Aplikace matematiky

Karel Segeth

Roundoff errors in the fast computation of discrete convolutions

Aplikace matematiky, Vol. 26 (1981), No. 4, 241–262

Persistent URL: http://dml.cz/dmlcz/103916

ROUNDOFF ERRORS IN THE FAST COMPUTATION OF DISCRETE CONVOLUTIONS

KAREL SEGETH

(Received June 20, 1979)

1. INTRODUCTION

In many applications, the problem of efficient evaluation of the formula

$$b_l = \sum_{j=0}^{L-1} \psi_j \varphi_{l-j} ; \quad l = 0, \ldots, M-1 , \tag{1.1}$$

is often to be solved, where L is a positive integer and the sequence b is called the discrete convolution of the sequences ψ and φ. The integer M is supposed to be much greater than L.

Computing e.g. the right-hand part b of the system of M linear algebraic equations with the Gram matrix in the finite element method and applying an L-point (composite) Newton-Cotes quadrature formula, we finally come to the expression (1.1), where φ is now the vector of the values of the right-hand part of the solved differential equation at some M equidistant nodes, and the L components of the vector ψ depend on the values of the basis function and the coefficients of the quadrature formula.

A very important application of (1.1) is called the linear filtering. Now φ is a sequence of measured data and ψ is a filter. In many branches of science, linear filtering is the fundamental way of processing measured data.

The evaluation of (1.1) can be carried out directly. Such a direct procedure requires L multiplications and L additions for each l; the complete sequence b is thus calculated with the help of LM multiplications and the same number of additions.

Choosing an integer N, we will consider a more specific problem, namely the problem of evaluating the discrete convolution $t = (t_0, \ldots, t_{N-1})$, the components of which are given as

$$t_l = \sum_{j=0}^{N-1} g_j f_{l-j} ; \quad l = 0, \ldots, N-1 , \tag{1.2}$$

where $g = (g_0, \ldots, g_{N-1})$ and $f = (f_0, \ldots, f_{N-1})$ are known vectors and $f_{j+N} = f_j$, i.e., f represents a "periodic sequence".

The problem (1.2) possesses two important properties. First, the evaluation of (1.1) can be carried out as a repeated substitution into the formula (1.2) with a proper choice of N, $L \leqq N \leqq M$, and f and g. Second, we can employ a fast method of evaluation of the expression (1.2). Such a method is based on the discrete Fourier transform realized through the algorithm of the fast Fourier transform (Cooley and Tukey [2]).

The ways of the fast evaluation of (1.1) requiring an optimum choice of N and a repeated fast calculation of a convolution of the type (1.2) have been studied e.g. by Helms [6], who presented two methods suitable to this end, the "select-saving" method and the "overlap-adding" method.

The fast evaluation of the convolution (1.2) consists in the fast calculation of the discrete Fourier transforms of g and f, the determination of the "product" (in a certain sense) of these transforms, and the fast calculation of the inverse discrete Fourier transform of this product. The whole procedure thus requires $O(N \log N)$ arithmetic operations. To obtain (1.1), we have to repeat this procedure about $M/(N - L + 1)$ times. Apparently, we can achieve considerable efficiency of this fast convolution procedure only when L is large.

Davis and Rabinowitz [3] claim that (p. 201) "*experience has shown that rounding error does not generally accumulate in the performance of fast convolutions when floating-point arithmetic is used*"; they are apparently concerned only with the rounded arithmetic. Unfortunately, a series of numerical experiments carried out by the author in chopped arithmetic in order to compare the fast and the direct procedure for computing convolutions has shown that the fast convolution (being sometimes really less time-consuming) is generally much more influenced by roundoff errors. The paper by Thong and Liu [16] concerned with an improvement of the algorithm of the fast Fourier transform with the aim of obtaining a less roundoff error also indicates that the accumulation of the roundoff error in this algorithm may be dangerous in practical computation. •

In the present paper we try to analyse the roundoff error in the fast convolution as defined by (1.2). To obtain a comparison with the classical (direct) convolution, we consider also the problem of the direct computation of

$$u_l(L) = \sum_{j=0}^{L-1} g_j f_{l-j}, \quad l = 0, \ldots, N-1, \quad L \leqq N,$$

(1.2) being a particular case of this formula with $L = N$.

Basic concepts necessary for the treatment of the subject are introduced in Section 2. We define the discrete and the inverse discrete Fourier transform and present Parseval's identity. Further, we state the fundamental assumptions on the arithmetic considered and the local roundoff errors. We consider two possibilities of performing machine operations, the rounding and chopping.

In the whole paper we employ the stochastic approach to the propagation of roundoff error. In the conclusion of Section 2, the behavior of the roundoff error in the discrete and inverse discrete Fourier transforms is presented as established by Kaneko and Liu [10] for the particular algorithm of the fast Fourier transform called the decimation in frequency.

The main result of the paper is given in Section 3. The discrete convolution is defined here and its roundoff error is examined for the case of the direct as well as the fast computation.

The theoretical results derived in Section 3 are compared with numerical results in Section 4, which contains a simple numerical example. The comparison shows that the behavior of the roundoff error agrees with that predicted theoretically.

The problem of efficient evaluation of the discrete convolution is even more important in two dimensions. The way of the fast computation shown here can be applied also to the two-dimensional case. The analysis of the accumulation of roundoff error is then performed similarly.

2. THE FAST FOURIER TRANSFORM

Basic concepts enabling us to formulate our problem and to study it are introduced in this section. Further, we present several known statements concerned with the roundoff error in the fast Fourier transform.

The notation introduced in the following definitions is kept throughout the paper.

Definition 2.1. *Let m be a positive integer and*

$$N = 2^m . \tag{2.1}$$

Put

$$w_{jk} = \exp\left(2\pi ijk/N\right) .$$

If $x = (x_0, \ldots, x_{N-1})$ is a complex vector, i.e. $x \in \mathbb{C}_N$, the complex vector $a = (a_0, \ldots, a_{N-1}) \in \mathbb{C}_N$ defined as

$$a_k = \sum_{j=0}^{N-1} w_{jk} x_j ; \quad k = 0, \ldots, N-1 ,$$

is called the discrete Fourier transform of x and denoted by $a = \mathscr{F}(x)$. Similarly, the vector x given as

$$x_j = N^{-1} \sum_{k=0}^{N-1} \overline{w}_{jk} a_k ; \quad j = 0, \ldots, N-1 ,$$

is called the inverse discrete Fourier transform of a and denoted by $x = \mathscr{F}^{-1}(a)$.

The following trivial statement is very useful in calculations with the discrete Fourier transform.

Lemma 2.1. *We have*

$$\sum_{j=0}^{N-1} w_{j0} = N\,,$$

$$\sum_{j=0}^{N-1} w_{jk} = 0 \quad \textit{for} \quad k \neq 0\,.$$

Proof is straightforward as the second series is a finite geometrical series.

Remark 2.1. The discrete Fourier transform $\mathscr{F}$ introduced in Definition 2.1 is apparently a linear continuous mapping from the N-dimensional complex space $\mathbb{C}_N$ into $\mathbb{C}_N$. A straightforward computation (based on Lemma 2.1) shows that the mapping inverse to $\mathscr{F}$ is the inverse discrete Fourier transform $\mathscr{F}^{-1}$ also introduced in Definition 2.1.

The assumption that N is a power of 2 will be used later for the fast implementation of the discrete Fourier transform. The discrete Fourier transform itself can be defined for any positive integer N. •

A discrete analog of the well-known Parseval's identity is established in the following theorem.

