2.6.

Formula (8) follows from (5) and the relation proved in theorem 2.3

$$B_n \cdot (fl(F_n x) - \hat{x}) = \sum_{j=1}^{k} H_k H_{k-1} \cdots H_{j+1} \Delta_j H_j H_{j-1} \cdots H_1 x + \mathcal{O}(\epsilon^2). \qquad \square$$

For the inverse FFT, we have results similar to those of theorems 2.3, 2.4, 2.6, 2.7 and 2.8. In particular, we have

$$\frac{1}{n} F_n^* x - fl\left(\frac{1}{n} F_n^* x\right) = \tilde{G}\left(\frac{1}{n} F_n^* x\right) + \mathcal{O}(\epsilon^2)$$

with

$$\|\tilde{G}\|_\infty < 10.7\sqrt{n}\epsilon + \mathcal{O}(\epsilon^2).$$

Moreover, the matrix $\tilde{G}$ satisfies properties similar to those satisfied by G.

3. Statistical error analysis

The above analysis is based on a "worst case" analysis where all rounding errors work in the same direction. In this section we will show that if we assume a statistical distribution of rounding errors, the relative error in the answers will grow as $\log_2(n)$ rather than the $\sqrt{n}$ bound derived above.

Theorem 2.3 gives an exact expression for the roundoff error, expressed in terms of unknown matrices Δ_j of the form

$$\Delta = \operatorname{diag}(\delta_0, \delta_1, \ldots, \delta_{n-1}).$$

In this section we make the following assumption about the statistical distribution of the complex numbers δ_i:

Assumption 1
The numbers δ_i are independent complex stochastic variables where the real and imaginary parts are binormally distributed with mean 0 and variance σ^2, i.e. the probability density of the real and imaginary parts is given by:

$$f(\Re(\delta_i), \Im(\delta_i)) = BiNo(0, \sigma^2) = \frac{1}{2\pi\sigma^2} e^{-|\delta_i|^2/2\sigma^2}.$$

This distribution is not realistic for the arithmetic model (6), since, in the statistical distribution, the rounding errors δ_i range in $[-\infty, \infty]$. A more realistic probability distribution would be a uniform distribution in a disk, i.e. the probability density

$$g(\Re(\delta_i), \Im(\delta_i)) = \begin{cases} \dfrac{1}{\pi u^2} & \text{for } |\delta_i|^2 \le u^2, \\ 0 & \text{otherwise,} \end{cases}$$

where u is the machine precision.

Due to the Central Limit Theorem, we can, however, expect that results derived from the distributions f and g will behave similarly for large n.

The right choice for the variance σ^2 in our model may be debated. A reasonable choice is

$$\sigma^2 = 8u^2/9\pi,$$

which will give the same value for $E[|\delta_i|]$ for both distributions f and g. We are however more interested in how the relative error scales for large n than in the exact value of the constant of the error expression, so for simplicity we choose the larger value

$$\sigma^2 = u^2.$$

The following two lemmas are based on elementary properties of the normal distribution:

Lemma 3.1
Let δ be a complex stochastic variable with distribution

$$\delta \sim BiNo(0, \sigma^2)$$

and let z be a complex number. Then $z\delta$ has a distribution

$$z\delta \sim BiNo(0, \sigma^2 |z|^2).$$

Lemma 3.2
Let δ_1 and δ_2 be independent stochastic variables with distributions

$$\delta_{1,2} \sim BiNo(0, \sigma_{1,2}^2).$$

Then $\delta = \delta_1 + \delta_2$ has the probability distribution

$$\delta \sim BiNo(0, \sigma_1^2 + \sigma_2^2).$$

From these and from the properties of orthogonality of F_n we derive [1]:

Lemma 3.3
Let $\delta = \{\delta_l\}_{l=0}^{n-1}$ be a stochastic vector where the components are independent complex stochastic variables with distributions

$$\delta_l \sim BiNo(0, \sigma^2).$$

Then the inverse FFT $\check{\delta}$ is a stochastic vector where the components are independent with distributions

$$\check{\delta}_l \sim BiNo(0, \sigma^2/n).$$

From theorem 2.3 we see that the relative error

$$\rho_\infty = \frac{\|\hat{x} - fl(F_n x)\|_\infty}{\|\hat{x}\|_\infty}$$

is largest when all the components of $\hat{x}$ are of the same size, and in the following we therefore assume that

$$|\hat{x}_i| = 1 \qquad \text{for all } i. \tag{10}$$

Let C_j be the elementary matrices of theorem 2.4. From theorem 2.4 and lemmas 3.1–3.3 we find that the components of $y = C_j x$ are distributed:

$$y_i \sim BiNo(0, \sigma^2).$$

Since there are $\log_2(n)$ such terms in the error expression we get:

Lemma 3.4
Let $y = G\hat{x}$, where $|\hat{x}_i| = 1$ and G is given in theorem 2.3. Then the components of y are independently distributed

$$y_i \sim BiNo(0, \sigma^2 \log_2(n)).$$

Let $z = \max_{i=0}^{n-1} |y_i| = \|y\|_\infty$. It remains to compute $E[z]$. It is well known that if x_1 and x_2 are real normally distributed stochastic variables with mean 0 and variance 1, then $Y = \sqrt{x_1^2 + x_2^2}$ is Rayleigh distributed, i.e.

$$Y \sim ye^{-y^2/2}; \qquad y > 0.$$

Thus

$$P(Y \leq y) = 1 - e^{-y^2/2}.$$

Let $Z = \max_{i=0}^{n-1} Y_i$. Then

$$P(Z \leq z) = (1 - e^{-z^2/2})^n.$$

We will show that this function approaches a step function as $n \to \infty$:

Lemma 3.5
For large n we have

$$F_n(y) = (1 - e^{-y^2/2})^n \to \Theta_{\sqrt{2\ln(n)}}(y),$$

where $\Theta_z(y)$ is the step function

$$\Theta_z(y) = \begin{cases} 0 & \text{if } y < z, \\ 1 & \text{otherwise.} \end{cases}$$

The width of the step in F_n decreases as $O(\ln(n)^{-1/2})$.