Theorem 2.1. *Parseval's identity. Putting* $a = \mathscr{F}(x)$ *for* $x \in \mathbb{C}_N$, *we have*

$$\sum_{j=0}^{N-1} |x_j|^2 = N^{-1} \sum_{k=0}^{N-1} |a_k|^2\,. \tag{2.2}$$

Proof follows from Lemma 2.1 through a straightforward calculation. •

The investigation of the accumulation of roundoff errors presented in this and the next chapters is based on the well-known book by Wilkinson [17]. We have chosen the stochastic approach to the problem, the fundamentals of which are explained e.g. by Henrici [7], Hamming [4], and Sterbenz [13]. On the other hand, we are aware of the objections to this approach expressed e.g. by Hartree [5] and Huskey [8].

Let us formulate the basic assumptions on the arithmetic considered.

Assumption. *Let* x *and* y *be real (machine) numbers and let*

$$\mathrm{fl}(x + y) = (x + y)(1 + \alpha)\,,$$

$$\mathrm{fl}(xy) = xy(1 + \beta)\,,$$

where $\mathrm{fl}(x + y)$ *and* $\mathrm{fl}(xy)$ *are the results of machine operations and* α *and* β *are the local relative roundoff errors, for which*

$$L^+ \leqq \alpha \leqq U^+\,, \quad L^* \leqq \beta \leqq U^*$$

hold with some constants L^+, U^+, L^* *and* U^*.

We assume that α *and* β *are random variables (independent of* x *and* y*) with their means*

$$\mathsf{E}(\alpha) = \mu_\alpha\,, \quad \mathsf{E}(\beta) = \mu_\beta \tag{2.3}$$

and their variances

$$\mathsf{E}(\alpha^2) - \mu_\alpha^2 = \sigma_\alpha^2\,, \quad \mathsf{E}(\beta^2) - \mu_\beta^2 = \sigma_\beta^2\,. \tag{2.4}$$

We further suppose that the roundoff errors made in individual arithmetic operations of an algorithm are independent random variables.

Remark 2.2. The nature of the random variables α and β (in particular, their distribution) is studied in more detail e.g. by Kaneko and Liu [11], Sterbenz [13], and Hamming [4]. For our further purposes, only the above assumption with the relations (2.3) and (2.4) is essential.

Definition 2.2. *We consider the following possibilities of performing machine operations*:

1. *If* $\mu_\alpha = \mu_\beta = 0$ *we speak about the rounded arithmetic (or rounding).*
2. *If* $\mu_\alpha \neq 0$ *and* $\mu_\beta \neq 0$ *the arithmetic is called chopped (we speak about chopping).*

Remark 2.3. It is well-known that the stochastic treatment of the propagation of roundoff errors is advantageous in the case of rounding while for the chopped arithmetic (which is the case of the example in Section 4) it can hardly give anything better that the deterministic approach (see e.g. Sterbenz [13]). •

The usual machine arithmetic is real. The complex arithmetic is realized through real operations with real and imaginary parts of the operands. The following statement is concerned with the roundoff errors in such complex arithmetic and we will use it later.

Lemma 2.2. *Let* a *be a real number, let* $x = x_1 + x_2 i$ *and* $y = y_1 + y_2 i$ *be complex numbers. Then*

$$\begin{aligned} \mathrm{fl}\,(x + y) &= (x + y)(1 + \gamma)\,, \\ \mathrm{fl}\,(ay) &= ay(1 + \eta)\,, \\ \mathrm{fl}\,(xy) &= xy(1 + \delta)\,, \end{aligned} \tag{2.5}$$

where $\mathrm{fl}\,(x + y)$, $\mathrm{fl}\,(ay)$ *and* $\mathrm{fl}\,(xy)$ *are the computed values of* $x + y$, ay *and* xy, *respectively, and* γ, η *and* δ *are random variables with their means and variances given by*

$$\begin{aligned} \mathsf{E}(\gamma) &= \mu_\alpha\,, \\ \mathsf{E}(|\gamma|^2) - \mu_\alpha^2 &= \sigma_\alpha^2\,, \end{aligned}$$

$$\mathsf{E}(\eta) = \mu_\beta\,,$$

$$\mathsf{E}(|\eta|^2) - \mu_\beta^2 = \sigma_\beta^2\,,$$

(2.6) $$\mathsf{E}(\delta) = \mu_\alpha + \mu_\beta = \mu_\delta\,,$$

(2.7) $$\mathsf{E}(|\delta|^2) - \mu_\delta^2 = \sigma_\alpha^2 + \sigma_\beta^2 = \sigma_\delta^2\,.$$

The equality sign in (2.6) *and* (2.7) *is used in an approximate sense since the error terms of higher orders are neglected.*

Proof. The statement of the lemma is an easy consequence of Assumption. We will establish only the relations (2.6) and (2.7) for the complex multiplication (2.5). The other relations can be proved in a similar way.

Put

$$xy = z = z_1 + z_2 i\,, \quad \mathrm{fl}\,(xy) = z' = z_1' + z_2' i\,.$$

The equality (2.5) is apparently true for $z = 0$. Suppose thus $z \neq 0$. We have

$$z_1' = x_1 y_1(1 + \beta_1)(1 + \alpha_1) - x_2 y_2(1 + \beta_2)(1 + \alpha_1)\,,$$

$$z_2' = x_1 y_2(1 + \beta_3)(1 + \alpha_2) + x_2 y_1(1 + \beta_4)(1 + \alpha_2)\,,$$

i.e., after neglecting the error terms of higher orders,

$$z' - z = x_1 y_1(\alpha_1 + \beta_1) - x_2 y_2(\alpha_1 + \beta_2) +$$

$$+ (x_1 y_2(\alpha_2 + \beta_3) + x_2 y_1(\alpha_2 + \beta_4))\, i\,,$$

$$|z' - z|^2 = x_1^2 y_1^2(\alpha_1 + \beta_1)^2 - 2x_1 x_2 y_1 y_2(\alpha_1 + \beta_1)(\alpha_1 + \beta_2) +$$

$$+ x_2^2 y_2^2(\alpha_1 + \beta_2)^2 + x_1^2 y_2^2(\alpha_2 + \beta_3)^2 +$$

$$+ 2x_1 x_2 y_1 y_2(\alpha_2 + \beta_3)(\alpha_2 + \beta_4) + x_2^2 y_1^2(\alpha_2 + \beta_4)^2\,.$$

Now

$$\mathsf{E}(z' - z) = z_1(\mu_\alpha + \mu_\beta) + z_2(\mu_\alpha + \mu_\beta)\, i = z(\mu_\alpha + \mu_\beta)\,,$$

$$\mathsf{E}(|z' - z|^2) = (\mu_\alpha^2 + \sigma_\alpha^2 + 2\mu_\alpha\mu_\beta + \mu_\beta^2 + \sigma_\beta^2)(x_1^2 y_1^2 + x_2^2 y_2^2 +$$

$$+ x_1^2 y_2^2 + x_2^2 y_1^2) = |z|^2((\mu_\alpha + \mu_\beta)^2 + \sigma_\alpha^2 + \sigma_\beta^2)\,.$$

On the other hand,

$$\mathsf{E}(z' - z) = \mathsf{E}(z\delta) = z\,\mathsf{E}(\delta)\,,$$

$$\mathsf{E}(|z' - z|^2) = |z|^2\,\mathsf{E}(|\delta|^2)\,,$$

from which (2.6) and (2.7) finally follow. The last statement of the lemma has been proved. ●

A very efficient way of computing the discrete (or inverse discrete) Fourier transform is based on the fact that N can be written as a product of a great number

of factors. The particular case $N = 2^m$ was first studied by Cooley and Tukey [2]. The efficient algorithm is called the fast Fourier transform. The number of arithmetic operations needed for the computation of $\mathscr{F}(a)$ or $\mathscr{F}^{-1}(x)$ is of order $N \log_2 N$ while the usual computation of the transform requires $2N^2$ operations (the values w_{jk} are supposed to be known).

In the whole paper we consider a particular algorithm of the fast Fourier transform called the decimation in frequency [10]. There is another commonly used version of the fast Fourier transform known as the decimation in time [2], [14].

The study of the accumulation of roundoff errors in the fast Fourier transform is the first step of our investigation. This problem has been solved by Ramos [12] by deterministic means and by Kaneko and Liu [10], Thong and Liu [15], and Alt [1] by stochastic means. We present here a theorem based on the main result of Kaneko and Liu [10]. We wish to point out that in the formulae expressing the error, the error terms of higher orders are neglected both here and in the subsequent statements.

Theorem 2.2. *Let $a = \mathscr{F}(x)$ and let a'_p be the value of a_p computed by the decimation-in-frequency algorithm, where*

$$p = 2^{m-1} p_0 + \ldots + 2p_{m-2} + p_{m-1}$$

and p_k is a binary digit. Put

$$\Gamma_p = 0 \quad \textit{for rounding}\,, \tag{2.8}$$

$$\Gamma_p = \sum_{k=0}^{m-1} \gamma_{kp} \quad \textit{for chopping}\,, \tag{2.9}$$

where

$$\begin{aligned} \gamma_{kp} &= 1 + p_{m-1-k}(1 + \mu_\beta/\mu_\alpha)\,, \quad 0 \leqq k \leqq m-3\,, \\ &= 1\,, \quad k = m-2\,, \quad m-1\,. \end{aligned} \tag{2.10}$$

Put further

$$R^2(a, p) = \sum_{l=0}^{m-1} 2^{-m+l+1} c_{lp} \sum_{i_{l+1}=0}^{1} \ldots \sum_{i_{m-1}=0}^{1} \left|a_{I(l,p)}\right|^2\,, \tag{2.11}$$

where

$$\begin{aligned} c_{lp} &= 1 + p_{m-1-l}(1 + \sigma_\beta^2/\sigma_\alpha^2)\,, \quad 0 \leqq l \leqq m-3\,, \\ &= 1\,, \quad l = m-2\,, \quad m-1\,, \end{aligned} \tag{2.12}$$

and

$$I(l, p) = 2^{m-1} i_{m-1} + \ldots + 2^{l+1} i_{l+1} + 2^l p_{m-1-l} + \cdots + p_{m-1}\,. \tag{2.13}$$

Then

$$\mathsf{E}(a_p - a'_p) = -a_p \mu_\alpha \Gamma_p\,, \tag{2.14}$$

$$\mathsf{E}(|a_p - a'_p|^2) = |a_p|^2 \mu_\alpha^2 \Gamma_p^2 + \sigma_\alpha^2 R^2(a, p)\,. \tag{2.15}$$

Let $x = \mathscr{F}^{-1}(a)$ *and let* x'_p *be the computed value of* x_p. *Analogously, then*

$$\mathsf{E}(x_p - x'_p) = -x_p \mu_\alpha \tilde{\Gamma}_p , \tag{2.16}$$

$$\mathsf{E}(|x_p - x'_p|^2) = |x_p|^2 \mu_\alpha^2 \tilde{\Gamma}_p^2 + \sigma_\alpha^2 \tilde{R}^2(x, p) , \tag{2.17}$$

where

$$\tilde{\Gamma}_p = 0 \quad \textit{for rounding}, \tag{2.18}$$

$$\tilde{\Gamma}_p = \mu_\beta/\mu_\alpha + \Gamma_p \quad \textit{for chopping} , \tag{2.19}$$

and

$$\tilde{R}^2(x, p) = |x_p|^2 \sigma_\beta^2/\sigma_\alpha^2 + R^2(x, p) . \tag{2.20}$$

Proof. The expressions (2.14) and (2.15) were proved by Kaneko and Liu [10]. The equality sign in (2.14) as well in (2.15) is used in a rather approximate sense since the error terms of higher orders are neglected.

The inverse Fourier transform is calculated with the help of the fast Fourier transform algorithm. Writting

$$y_p = \overline{\sum_{k=0}^{N-1} w_{pk} \bar{a}_k} ; \quad p = 0, \ldots, N - 1 ,$$

and denoting the computed value by y'_p, we then have from (2.14) and (2.15) that

$$\mathsf{E}(y_p - y'_p) = -y_p \mu_\alpha \Gamma_p , \tag{2.21}$$

$$\mathsf{E}(|y_p - y'_p|^2) = |y_p|^2 \mu_\alpha^2 \Gamma_p^2 + \sigma_\alpha^2 R^2(y, p) . \tag{2.22}$$

The last step consists in the multiplication of the complex number y_p by N^{-1}, i.e.

$$x_p = N^{-1} y_p .$$

Denote the exact result of the multiplication of y'_p by N^{-1} by

$$x_p^* = N^{-1} y'_p$$

and the computed value of x_p^* by x'_p. Then

$$x_p - x'_p = x_p - x_p^* + x_p^* - x'_p . \tag{2.23}$$

We will study these two differences separately. From (2.21) and (2.22) we obtain

$$\mathsf{E}(x_p - x_p^*) = -x_p \mu_\alpha \Gamma_p ,$$

$$\mathsf{E}(|x_p - x_p^*|^2) = |x_p|^2 \mu_\alpha^2 \Gamma_p^2 + \sigma_\alpha^2 R^2(x, p) ,$$

where we employed also (2.11).

Further, the multiplication performed gives

$$\mathsf{E}(x_p^* - x'_p) = -\mu_\beta x_p ,$$

$$\mathsf{E}(|x_p^* - x'_p|^2) = (\mu_\beta^2 + \sigma_\beta^2) |x_p|^2 ,$$

where we put approximately $x_p^* \sim x_p$, i.e., we neglect the error terms of higher orders.

In fact, x_p^* is a random variable with its mean $\mathsf{E}(x_p^*) = \mathsf{E}(x_p^* - x_p + x_p) = \mathsf{E}(x_p^* - x_p) + x_p = x_p(\mu_\alpha \Gamma_p + 1)$. Putting now $\mu_\alpha \Gamma_p + 1 \sim 1$, we come to the above formula. We proceed in an analogous way in all similar situations.

Using finally (2.23), we obtain (2.16) and (2.17) by a straightforward calculation. •

In the following section we will need a result on the discrete Fourier transform of a vector with a random perturbation. This result is presented in the next theorem.

Theorem 2.3. *Let* $a = \mathscr{F}(x)$. *Putting* $a^* = \mathscr{F}(x^*)$, *where*

$$x_j^* = x_j + \Delta_j\,; \quad j = 0, \ldots, N-1\,, \tag{2.24}$$

and Δ_j *are mutually independent random variables with*

$$\mathsf{E}(\Delta_j) = \mu_j\,, \quad \mathsf{E}(|\Delta_j|^2) = \omega_j^2\,; \quad j = 0, \ldots, N-1\,,$$

we obtain that

$$\mathsf{E}(a_k - a_k^*) = \sum_{j=0}^{N-1} w_{jk}\mu_j\,, \tag{2.25}$$

$$\mathsf{E}(|a_k - a_k^*|^2) = \Big|\sum_{j=0}^{N-1} w_{jk}\mu_j\Big|^2 + \sum_{j=0}^{N-1}\omega_j^2 - \sum_{j=0}^{N-1}|\mu_j|^2\,, \quad k = 0, \ldots, N-1\,. \tag{2.26}$$