Proof
Fix the function at a value $0 < c < 1$: $(1 - e^{-y^2/2})^n = c$. Solving for y, we get:

$$y(c) = \sqrt{-2\ln(1 - c^{1/n})}.$$

Assuming n large, we find by Taylor expansion:

$$y(c) = \sqrt{-2\ln(1 - c^{-1/n})} \approx \sqrt{2\ln(n) - \ln(-\ln(c))}$$
$$\approx \sqrt{2\ln(n)} - \ln(-\ln(c))/(2\sqrt{2\ln(n)}) = \sqrt{2\ln(n)} + \mathcal{O}(\ln(n)^{-1/2}).$$

Fixing two function values c_1 and c_2 we see that

$$y(c_1) - y(c_2) = \mathcal{O}(\ln(n)^{-1/2}) \to 0 \qquad \text{when } n \to \infty.$$

Thus $F_n(y)$ approaches a step function $\Theta_z(y)$ with a step at $z = \sqrt{2\ln(n)}$, and the width of the step in $F_n(y)$ decreases as $\mathcal{O}(\ln(n)^{-1/2})$. □

Thus the discussion in this section can be summarized in the following theorem, which is the stochastic analogue of theorem 2.6:

Theorem 3.6
Under assumption 1, where $\sigma = \epsilon$, the relative errors in the computed results satisfy

$$E[\rho_\infty] = E[\| fl(F_n x) - \hat{x}\|_\infty / \|\hat{x}\|_\infty]$$

$$\leq \epsilon \sqrt{\frac{2}{\ln(2)}} \ln(n) + \mathcal{O}(\epsilon),$$

where $E[z]$ denotes the expected value of the stochastic variable z. Furthermore:

$$\text{Var}[\rho_\infty] \to 0 \qquad \text{when } n \to \infty.$$

Proof
It is evident that if a stochastic variable X satisfies

$$P(X \leq x) = \Theta_{x_0}(x),$$

then

$$E(X) = x_0$$

and

$$\text{Var}(X) = 0.$$

Now let y be as in lemma 3.4 and let $z = \|y\|_\infty$. From lemma 3.5 we find that

$$E\left[\frac{z}{\sigma\sqrt{\log_2(n)}}\right] = \sqrt{2n\ln(n)} + \mathcal{O}(\ln(n)^{-1/2}).$$

Thus

$$E[z] = \sigma \sqrt{\frac{2}{\ln(2)}} \ln(n) + \mathcal{O}(\sigma).$$

This is derived under the assumption in equation (10) about $\hat{x}$. For general $\hat{x}$, the relative error becomes smaller. □

The derivation is based on Taylor expansions, and is valid for "sufficiently" large n. To check the validity of the derivation we have compared the bound in theorem 3.6 with the correct value for the statistical expectation, computed by numerical integration.

The statistical error bounds were based on assumption 1. We have done several numerical experiments to check if this assumption is realistic and if the $\log_2(n)$ behaviour is also seen in practice. The numerical computations were done on Alliant FX80, IBM-RS6000 and CRAY2 starting from a complex signal having random uncorrelated real and imaginary parts distributed uniformly between -1 and 1; the direct transform of the signal was performed both in double (Alliant FX80 and IBM-RS6000) and quadruple precision (CRAY2 and IBM-RS6000).

The infinity norm error of the double precision FFT was then computed as:

$$\rho_\infty = \frac{\|F_n^d x - F_n^q x\|}{\|F_n^q x\|}.$$

(We label with F^d the double precision FFT and with F^q the quadruple precision FFT.)

In figure 1 we plot the ratio between ρ_∞ and the $\sqrt{n}$ bound, derived in the "worst case" analysis, vs $\log_2 n$. We clearly see that this bound overestimates the actual error. The fit with the $\log_2 n$ law, shown in figure 2 where we plot $\rho_\infty / \log_2 n$ vs $\log_2 n$, is clearly much better.

When all the computations were performed using the CRAY2 and the IBM-RS6000, we obtained results similar to those displayed in figures 1 and 2.

4. Application to turbulence

Turbulent flows contain a large range of length scales. The largest scale is determined by the geometric dimension of the fluid, the smallest scale is the one at which the molecular viscosity is effective. The intermediate scales, which are not directly affected by the energy injection and dissipation, display an energy spectrum exhibiting an universal scaling behaviour:

$$\mathcal{E}(k) = k^{-n}.$$

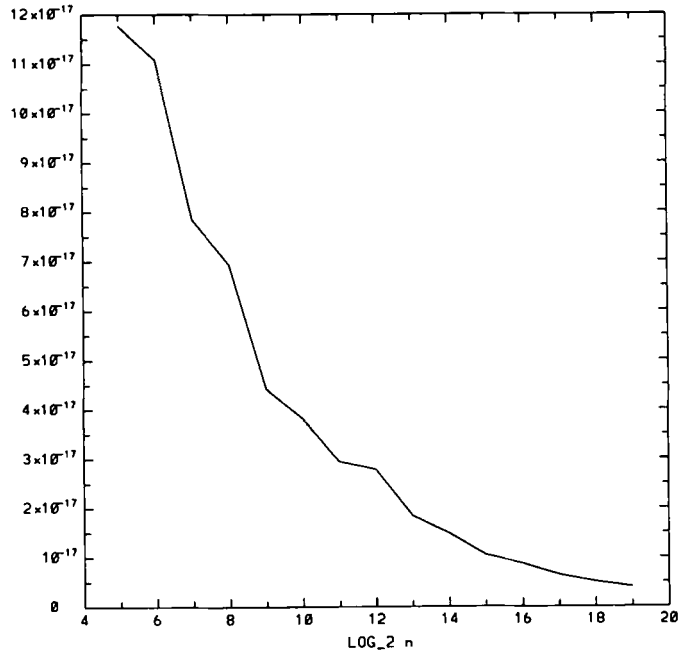

Figure 1. Computed infinity norm error of the double precision FFT of a random signal: $\rho_\infty / \sqrt{n}$ is plotted versus $\log_2 n$.

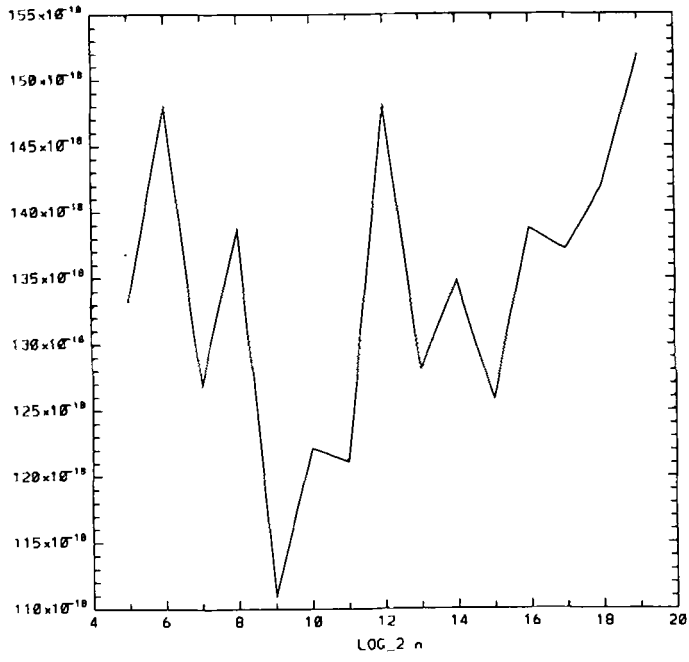

Figure 2. Computed infinity norm error of the double precision FFT of a random signal: $\rho_\infty/\log_2 n$ is plotted versus $\log_2 n$.

For three-dimensional turbulence this spectrum is close to the well known Kolmogorov spectrum with $n = 5/3$, while for bidimensional turbulence the exponent is close to $n = 3$ (see [9]).

In order to simulate effectively turbulent flows, a numerical method with high accuracy is needed. It is generally agreed that spectral methods are quite successful for this problem [6,2]. In these methods, the unknown function u is expanded in terms of an infinite sequence of orthogonal functions ϕ_i: $u = \sum_{k=-\infty}^{\infty} \hat{u}_k \phi_k$.

The numerical implementation of spectral methods requires a routine that computes the coefficients of the expansion, starting from the value of the function on a discretized grid, and the reciprocal routine which gives the value of the function on the mesh points from the spectral coefficients. It is required that these routines do not pollute the spectrum, i.e. that all the coefficients of the expansion are computed with a good accuracy. We want to study here the effects of the roundoff error on such a transformation.

We will restrict ourselves to the Fourier spectral method, which uses a Fourier series on a regular grid (e.g. $u = \sum_{k=-\infty}^{\infty} \hat{u}_k e^{ikx}$ in one dimension). The coefficients of the series, $\hat{u}_k$, are obtained by computing the d-dimensional FFT of the mesh points (d is the dimensionality of the space); conversely, the values at the mesh points are found by performing the inverse FFT of the coefficients of the Fourier series.

In order to be consistent with turbulence data we consider a signal with random

phases and with the following spectrum:

$$\mathcal{E}(k) = k^{-n} e^{-\beta k},$$

where n represents the scaling behaviour of the inertial region and β is the viscous cutoff. The actual choice of the value of n and β for the experiment is not very important: the relevant thing is that there is a strong difference in magnitude between the large and the small scales. We do the numerical experiment in the following way: we start with the signal in spectral space, we perform the inverse FFT of this signal in quadruple precision, and we go back to spectral space using a double precision FFT. Since the errors arising from the computations in extended precision are negligible, the difference between the original signal and the result of the computation is precisely the error of the double precision FFT.

From the error analysis of section 2 we expect that the absolute errors will be distributed among all the components; this is because the error matrix G of theorem 1 mixes up all the components.

In figure 3 we show the difference between the original signal and the result of the computation. We see that indeed the absolute errors are distributed throughout the spectrum. As a consequence, the relative errors on the high harmonics are quite strong: from figure 4 we see that all the components which are smaller by a factor of $\sim 10^{-16}$ than the largest components are completely polluted by the double precision FFT.