On the other hand, let $x = \mathscr{F}^{-1}(a)$. *Putting* $x^{**} = \mathscr{F}^{-1}(a^{**})$, *where*

$$a_k^{**} = a_k + \tilde{\Delta}_k\,; \quad k = 0, \ldots, N-1\,,$$

and $\tilde{\Delta}_k$ *are mutually independent random variables with*

$$\mathsf{E}(\tilde{\Delta}_k) = \tilde{\mu}_k\,, \quad \mathsf{E}(|\tilde{\Delta}_k|^2) = \tilde{\omega}_k^2\,; \quad k = 0, \ldots, N-1\,,$$

we now obtain that

$$\mathsf{E}(x_j - x_j^{**}) = N^{-1}\sum_{k=0}^{N-1} \overline{w}_{jk}\tilde{\mu}_k\,, \tag{2.27}$$

$$\mathsf{E}(|x_j - x_j^{**}|^2) = N^{-2}\Big(\Big|\sum_{k=0}^{N-1} \overline{w}_{jk}\tilde{\mu}_k\Big|^2 + \sum_{k=0}^{N-1}\tilde{\omega}_k^2 - \sum_{k=0}^{N-1}|\tilde{\mu}_k|^2\Big)\,, \quad k = 0, \ldots, N-1\,. \tag{2.28}$$

Proof. Substituting (2.24) into the expression for $a_k - a_k^*$, we readily obtain (2.25). Employing moreover the independence of Δ_j, we establish (2.26) as well. The relations (2.27) and (2.28) follow by the same argument.

3. DISCRETE CONVOLUTION

We now turn to the calculation of the discrete convolution of the form considered in the introduction. We first define the discrete convolution directly and then show that it can be calculated with the help of the discrete and inverse discrete Fourier transforms.

Definition 3.1. *Let f_j and g_j, $j = 0, \ldots, N-1$, be complex numbers. Putting*

$$f_{j+N} = f_j \tag{3.1}$$

for any integer j and choosing a positive integer L such that $L \leqq N$, we write

$$u_l(L) = \sum_{j=0}^{L-1} g_j f_{l-j}\,; \quad l = 0, \ldots, N-1\,. \tag{3.2}$$

Definition 3.2. *Let f_j and g_j be the complex numbers from Definition* 3.1. *Put*

$$q = \mathscr{F}(f)\,, \quad r = \mathscr{F}(g)\,, \tag{3.3}$$

$$s_k = q_k r_k\,, \quad k = 0, \ldots, N-1\,, \tag{3.4}$$

and

$$t = \mathscr{F}^{-1}(s)\,. \tag{3.5}$$

Theorem 3.1. *Let u and t be given by Definitions* 3.1 *and* 3.2. *Then $t = u(N)$.*

Proof. The statement of the theorem can be verified by a straightforward calculation. Substituting (3.4) and (3.3) into (3.5) and using Definition 2.1, we obtain

$$t_l = N^{-1} \sum_{k=0}^{N-1} \bar{w}_{kl} \Big(\sum_{p=0}^{N-1} w_{pk} f_p \sum_{j=0}^{N-1} w_{jk} g_j \Big) =$$

$$= N^{-1} \sum_{j=0}^{N-1} \sum_{p=0}^{N-1} g_j f_p \sum_{k=0}^{N-1} w_{k,p+j-l}\,.$$

Applying now Lemma 2.1, we finally have

$$t_l = \sum_{j=0}^{N-1} g_j f_{l-j}\,, \quad l = 0, \ldots, N-1\,,$$

which is (3.2) with $L = N$. The theorem has been proved.

Remark 3.1. For $L = N$, the formulae (3.2) and (3.5) thus represent the same unique quantity. It is advantageous to keep both the notations of Definitions 3.1 and 3.2 since we are interested in the numerical evaluation of this quantity which is different in the two individual cases mentioned.

The parameter L introduced in (3.2) plays an important role. Formally, $u_l(L)$ can be expressed for $L < N$ as $u_l(N)$ with a particular choice of the values g_j, i.e., as

$$\sum_{j=0}^{N-1} g_j f_{l-j}$$

with $g_j = 0$ for $j = L, \ldots, N-1$. Nobody, however, would really compute the convolution $u(L)$ in this way since it costs many superfluous arithmetic operations. The notation introduced in (3.2) is thus quite justified from this computational point of view.

On the other hand, computing the convolution t via the discrete Fourier transform (see Definition 3.2), we are not able to make use of the information that $g_j = 0$ for $j \geqq L$, where L is an integer less than N. We must always work with all the N components of g. No parameter L thus appears in this algorithm. •

We will show the influence of the roundoff error on the values computed according to Definitions 3.1 and 3.2.

Theorem 3.2. *Let $u_l'(L)$ be the computed value of $u_l(L)$ (cf. Definition 3.1). Neglecting the error terms of higher orders, we obtain*

$$\mathsf{E}(u_l(L) - u_l'(L)) = -\sum_{j=0}^{L-1} g_j f_{l-j}\, \pi_j(L)\,; \quad l = 0, \ldots, N-1\,, \tag{3.6}$$

where

$$\pi_j(L) = \mu_\beta + (L - j + 1)\,\mu_\alpha\,,$$

and

$$\mathsf{E}(|u_l(L) - u_l'(L)|^2) = \sum_{j=0}^{L-1}\sum_{p=0}^{L-1} g_j \bar{g}_p f_{l-j} \bar{f}_{l-p}\, \varrho_{jp}^2(L)\,; \quad l = 0, \ldots, N-1\,, \tag{3.7}$$

where now

$$\varrho_{jp}^2(L) = \mu_\beta + (2L - j - p + 2)\,\mu_\alpha\mu_\beta + (L - j + 1)(L - p + 1)\,\mu_\alpha^2 + \\ + (L - j)\,\sigma_\alpha^2 \quad \textit{for} \quad p < j\,, \quad \varrho_{jp}^2(L) = \varrho_{pj}^2(L)\,, \tag{3.8}$$

$$\varrho_{jj}^2(L) = (\mu_\beta + (L - j + 1)\,\mu_\alpha)^2 + \sigma_\beta^2 + (L - j + 1)\,\sigma_\alpha^2\,. \tag{3.9}$$

Proof. Computing the value of $u_l(L)$ in the usual way, we obtain from Wilkinson [17] Ch. 1, Sec. 26 that

$$u_l'(L) = \sum_{j=0}^{L-1} g_j f_{l-j}(1 + \varepsilon_j)\,,$$

where

$$1 + \varepsilon_0 = (1 + \delta_0)(1 + \gamma_1)\ldots(1 + \gamma_{L-1})\,,$$

$$1 + \varepsilon_j = (1 + \delta_j)(1 + \gamma_j)\ldots(1 + \gamma_{L-1})\,, \quad j = 1, \ldots, L-1\,,$$

and γ_k and δ_k are the roundoff errors introduced in Lemma 2.2. Neglecting the error terms of higher orders and replacing ε_0 by a slightly more pessimistic expression, we come to the formula

$$\varepsilon_j = \delta_j + \gamma_j + \gamma_{j+1} + \ldots + \gamma_{L-1}\,; \quad j = 0, \ldots, L-1\,.$$

Applying now Lemma 2.2 and making use of the independence of the roundoff errors, we finally obtain (3.6) and (3.7). •

The next theorem is concerned with the fast computation of the convolution according to Definition 3.2. For the sake of simplicity we suppose that one of the two discrete Fourier transforms is computed exactly, i.e. without the roundoff error accumulation. Such a situation is common in practice. In the computation

of the right-hand part of the system of linear algebraic equations in the finite element method it appears in the case when the right-hand part of the solved differential equation is simple (e.g. constant). Then the discrete Fourier transform of such a sequence is computed very accurately.

Similarly, the coefficients of a linear filter are usually small integers. The discrete Fourier transform of such a filter is then computed without a considerable roundoff error accumulation.

A general theorem taking into account the complete roundoff error accumulation in the computation of the discrete Fourier transforms can be proved in an analogous way.