Current three-dimensional computations of turbulent flows do not display such

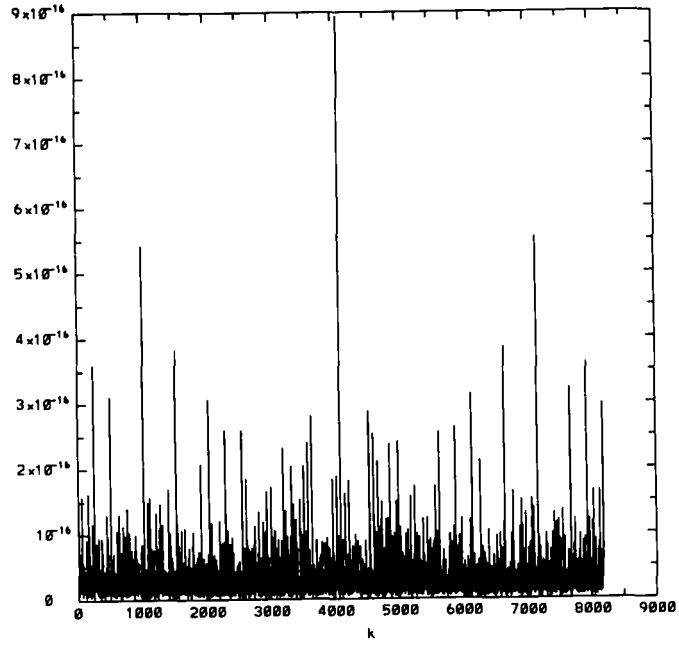

Figure 3. Roundoff error distribution in the computation of the double precision FFT of a signal having random phase and spectra $k^{-3}e^{-0.00232k}$. The absolute error is plotted versus k.

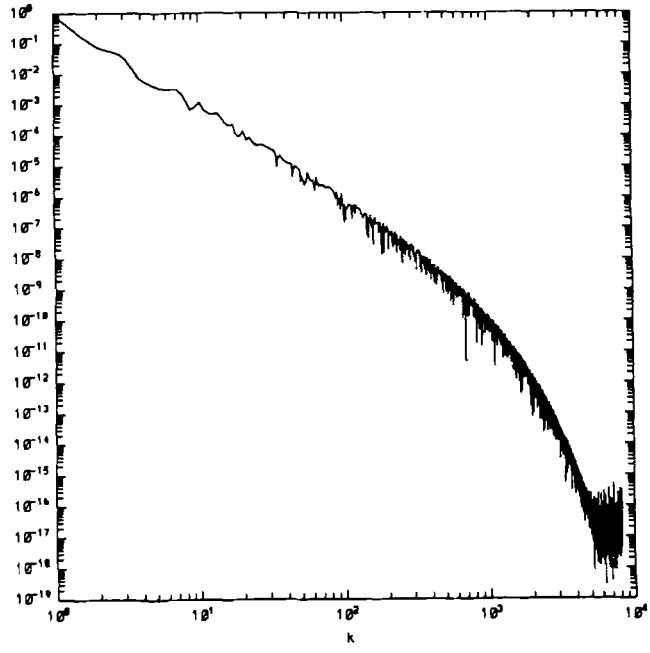

Figure 4. Double precision FFT of a signal having random phase and spectra $k^{-3}e^{-0.00232k}$, versus k. Note the complete pollution of the spectra for $k \geq 5 \times 10^3$.

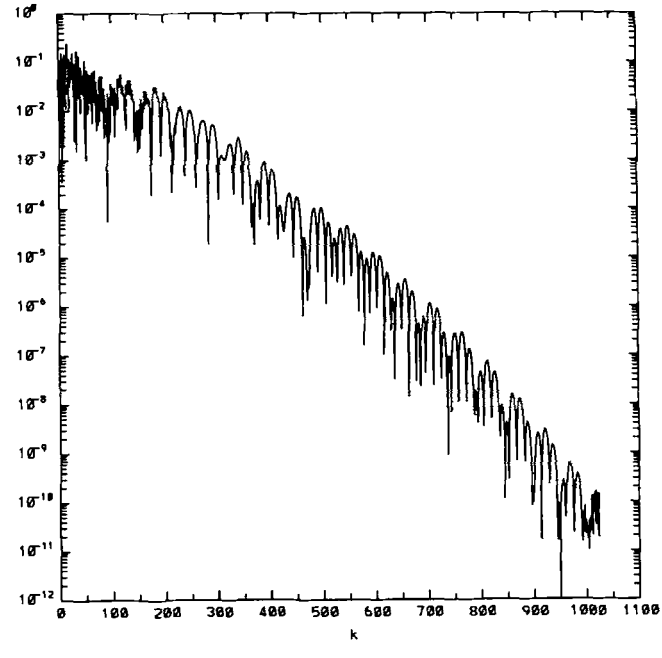

Figure 5. Plot of the real part of the spectral coefficients of the nonlinear term of the Navier–Stokes equation coming from two-dimensional turbulence simulations, as a function of the horizontal wave-number k.

a strong difference in magnitude between the large and the small scales: the largest available computation [15] displays spectra that vary by 6 orders of magnitude. The situation is however different for two-dimensional computations, where higher resolutions are reachable. We present in figure 5 a typical profile, in spectral space, of the nonlinear term of the Navier–Stokes equation (courtesy M. Manzini). This is the term which must be transformed to spectral space at each time step. The figure is obtained by fixing one of the two wavenumbers and plotting the spectral coefficients of the nonlinear term as a function of the second wavenumber. We see that the last components are smaller by a factor of 10^{-10} than the largest components; therefore they are computed with a relative precision of only $\sim 10^{10}\epsilon \sim 10^{-5}$ (ϵ is the machine precision).

5. Error analysis of fast Helmholtz solvers based on FFT

One of the major areas of application of the FFT is in the solution of the discretization of the Helmholtz problem on a square or a cube domain Ω

$$\nabla^2 u + \sigma u = f$$

with $u = g$ on $\partial\Omega$. For simplicity we will go into details only for the 2D case.

We will consider two kinds of discretization, central finite differences and the spectral method, and three algorithms for solving the corresponding discrete problem. In particular, we will give bounds for the roundoff error when we use the FFT for solving the discretized problems.

The finite difference discretization can be easily expressed in terms of the tridiagonal matrix of order N

$$T = \begin{bmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 2 \end{bmatrix},$$

where $N = 2^p$ is the number of points of discretization in one direction. Then the matrix A, discretizing the Helmholtz operator on a square, can be expressed as

$$A = I_N \otimes T + T \otimes I_N + \sigma I_N \otimes I_N.$$

The eigenvector matrix P of the matrix T is the imaginary part of a minor of F_N normalized by the factor $(2/(N+1))^{1/2}$ (see [14]). The corresponding eigenvalue matrix D is diagonal with entries

$$d_i = -2 + 2\cos(i\pi/(N+1)).$$

Moreover [14], the matrix $P \otimes P$ is the eigenvector matrix of A and the corresponding eigenvalue matrix is

$$I_N \otimes D + D \otimes I_N + \sigma I_N \otimes I_N.$$

The product of a vector y by P is the sine transform of y. In both cases, the products Py and $P \otimes Pz$, where y is an N-vector and z is an N^2 vector, can be performed by the FFT algorithm using the F_{2N} matrix.

Then, from the results of section 2, we have that

$$fl(Py) = Py + Gy$$

and

$$fl((P \otimes P)z) = (P \otimes P)z + \bar{G}z$$

with

$$\|G\|_\infty \leq \mu\epsilon\sqrt{N} + \mathcal{O}(\epsilon^2)$$

and

$$\|\bar{G}\|_\infty \leq \mu\epsilon N + \mathcal{O}(\epsilon^2),$$

where $\mu \leq 10.7$.

Two of the most common algorithms for solving the discrete problem

$$Ax = b$$

are based on the possibility of diagonalizing A and T by $P \otimes P$ and P and can be described as follows:

Algorithm 1

Step 1. $y = (P \otimes P)b$.
Step 2. $u = (I_N \otimes D + D \otimes I_N + \sigma I_N \otimes I_N)^{-1} y$.
Step 3. $x = (P \otimes P)^T u$.

Algorithm 2

Step 1. $y = (P \otimes I_N)b$.
Step 2. Solve $Vu = (I_N \otimes T + D \otimes I_N + \sigma I_N \otimes I_N)u = y$.
Step 3. $x = (P \otimes I_N)^T u$.

Since the matrix V is a positive definite tridiagonal matrix, the system $Vu = y$ is solved by the Cholesky algorithm and the computed solution $\bar{u}$ is the exact solution of the system [7]

$$(V + \delta V)\bar{u} = y, \qquad |\delta V| \leq 6\epsilon|V|. \tag{11}$$

If we assume that we have a periodic boundary condition for the Helmholtz equation, the problem can be discretized by truncating the Fourier series. For simplicity, we will assume $g = 0$ in the following. If we discretize the space variable in the Fourier nodes

$$x_j = 2j\pi/(N+1), \qquad y_k = 2k\pi/(N+1), \qquad j,k = 1,\ldots,N,$$

then the Fourier coefficients $\tilde{u}_{jk}$ and $\tilde{f}_{jk}$ of the functions u and f are obtained by

applying an inverse FFT to the values $u(x_j, y_k)$ and $f(x_j, y_k)$. In [2], it is proved that, using this property, the solution of a Helmholtz equation can be obtained by solving the system

$$Sx = b$$

with

$$S = (B_N \cdot F_N \otimes B_N \cdot F_N)^{-1} D_F (B_N \cdot F_N \otimes B_N \cdot F_N),$$

where we denote by D_F the diagonal matrix of entries

$$(D_F)_{j,k} = -(j^2 + k^2) + \sigma.$$

Algorithm 3

Step 1. $y = (B_N \cdot F_N \otimes B_N \cdot F_N)^{-1} b.$
Step 2. $u = (D_F)^{-1} y.$
Step 3. $x = (B_N \cdot F_N \otimes B_N \cdot F_N) u.$

We first establish a technical lemma that will be used in the proof of the main results stated in theorem 5.2.

From the algebraic point of view, the three algorithms that we have described, can be seen as three specific cases of the following sequence of matrix operations

$$x = W^{-1} U^{-1} W b, \tag{12}$$

where W is a matrix related to F_N and U is either a tridiagonal or a diagonal positive definitive matrix. Taking into account formula (11) we can also assume that

$$\delta U \leq c\epsilon |U|, \tag{13}$$

with c independent of N. Moreover, we can assume without loss of generality that $W^{-1} = W^T$.

The following result gives a general expression of the error δx affecting the computed value $\bar{x}$ of x in (12).

Lemma 5.1
Let

$$\bar{y} = fl(Wb) = Wb + \Gamma Wb, \qquad \bar{u} = fl(W^T v) = W^T v + \tilde{\Gamma} W^T v,$$

and (13). Then

$$\delta x = [\tilde{\Gamma} - W^T U^{-1} \delta U W + W^T U^{-1} \Gamma U W] x + \mathcal{O}(\epsilon^2).$$

Proof
From (11) we have

$$\bar{z} = fl(U^{-1}\bar{y}) = (U + \delta U)^{-1} \bar{y}.$$

The result is then obtained by expanding the right hand side of

$$\bar{x} = fl(W^T fl(U^{-1} fl(Wb))),$$

by linearizing the errors and by the fact that

$$(U + \delta U)^{-1} = (I - U^{-1}\delta U)U^{-1} + \mathcal{O}(\epsilon^2). \qquad \square$$

We are now ready to state the general expression of the error δx affecting the computed value $\bar{x}$ of x in (12).