Theorem 3.3. *We preserve the notation introduced in Theorem* 2.2 *and Definition* 3.2 *and neglect the error terms of higher orders. Let* r'_k *be the computed value of* r_k. *Then we have*

$$\mathsf{E}(r_k - r'_k) = -r_k \mu_\alpha \Gamma_k \,, \tag{3.10}$$

$$\mathsf{E}(|r_k - r'_k|^2) = |r_k|^2 \mu_\alpha^2 \Gamma_k^2 + \sigma_\alpha^2 R^2(r, k) \,, \quad k = 0, \ldots, N-1 \,, \tag{3.11}$$

where Γ_k *and* $R^2(r, k)$ *are given by* (2.8), (2.9), *and* (2.11).

Suppose that q_k *is computed exactly. Further, let* s'_k *be the computed (from* r'_k*) value of* s_k. *Then*

$$\mathsf{E}(s_k - s'_k) = -s_k \tau_k \,, \tag{3.12}$$

$$\mathsf{E}(|s_k - s'_k|^2) = |q_k|^2 (\mathsf{E}(|r_k - r'_k|^2) + |r_k|^2 \varphi_k) \,, \quad k = 0, \ldots, N-1 \,, \tag{3.13}$$

where

$$\tau_k = (1 + \Gamma_k) \mu_\alpha + \mu_\beta \,, \tag{3.14}$$

$$\varphi_k = \sigma_\alpha^2 + \sigma_\beta^2 + (\mu_\alpha + \mu_\beta)^2 + 2\mu_\alpha \Gamma_k (\mu_\alpha + \mu_\beta) \,. \tag{3.15}$$

Denoting finally by t'_l *the computed (from* s'_k*) value of* t_l, *we find that*

$$\mathsf{E}(t_l - t'_l) = -N^{-1} \sum_{k=0}^{N-1} \bar{w}_{kl} s_k \tau_k - \mu_\alpha t_l \tilde{\Gamma}_l \,, \tag{3.16}$$

$$\begin{aligned} \mathsf{E}(|t_l - t'_l|^2) &= N^{-2} \sum_{k=0}^{N-1} \sum_{\substack{p=0 \\ p \neq k}}^{N-1} \bar{w}_{kl} w_{pl} s_k \bar{s}_p \tau_k \tau_p + \\ &+ N^{-2} \sum_{k=0}^{N-1} \mathsf{E}(|s_k - s'_k|^2) + \mu_\alpha^2 |t_l|^2 \tilde{\Gamma}_l^2 + \sigma_\alpha^2 \tilde{R}^2(t, l) + \\ &+ N^{-1} \mu_\alpha \bar{t}_l \tilde{\Gamma}_l \sum_{k=0}^{N-1} \bar{w}_{kl} s_k \tau_k + N^{-1} \mu_\alpha t_l \tilde{\Gamma}_l \sum_{k=0}^{N-1} w_{kl} \bar{s}_k \tau_k \,, \quad l = 0, \ldots, N-1 \,, \end{aligned} \tag{3.17}$$

where $\tilde{\Gamma}_l$ *and* $\tilde{R}^2(t, l)$ *are given by* (2.18), (2.19), *and* (2.20).

Proof. The formulae (3.10) and (3.11) characterizing the error of r'_k follow immediately from Theorem 2.2. Putting

$$s^*_k = q_k r'_k$$

and recalling Definition 3.2, we have

$$s_k - s'_k = s_k - s^*_k + s^*_k - s'_k = q_k(r_k - r'_k) + s^*_k - s'_k .$$

Using now (3.10) and (3.11) and applying Lemma 2.2, we come to (3.12) and (3.13) after a straightforward calculation. We neglect error terms of higher orders in the whole proof.

Putting finally

$$t^* = \mathscr{F}^{-1}(s') ,$$

we can write

$$t_l - t'_l = t_l - t^*_l + t^*_l - t'_l . \tag{3.18}$$

Employing now Theorem 2.3 and the formulae (3.12) and (3.13), we readily obtain

$$\mathsf{E}(t_l - t^*_l) = -N^{-1} \sum_{k=0}^{N-1} \bar{w}_{kl} s_k \tau_k ,$$

$$\mathsf{E}(|t_l - t^*_l|^2) = N^{-2} \sum_{k=0}^{N-1} \sum_{\substack{p=0 \\ p \neq k}}^{N-1} \bar{w}_{kl} w_{pl} s_k \bar{s}_p \tau_k \tau_p + N^{-2} \sum_{k=0}^{N-1} \mathsf{E}(|s_k - s'_k|^2) ,$$

$$l = 0, \ldots, N-1 .$$

Moreover, the error $t^*_l - t'_l$ is characterized by Theorem 2.2. Combining these errors according to (3.18), we finally obtain (3.16) and (3.17).

4. AN EXAMPLE

We established Theorems 3.2 and 3.3 in order to compare the accumulation of roundoff errors in the direct calculation of the discrete convolution according to Definition 3.1 and in the fast calculation according to Definition 3.2. The formulae presented, however, are so complex that we can obtain such a comparison only for a relatively simple example. The theoretical statements are supported by numerical results computed on an IBM System/370 Model 135 computer. Since this machine has chopped arithmetic we consider only chopping throughout this section.

We first show some bounds for the quantities appearing in Theorem 3.3.

Lemma 4.1. *Consider the chopped arithmetic. Let* Γ_p, $\tilde{\Gamma}_p$, τ_p, *and* φ_p *be given by* (2.9), (2.19), (3.14) *and* (3.15). *Then there exist positive constants* C_Γ, $\tilde{C}_\Gamma$, C_τ *and* C_φ *independent of* N *and* p *such that, for* $p = 0, \ldots, N-1$,

$$m \leqq \Gamma_p \leqq C_\Gamma m \,, \tag{4.1}$$

$$\tilde{\Gamma}_p \leqq \tilde{C}_\Gamma m \,, \tag{4.2}$$

$$\tau_p \leqq C_\tau m \,, \tag{4.3}$$

$$\varphi_p \leqq C_\varphi m \,. \tag{4.4}$$

Further, let $R^2(a, p)$ *and* $\tilde{R}^2(x, p)$ *be given by* (2.11) *and* (2.20). *Then there exist positive constants* $C_{\Sigma R}$ *and* $\tilde{C}_{\Sigma R}$ *independent of* N *such that*

$$\sum_{p=0}^{N-1} R^2(y, p) \leqq C_{\Sigma R} m \sum_{p=0}^{N-1} |y_p|^2 \,, \tag{4.5}$$

$$\sum_{p=0}^{N-1} \tilde{R}^2(y, p) \leqq \tilde{C}_{\Sigma R} m \sum_{p=0}^{N-1} |y_p|^2 \,. \tag{4.6}$$

If, moreover,

$$|y_p|^2 \leqq C_y \,, \quad p = 0, \ldots, N-1 \,, \tag{4.7}$$

then

$$R^2(y, p) \leqq m C_y C_{\Sigma R} \,,$$

$$\tilde{R}^2(y, p) \leqq m C_y \tilde{C}_{\Sigma R} \,, \quad p = 0, \ldots, N-1 \,.$$

Proof. Putting $C_\Gamma = 2 + \mu_\beta/\mu_\alpha$ and considering the relations (2.9) and (2.10), we immediately obtain (4.1). Further, (4.2) holds with $\tilde{C}_\Gamma = 2(1 + \mu_\beta/\mu_\alpha)$ with respect to (2.19) and (4.1). Now putting $C_\tau = 3\mu_\alpha + 2\mu_\beta$ and $C_\varphi = \sigma_\alpha^2 + \sigma_\beta^2 + 5\mu_\alpha^2 + 8\mu_\alpha\mu_\beta + 3\mu_\beta^2$, and taking (3.14), (3.15) and (4.1) into account, we come to (4.3) and (4.4).