Theorem 5.2
Let $\bar{x}^{(1)}$, $\bar{x}^{(2)}$ and $\bar{x}^{(3)}$ be the computed values of the solution x obtained by algorithm 1, algorithm 2 and algorithm 3, respectively. Then

$$\frac{\|\bar{x}^{(1)} - x\|_\infty}{\|x\|_\infty} \leq c_1 \epsilon N[2 + \kappa_\infty(A)] + \mathcal{O}(\epsilon^2), \tag{14}$$

$$\frac{\|\bar{x}^{(2)} - x\|_\infty}{\|x\|_\infty} \leq c_2 \epsilon \sqrt{N}[1 + \kappa_\infty(A) + (8 + |\sigma|)\|A^{-1}\|_\infty] + \mathcal{O}(\epsilon^2), \tag{15}$$

$$\frac{\|\bar{x}^{(3)} - x\|_\infty}{\|x\|_\infty} \leq c_3 \epsilon N[2 + \kappa_\infty(S)] + \mathcal{O}(\epsilon^2), \tag{16}$$

with c_1, c_2 and c_3 constant and

$$\kappa_\infty(A) = \|A^{-1}\|_\infty \|A\|_\infty, \qquad \kappa_\infty(S) = \|S^{-1}\|_\infty \|S\|_\infty.$$

Moreover, for the Poisson equation ($\sigma = 0$) we have

$$\frac{\|\bar{x}^{(1)} - x\|_\infty}{\|x\|_\infty} \leq \mathcal{O}(\epsilon N^3), \tag{17}$$

$$\frac{\|\bar{x}^{(2)} - x\|_\infty}{\|x\|_\infty} \leq \mathcal{O}(\epsilon \sqrt{N} N^2), \tag{18}$$

$$\frac{\|\bar{x}^{(3)} - x\|_\infty}{\|x\|_\infty} \leq \mathcal{O}(\epsilon N^3 \log_2^2 N). \tag{19}$$

Remark
Formulas (17)–(19) are, to our knowledge, the first bounds given on the roundoff error for the Poisson problem. Our bounds take into account the novelty of the results of section 1. When using finite differences it is known that the condition number $\kappa_\infty(A)$ does not depend on the dimensionality of the problem. This property does not hold when the FFT algorithm is used, as is shown in appendix A. As a consequence, the roundoff error (19) has a different expression for 1D and 3D problems (the behaviour would be, respectively, $\epsilon N^{5/2} \log_2 N$ and $\epsilon N^{9/2} \log_2 N$ in 1D and 3D).

We note also that all these bounds are based on the upper bound worst case analysis of section 2. Using the statistical analysis of section 3 the bounds

(17)–(18) drop to $\mathcal{O}(\epsilon N^3 \log_2^2 N)$, and the bound (19) becomes: $\mathcal{O}(\epsilon N^2 \log_2^2 N)$ in 1D, $\mathcal{O}(\epsilon N^2 \log_2^3 N)$ in 2D and $\mathcal{O}(\epsilon N^3 \log_2^2 N)$ in 3D.

For the Helmholtz problem (formulas (14)–(16)), the condition numbers $\kappa_\infty(A)$ and $\kappa_\infty(S)$ depend heavily on the value of σ. We can only say that for σ positive the condition number goes closer to 1.

Proof of theorem 5.2

For algorithm 1 and algorithm 2 we can write the expression of δU in (13) as

$$\delta U = \Delta D$$

with D the corresponding diagonal matrix of each algorithm and $|\Delta| \leq \epsilon$. Then, from lemma 5.1 we have

$$\delta x = [\tilde{\Gamma} - W^T \Delta W + A^{-1} W^T \Gamma W A] x + \mathcal{O}(\epsilon^2).$$

Using the infinity norm we obtain

$$\|\delta x\|_\infty \leq [\|\tilde{\Gamma}\|_\infty + \|W^T \Delta W\|_\infty + \kappa_\infty(A)\|W^T \Gamma W\|_\infty] \|x\|_\infty + \mathcal{O}(\epsilon^2),$$

where $\kappa_\infty(A) = \|A\|_\infty \|A^{-1}\|_\infty$.

Formulae (14) and (16) follow directly from the fact that

$$\|W^T \Gamma W\|_\infty \leq 10.7 N\epsilon + \mathcal{O}(\epsilon^2)$$

(the proof of this is equivalent to the one of theorem 2.6), and from the property

$$\|\tilde{\Gamma}\|_\infty \leq 10.7 N\epsilon + \mathcal{O}(\epsilon^2).$$

For proving (15) it is necessary to remember that the structure of δU is block diagonal with diagonal blocks of size N which are tridiagonal.

From lemma 5.1 we have

$$\|\delta x\|_\infty \leq [\|\tilde{\Gamma}\|_\infty + \kappa_\infty(A)\|W^T \Gamma W\|_\infty + \|W^T U^{-1} \delta U W\|_\infty] + \mathcal{O}(\epsilon^2).$$

The first two terms, since $\tilde{\Gamma}$ and Γ are block diagonal with diagonal blocks of size N, can be bounded as follows

$$\|\tilde{\Gamma}\|_\infty \leq 10.7\sqrt{N}\epsilon + \mathcal{O}(\epsilon^2)$$

and

$$\|\Gamma\|_\infty \leq 10.7\sqrt{N}\epsilon + \mathcal{O}(\epsilon^2).$$

For the third term we have

$$\|W^T U^{-1} \delta U W\|_\infty \leq \|A^{-1}\|_\infty \|(P \otimes I_N) \delta U (P \otimes I_N)^T\|_\infty.$$

Because there exists a permutation matrix Q such that $Q(P \otimes I_N)Q^T = I_N \otimes P$ and $Q \delta U Q^T = \widetilde{\delta U}$ and $\widetilde{\delta U}$ having a block tridiagonal structure with blocks diagonal, we have

$$\|(P \otimes I_N) \delta U (P \otimes I_N)^T\|_\infty = \|(I_N \otimes P) \widetilde{\delta U} (I_N \otimes P)^T\|_\infty.$$

The value of $\|(I_N \otimes P)\widetilde{\delta U}(I_N \otimes P)^T\|_\infty$ is attained on one block of N lines of the matrix. From this we derive

$$\|(I_N \otimes P)\widetilde{\delta U}(I_N \otimes P)^T\|_\infty \leq \sqrt{N}\|(I_N \otimes P)\widetilde{\delta U}(I_N \otimes P)^T\|_2 \leq \sqrt{N}\|\delta U\|_2$$

and then

$$\|\delta U\|_2 \leq \||\delta U|\|_2 \leq 6\epsilon\||U|\|_2 + \mathcal{O}(\epsilon^2) \leq 6\epsilon(8 + |\sigma|) + \mathcal{O}(\epsilon^2).$$

Finally, (17) and (18) follow from [14]

$$\kappa_\infty(A) \leq cN^2$$

and (19) follows from the following result established in appendix A:

$$\kappa_\infty(S) \leq cN^2 \log_2^2 N. \qquad \square$$

Appendix A

In this appendix we show that the condition number in infinity norm of the matrix S of theorem 5.2 approaches, for $\sigma = 0$, the value $2N^2(\log_2^2 N)/\pi$ in the limit of large N. Since both S and S^{-1} are Toeplitz circulant matrices, their infinity norms are equal to the L_1 norm of their rows. In particular, for S^{-1} we have

$$\|S^{-1}\|_\infty = \frac{1}{N^2} \sum_{p=0,q=0}^{N-1} \left| \sum_{j=0,k=0}^{N-1} \frac{1}{(j+1)^2 + (k+1)^2} e^{-2\pi i(pj+qk)/N} \right|.$$

It is easy to prove that

$$\|S^{-1}\|_\infty \sim \log_2 N.$$

The matrix $S_1 = F_N \Delta F_N^{-1}$, with Δ diagonal matrix of entries $\delta_k = k^2$, is the one-dimensional equivalent of S and

$$\|S\|_\infty \leq 2\|S_1\|_\infty.$$

Since S_1 is a Toeplitz circulant matrix, its infinity norm is equal to the L_1 norm of its rows. We thus have:

$$\|S_1\|_\infty = \frac{1}{N} \sum_{j=0}^{N-1} \left| \sum_{k=0}^{N-1} (k+1)^2 e^{-2\pi ijk/N} \right|.$$

From the well known property

$$\sum_{k=0}^{n} \sin kx = \frac{\sin((n+1)x/2)\sin(nx/2)}{\sin(x/2)},$$

$$\sum_{k=0}^{n} \cos kx = \frac{\sin((n+1)x/2)\cos(nx/2)}{\sin(x/2)},$$

the inner sum can be evaluated exactly:

$$\|S_1\|_\infty = \frac{N(2N+1)}{6} + \frac{1}{2}\sum_{j=1}^{N-1} \frac{\sqrt{1+(N^2-1)\sin^2 x_j}}{\sin^2 x_j} \qquad \left(x_j \equiv \frac{\pi j}{N}\right).$$

We rewrite this expression in the following form:

$$\|S_1\|_\infty = \frac{N(2N+1)}{6} + \frac{\sqrt{(N^2-1)}}{2}\sum_{j=1}^{N-1} \frac{1}{\sin x_j}\sqrt{1+\frac{1}{(N^2-1)\sin^2 x_j}}$$

and we expand the square root in a Taylor series around 1:

$$\sqrt{1+\frac{1}{(N^2-1)\sin^2 x_j}} = \sum_{i=0}^{\infty} c_i[(N^2-1)]^{-i}\sin^{-2i} x_j$$

(c_i are the coefficients of the Taylor expansion of $\sqrt{1+x}$).

We then have:

$$\|S_1\|_\infty = \frac{N(2N+1)}{6} + \frac{\sqrt{(N^2-1)}}{2}\sum_{j=1}^{N-1}\sum_{i=0}^{\infty} c_i[(N^2-1)]^{-i}\sin^{-2i-1} x_j.$$

We interchange the inner and the outer sum:

$$\|S_1\|_\infty = \frac{N(2N+1)}{6} + \frac{\sqrt{(N^2-1)}}{2}\sum_{i=0}^{\infty} c_i[(N^2-1)]^{-i}\sum_{j=1}^{N-1}\sin^{-2i-1} x_j. \tag{20}$$

We will show that the inner sum is dominated by the contribution $i = 0$ in the limit of large N, and that this contribution is of order $2N(\log_2 N)/\pi + \mathcal{O}(N)$.

We begin by proving that the contribution for $i \neq 0$ is of order $\mathcal{O}(N)$.