To proceed further, we use the relation (Kaneko and Liu [10])

$$\sum_{p=0}^{N-1} \sum_{i_1+1=0}^{1} \cdots \sum_{i_{m-1}=0}^{1} |a_{I(l,p)}|^2 = 2^{m-1-l} \sum_{p=0}^{N-1} |a_p|^2 \,, \quad l = 0, \ldots, m-1 \,, \tag{4.8}$$

where $I(l, p)$ is given by (2.13). Considering the definition (2.11) of R^2 and (2.12), and putting $C_{\Sigma R} = 2 + \sigma_\beta^2/\sigma_\alpha^2$, we obtain (4.5). Analogously, using (2.20), (4.5) and (4.8), we establish (4.6) with $\tilde{C}_{\Sigma R} = 2(1 + \sigma_\beta^2/\sigma_\alpha^2)$.

The last two inequalities follow directly from (2.11) (and (2.20)) if we apply the assumption (4.7) to each term appearing in the sum over l and also employ (2.12). ●

We now give an example illustrating the statements of the previous section.

Choosing positive integers P and N, $P \leqq N$, and turning to Definitions 3.1 and 3.2, we put

$$f_j = j \,; \qquad j = 0, \ldots, N-1 \,, \tag{4.9}$$

$$g_j = 1/P \,; \quad j = 0, \ldots, P-1 \,, \tag{4.10}$$

$$= 0 \,; \qquad j = P, \ldots, N-1 \,.$$

For this choice of f and g, all the quantities introduced in Definitions 3.2 and 3.1 are presented in the following lemma.

Lemma 4.2. *Let P and N be positive integers, $P \leqq N$. Let f and g be given by* (4.9) *and* (4.10). *The formulae* (3.3), (3.4), (3.5) *and* (3.2) *then read*

$$q_0 = \tfrac{1}{2}N(N-1)\,, \tag{4.11}$$

$$q_k = -\frac{N}{1-\exp(2\pi ik/N)}\,, \quad k = 1, \ldots, N-1\,, \tag{4.12}$$

$$r_0 = 1\,, \tag{4.13}$$

$$r_k = \frac{1-\exp(2\pi ikP/N)}{P(1-\exp(2\pi ik/N))}\,, \quad k = 1, \ldots, N-1\,, \tag{4.14}$$

$$s_0 = \tfrac{1}{2}N(N-1)\,, \tag{4.15}$$

$$s_k = -\frac{N(1-\exp(2\pi ikP/N))}{P(1-\exp(2\pi ik/N))^2}\,; \quad k = 1, \ldots, N-1\,, \tag{4.16}$$

$$\begin{aligned} t_l = u_l(L) &= l - \tfrac{1}{2}P + \tfrac{1}{2} + N(-l+P-1)/P\,, \quad l = 0, \ldots, P-2\,, \\ &= l - \tfrac{1}{2}P + \tfrac{1}{2}\,, \quad l = P-1, \ldots, N-1\,; \end{aligned} \tag{4.17}$$

for $L = P, \ldots, N$.

Proof. Using the formulae for the sum of a finite arithmetical-geometrical (Jolley [9]) and geometrical series, we readily obtain (4.11) to (4.14). The quantities (4.15) or (4.16) are simple products of (4.11) and (4.13) or (4.12) and (4.14).

We have $u(L) = u(N)$; $L = P, \ldots, N$, with respect to (4.10). Finally, $t = u(L)$; $L = P, \ldots, N$, since Theorem 3.1 states that $t = u(N)$. We can thus use Definition 3.1 with $L = P$ to calculate (4.17) and obtain

$$t_l = \sum_{j=0}^{P-1} g_j f_{l-j} = P^{-1} \sum_{j=l-P+1}^{l} f_j$$

after the substitution of (4.10). Employing now (3.1) and (4.9) we come to (4.17) with the help of a straightforward computation. •

Several inequalities for q_k, r_k and s_k are established in the next lemma.

Lemma 4.3. *Let q_k, r_k and s_k, $k = 0, \ldots, N-1$, be given by* (4.11) *to* (4.16). *Then there exist positive constants C_{1q}, C_{2q}, C_{1r}, C_{2r}, $C_{\Sigma r}$ and $C_{\Sigma s}$ independent of N such that*

$$|q_k|^2 \leqq C_{2q}^2 N^4\,, \quad k = 0, \ldots, N-1\,, \tag{4.18}$$

$$C_{1q}^2 N^4 \leqq |q_1|^2\,, \tag{4.19}$$

$$|r_k|^2 \leqq C_{2r}^2\,, \quad k = 0, \ldots, N-1\,, \tag{4.20}$$

(4.21) $$C_{1r}^2 \leqq |r_1|^2 ,$$

(4.22) $$\sum_{k=0}^{N-1} |q_k|^2 = \tfrac{1}{3}N^4 + O(N^3) \quad as \quad N \to \infty , \quad \sum_{k=0}^{N-1} |r_k|^2 = C_{\Sigma r} N ,$$

(4.23) $$\sum_{k=0}^{N-1} |s_k| \leqq C_{\Sigma s} m N^2$$

and

$$\sum_{k=0}^{N-1} |s_k|^2 = \tfrac{1}{3}N^4 + O(N^3) \quad as \quad N \to \infty .$$

Proof. Using elementary trigonometric formulae, we find that

(4.24) $$|q_k|^2 = \frac{N^2}{4 \sin^2 (\pi k/N)}, \quad k = 1, \ldots, N/2 ,$$

follows from (4.12). There exist positive constants C_l and C_u such that

$$0 < C_l \leqq \frac{x}{\sin x} \leqq C_u \quad \text{for} \quad x \in \langle 0, \pi/2 \rangle$$

as the function $S(x) = x/\sin x$ is continuous in $\langle 0, \pi/2 \rangle$, has no zeros in this interval, and $S(0) = 1$, $S(\pi/2) = \pi/2$. Thus

(4.25) $$\frac{C_l}{x} \leqq \frac{1}{\sin x} \leqq \frac{C_u}{x} \quad \text{for} \quad x \in (0, \pi/2 \rangle .$$

The inequality (4.25) gives now the upper bound (4.18) for $|q_k|^2$, $k = 1, \ldots, N/2$, when applied to (4.24). For $k = 0$, (4.18) is evident. As $q_{N-k} = \bar{q}_k$, we can choose a suitable constant C_{1q} for (4.18) to hold for $k = 0, \ldots, N - 1$. Analogously, (4.24) and (4.25) imply (4.19).

To prove (4.20) and (4.21), we proceed similarly. We can rewrite (4.14) as

(4.26) $$|r_k|^2 = \frac{\sin^2 (\pi k P/N)}{P^2 \sin^2 (\pi k/N)} ; \quad k = 1, \ldots, N - 1 .$$

Examining now the behavior of the function $T(x) = |\sin xP|/\sin x$ in $\langle 0, \pi \rangle$ and considering (4.26), we come finally to (4.20) and (4.21). Recalling the definition (3.3) of q_k and r_k and (4.9), (4.10), we readily obtain (4.22) as a consequence of Parseval's identity (2.2).

According to (3.4) we now combine the expressions (4.24) and (4.26) for $|q_k|$ and $|r_k|$ and their upper bounds used above to obtain

(4.27) $$|s_k| \leqq C' N^2/k , \quad k = 1, \ldots, N - 1 ,$$

with a constant C' independent of N. The estimate (4.23) then follows from (4.15), (4.27) and the well-known formula

$$\sum_{k=1}^{K} \frac{1}{k} \leqq \mathrm{C} + \log K + \frac{1}{2K} ,$$

where C is the Euler constant. The last inequality is again a consequence of Parseval's identity (2.2) after the substitution of (4.17) for t_l, which completes the proof. •

We now turn to the results of the numerical example. We computed the values of $u(P)$ and t according to Definitions 3.1 and 3.2 with f and g given by (4.9) and (4.10) in single precision on an IBM System/370 Model 135 computer (with chopped arithmetic). We used the decimation-in-frequency algorithm of the fast Fourier transform. We recall that the choice of g implies that $u(L) = t$ for $L = P, \ldots, N$ (see Lemma 4.2).