Lemma A.1

$$\sin^{-1} x \leq \frac{\pi}{2x}, \qquad \forall x \in \left]0, \frac{\pi}{2}\right[.$$

From this we derive:

$$\sum_{i=1}^{\infty} c_i[(N^2-1)]^{-i}\sum_{j=1}^{N-1}\sin^{-2i-1}\frac{\pi j}{N} \sim 2\sum_{i=1}^{\infty} c_i[(N^2-1)]^{-i}\sum_{j=1}^{(N-1)/2}\sin^{-2i-1}\frac{\pi j}{N}$$

$$\leq 2\sum_{i=1}^{\infty}\left(\frac{N}{2}\right)^{2i+1} c_i[(N^2-1)]^{-i}\sum_{j=1}^{(N-1)/2}\frac{1}{j^{2i+1}}$$

$$\leq 2C\sum_{i=1}^{\infty}\left(\frac{N}{2}\right)^{2i+1} c_i[(N^2-1)]^{-i}$$

$$\sim 2CN\sum_{i=1}^{\infty}\frac{c_i}{2^{i+1}} = \mathcal{O}(N)$$

(C is a constant such that $\sum_{j=1}^{\infty} 1/j^{2i+1} < C \qquad \forall i \geq 1$).

We show now that the contribution for $i = 0$ is of order $2N(\log_2 N)/\pi$.

Lemma A.2

$$\frac{1}{x} \leq \frac{1}{\sin x} \leq \frac{1}{x} + x, \qquad \forall x \in]0, \frac{\pi}{2}].$$

From this we derive:

$$\sum_{j=1}^{(N-1)/2} \frac{N}{\pi j} \leq \sum_{j=1}^{(N-1)/2} \frac{1}{\sin(\pi j/N)} \leq \sum_{j=1}^{(N-1)/2} \frac{N}{\pi j} + \frac{\pi j}{N},$$

$$\frac{N \log_2(N/2)}{\pi} \leq \sum_{j=1}^{(N-1)/2} \frac{1}{\sin(\pi j/N)} \to \frac{N \log_2(N/2)}{\pi} + \frac{\pi}{2}(N+1).$$

In the limit $N \to \infty$ we have:

$$\sum_{j=1}^{N-1} \frac{1}{\sin(\pi j/N)} \sim 2 \sum_{j=1}^{(N-1)/2} \frac{1}{\sin(\pi j/N)} \to \frac{2N \log_2 N}{\pi} + \mathcal{O}(N). \tag{21}$$

From equations (20) and (21) the infinity norm of the matrix S_1 in the limit of large N can be evaluated as:

$$\| S_1 \|_\infty \sim \frac{\sqrt{(N^2-1)}}{2} \frac{2N \log_2 N}{\pi} \sim \frac{N^2 \log_2 N}{\pi}.$$

It is interesting to note that S_1 is the spectral matrix in 1D corresponding to the discretization of the second derivative. Its condition number is

$$\kappa_\infty(S_1) \sim N^2 \log_2 N.$$

Analogously, denoting by S_3 the $N^3 \times N^3$ spectral matrix corresponding to the 3D discretization of the Laplace operator we have

$$\kappa_\infty(S_3) \sim N^3 \log_2 N.$$

The proof is similar to the one given for S.

It is relevant to note that the infinity norm condition number of the matrices coming from the finite difference approximation of the Laplace operator behaves as $\mathcal{O}(N^2)$, whereas that coming from the spectral approximation matrices S_1, S and S_3 increases from $\mathcal{O}(N^2 \log_2 N)$ to $\mathcal{O}(N^3 \log_2 N)$ with the dimensionality of the problem.

References

[1] D. Brillinger, *Time Series. Data Analysis and Theory* (Holt, Rinehart and Wilson, 1975).

[2] C. Canuto, M.Y. Hussaini, A. Quarteroni and T.A. Zang, *Spectral Methods in Fluid Dynamics* (Springer, New York, 1988).

[3] C.Y. Chu, The Fast Fourier Transform on hypercube parallel computers, Technical Report 87-882, Department of Computer Science, Cornell University (1987).
[4] J.W. Cooley and J.W. Tukey, An algorithm for the machine calculation of complex Fourier series, Math. Comp. 19 (1965) 297–301.
[5] W.M. Gentleman and G. Sande, Fast Fourier Transforms – For fun and profit, in: *1966 Fall Joint Computer Conf.*, AFIPS Conf. Proc. vol. 29 (Spartan, Washington, DC, 1966) pp. 563–578.
[6] D. Gottlieb and S.A. Orszag, *Numerical Analysis of Spectral Methods: Theory and Applications* (SIAM-CBMS, Philadelphia, 1977).
[7] N.J. Higham, Bounding the error in Gaussian elimination for tridiagonal systems, SIAM J. Matrix Anal. Appl. 11 (1990) 521–530.
[8] T. Kaneko and B. Liu, Accumulation of round-off error in Faster Fourier Transforms, J. ACM 17 (1970) 637–654.
[9] R.H. Kraichnan, Inertial ranges in two-dimensional turbulence, Phys. Fluids 10 (1967) 1417.
[10] C. Van Loan, *Computational Frameworks for the Fast Fourier Transforms* (SIAM, Philadelphia, 1992).
[11] H. Munthe-Kaas, Super parallel FFTs, Rep. no. 52 1991, Dept. of Informatics, University of Bergen, Norway (1991).
[12] A.V. Oppenheim and C.J. Weinstein, Effects of finite register length in digital filtering and Fast Fourier Transform, Proc. IEEE 60 (1972) 957–976.
[13] G.U. Ramos, Roundoff error analysis of the Fast Fourier Transform, Math. Comp. 25 (1971) pp. 757–768.
[14] P.N. Swarztrauber, *Fast Poisson Solvers*, vol. 24, MAA Studies in Numerical Analysis (1977).
[15] A. Vincent and M. Meneguzzi, The spatial structure and statistical properties of homogeneous turbulence, J. Fluid Mech. 225 (1991) 1.
[16] C.J. Weinstein, Roundoff noise in floating point Fast Fourier Transform computation, IEEE Trans. Audio and Electroacoustic AU-17 (1969) 209–215.