The parameter N varied from P to 4096. We tested the values $P = 16$, 32 and 64, which all showed a very similar behavior. In the following the results with $P = 16$ are presented.

We assumed that r was computed exactly. The experiments carried out confirmed that this assumption was reasonable. Moreover, to obtain the expected roundoff error in the computation of q and u, we "spoiled" the values of g_j by small random perturbations (uniformly distributed between $-0{\cdot}5 \times 10^{-6}$ and $0{\cdot}5 \times 10^{-6}$). Otherwise, the number $1/P$ was represented in the machine exactly and the expected roundoff error accumulation did not occur.

The error of the computed values r, s, t, and u was measured by the norms introduced in the following definition.

Definition 4.1. *Let v and v' be vectors from $\mathbb{C}_N$. We put*

$$\|v - v'\|_2 = \left(N^{-1} \sum_{k=0}^{N-1} |v_k - v'_k|^2\right)^{1/2},$$

$$\|v - v'\|_\infty = \max_{k=0,\ldots,N-1} |v_k - v'_k|.$$

If $v_k \neq 0$, $k = P - 1, \ldots, N - 1$, we further write

$$\|v - v'\|_2^* = \left((N - P + 1)^{-1} \sum_{k=P-1}^{N-1} |v_k - v'_k|^2 / |v_k|^2\right)^{1/2},$$

$$\|v - v'\|_\infty^* = \max_{k=P-1,\ldots,N-1} |v_k - v'_k| / |v_k|.$$

Employing Theorems 3.2 and 3.3 and the lemmas of this section. we now estimate the order of the roundoff error (depending on N) in the computation of our particular example of the discrete convolution. The estimates can be compared with the computed experimental results. Their agreement is good. (Let us note that another example, where we put

$$f_j = 1; \quad j = 0, \ldots, N - 1;$$

instead of (4.9), gave also a good agreement of the theoretical estimates of roundoff errors with the actual experimental values.)

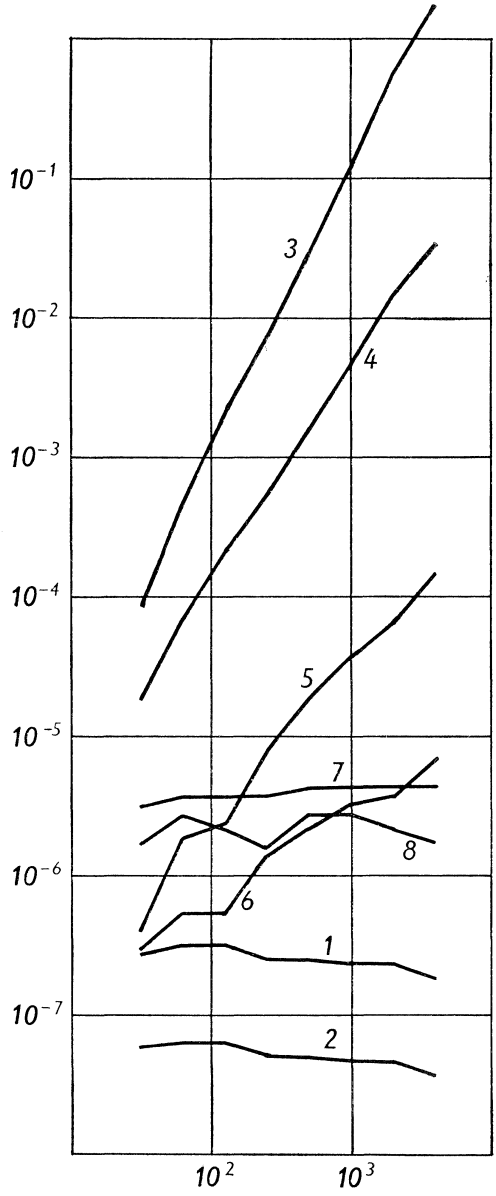

Fig. 4.1.

1	$\|r - r'\|_\infty/m$	2	$\|r - r'\|_2/m$
3	$\|s - s'\|_\infty/m$	4	$\|s - s'\|_2/m$
5	$\|t - t'\|^*_\infty/m$	6	$\|t - t'\|^*_2/m$
7	$\|u - u'\|^*_\infty$	8	$\|u - u'\|^*_2$

The errors $\|r - r'\|_k$, $\|s - s'\|_k$, $\|t - t'\|_k^*$ and $\|u(P) - u'(P)\|_k^*$ are presented for $k = \infty$ and $k = 2$ in Fig. 4.1, where N is the independent variable, $32 \leqq N \leqq 4096$. The scale of the variable N (horizontal) as well as that of the error (vertical) are logarithmic. The lines 1, 3, 5 and 7 correspond to the quantities $\|r - r'\|_\infty/m$, $\|s - s'\|_\infty/m$, $\|t - t'\|_\infty^*/m$ and $\|u(P) - u'(P)\|_\infty^*$, respectively, while the lines 2, 4, 6 and 8 correspond to the same errors but this time with the subscript 2. The errors the order of which depends also on $m = \log_2 N$ are divided by this value to make the slope of the lines clearer in the figure with both axes having logarithmic scales.

The quantity $\mathsf{E}(|r_k - r'_k|^2)$ is given by (3.11). Using (4.13), (4.20) and Lemma 4.1, we come to the estimate

$$\mu_\alpha^2 m^2 \leqq \mathsf{E}(|r_0 - r'_0|^2) \leqq \max_{k=0,\ldots,N-1} \mathsf{E}(|r_k - r'_k|^2) \leqq D_{2r}^2 m^2 , \tag{4.28}$$

where D_{2r} is a constant independent of N. Summing further the formula (3.11) over k and employing (4.1), (4.5) and (4.22), we finally obtain

$$D_{1\Sigma r}^2 m^2 \leqq N^{-1} \sum_{k=0}^{N-1} \mathsf{E}(|r_k - r'_k|^2) \leqq D_{2\Sigma r}^2 m^2 , \tag{4.29}$$

where $D_{1\Sigma r}$ and $D_{2\Sigma r}$ are constants independent of N. The estimates (4.28) and (4.29) thus show that $\|r - r'\|_2$ as well as $\|r - r'\|_\infty$ should be of order m. This fact is confirmed by lines 1 and 2 in Fig. 4.1.

The quantity $\mathsf{E}(|s_k - s'_k|^2)$ is given by (3.13). Proceeding similarly to the previous considerations and using (4.28) and Lemmas 4.1 and 4.3, we find that

$$D_{1s}^2 m^2 N^4 \leqq \mathsf{E}(|s_1 - s'_1|^2) \leqq \max_{k=0,\ldots,N-1} \mathsf{E}(|s_k - s'_k|^2) \leqq D_{2s}^2 m^2 N^4 , \tag{4.30}$$

where D_{1s} and D_{2s} are constants independent of N. Summing again (3.13) over k and employing (2.11), (3.11) and also Lemmas 4.1 and 4.3, we come to the estimate

$$D_{1\Sigma s}^2 m^2 N^3 \leqq N^{-1} \sum_{k=0}^{N-1} \mathsf{E}(|s_k - s'_k|^2) \leqq D_{2\Sigma s}^2 m^2 N^3 , \tag{4.31}$$

where $D_{1\Sigma s}$ and $D_{2\Sigma s}$ are constants independent of N. The inequality (4.30) now shows that $\|s - s'\|_\infty$ should be of order mN^2 while (4.31) states that $\|s - s'\|_2$ is to be of order $mN^{3/2}$. Both these orders are confirmed by lines 3 and 4 in Fig. 4.1.

We now turn to the statement of Theorem 3.3 concerned with $t - t'$. Recalling that in the select-saving method (cf. Helms [6]) only those t_l's with $l = P - 1, \ldots, N - 1$ are saved, we will examine $\mathsf{E}(|t_l - t'_l|^2)$ and $\mathsf{E}(|u_l - u'_l|^2)$ only for these values of l in the following. The corresponding norms are introduced in Definition 4.1. The analysis for $l = 0, \ldots, N - 1$, however, could be readily performed in the same manner.

Moreover, we will study the relative error, which enables us to distinguish better between the behavior of roundoff errors in the direct and the fast computation.

Considering (3.17) and applying (4.31) and Lemmas 4.1 to 4.3, we come to the estimate

$$\mathsf{E}(|t_l - t'_l|^2) \leqq C_{\Sigma s}^2 C_\tau^2 m^4 N^2 + D_{2\Sigma s}^2 m^2 N^2 + \mu_\alpha^2 \tilde{C}_T^2 |t_l|^2 m^2 + \\ + \sigma_\alpha^2 \tilde{C}_{\Sigma R} |t_{N-1}|^2 m + 2\mu_\alpha \tilde{C}_T C_\tau C_{\Sigma s} |t_l| m^3 N$$

and thus

$$\mathsf{E}(|t_l - t'_l|^2) |t_l|^{-2} \leqq |t_l|^{-2} (\Delta_{2t}^2 m^4 N^2 + \Delta_{2t}^{*2} m^2 N^2) \leqq \\ \leqq D_{2t}^2 m^4 N^2 + D_{2t}^{*2} m^2 N^2 ; \quad l = P - 1, \ldots, N - 1 , \tag{4.32}$$

where Δ_{2t}, Δ_{2t}^*, D_{2t} and D_{2t}^* are constants independent of N and l. The reason for keeping two terms in the estimate will be apparent from the following. It would be rather difficult to obtain also a lower bound for the relative error; we are not going to do it. Summing now (4.32) from $L - 1$ to $N - 1$, we arrive at

$$(N - P + 1)^{-1} \sum_{l=P-1}^{N-1} \mathsf{E}(|t_l - t'_l|^2) |t_l|^{-2} \leqq D_{2\Sigma t}^2 m^4 N + D_{2\Sigma t}^{*2} m^2 N , \tag{4.33}$$

where $D_{2\Sigma t}$ and $D_{2\Sigma t}^*$ are constants independent of N. We see from (4.32) that the upper bound for $\|t - t'\|_\infty^*$ is of order $m^2 N$ and from (4.33) that the upper bound for $\|t - t'\|_2^*$ is of order $m^2 N^{1/2}$. Lines 5 and 6 in Fig. 4.1 show that these estimates are pessimistic. The experimental orders seem to be mN and $mN^{1/2}$ for $\|t - t'\|_\infty^*$ and $\|t - t'\|_2^*$, respectively, which corresponds to the second terms in (4.32) and (4.33).

Finally, we examine the roundoff error in the direct computation as established in Theorem 3.2. We have

$$\mathsf{E}(|u_l(P) - u'_l(P)|^2) = P^{-2} \sum_{j=0}^{P-1} \sum_{p=0}^{P-1} (l - j)(l - p) \varrho_{jp}^2(P) , \quad l = 0, \ldots, N - 1 ,$$

where $\varrho_{jp}^2(P)$ is independent of l and N in accordance with (3.8) and (3.9), i.e.,

$$\mathsf{E}(|u_l(P) - u'_l(P)|^2) = C_u^2 l^2 + C_u^{*2} l + C_u^{**2} ; \quad l = P - 1, \ldots, N - 1 ,$$

where C_u, C_u^* and C_u^{**} are constants independent of l and N. Lemma 4.2 now gives

$$D_{1u}^2 \leqq \mathsf{E}(|u_l(P) - u'_l(P)|^2) |u'_l(P)|^{-2} \leqq D_{2u}^2 , \quad l = P - 1, \ldots, N - 1 , \tag{4.34}$$

where D_{1u} and D_{2u} are constants independent of l and N. Finally,

$$D_{1\Sigma u}^2 \leqq (N - P + 1)^{-1} \sum_{l=P-1}^{N-1} \mathsf{E}(|u_l(P) - u'_l(P)|^2) |u_l(P)|^{-2} \leqq D_{2\Sigma u}^2 , \tag{4.35}$$

where $D_{1\Sigma u}$ and $D_{2\Sigma u}$ are constants independent of N, is an easy consequence of (4.34). The estimates (4.34) and (4.35) show that $\|u(P) - u'(P)\|_2^*$ as well as $\|u(P) - u'(P)\|_\infty^*$ should be of order 1. This fact is confirmed by lines 7 and 8 in Fig. 4.1.

As the error of the fast computation grows with N increasing, a comparison with the (bounded) error of the direct computation shows that the relations

$$\|t - t'\|_\infty^* \geqq \|u(P) - u'(P)\|_\infty^* ,$$

$$\|t - t'\|_2^* \geqq \|u(P) - u'(P)\|_2^*$$

hold for $N \geqq 128$ for the actually computed results.

In Theorems 2.2 and 3.3 we supposed that the values of w_{jk} were computed exactly, i.e. without the accumulation of roundoff errors. The decimation-in-frequency algorithm employed to obtain the results described above possesses this property. However, many decimations in frequency commonly used compute w_{jk} from recurrence formulae and the roundoff error in the computation of w_{jk} is accumulated. The experiments performed show that the influence of these accumulated roundoff errors may be substantial. The quantities $\|t - t'\|_\infty^*$ and $\|t - t'\|_2^*$ may be much greater than those computed with exact w_{jk} and their growth (with N increasing) may be much quicker.

The algorithm of the fast Fourier transform available from the IBM Scientific Subroutine Package [14] is the decimation in time and the results of Kaneko and Liu [10] and of this paper do not apply to it. The experiments performed, however, give results very similar to Fig. 4.1. It is in agreement with Ramos [12] who obtained bounds for the roundoff error in the fast Fourier transform inedpendently of the particular version of the algorithm.

References

[1] *R. Alt:* Error propagation in Fourier transforms. Math. Comput. Simulation 20 (1978), 37–43.

[2] *J. W. Cooley, J. W. Tukey:* An algorithm for the machine calculation of complex Fourier series. Math. Comp. 19 (1965), 297–301.

[3] *P. J. Davis, P. Rabinowitz:* Methods of Numerical Integration. Academic Press, New York 1975.

[4] *R. W. Hamming:* Numerical Methods for Scientists and Engineers. McGraw-Hill, New York 1962.

[5] *D. R. Hartree:* Note on systematic roundoff errors in numerical integration. J. Res. Nat. Bur. Standards 42 (1949), 62.

[6] *H. D. Helms:* Fast Fourier transform method of computing difference equations and simulating filters. IEEE Trans. Audio Electroacoust. AU-15 (1967), 85–90.

[7] *P. Henrici:* Elements of Numerical Analysis. Wiley, New York 1964.

[8] *H. D. Huskey:* On the precision of a certain procedure of numerical integration. J. Res. Nat. Bur. Standards 42 (1949), 57–62.

[9] *L. Jolley:* Summation of Series. Chapman and Hall, London 1925.

[10] *T. Kaneko, B. Liu:* Accumulation of round-off error in fast Fourier transforms. J. Assoc. Comput. Mach. 17 (1970), 637–654.

[11] *T. Kaneko, B. Liu:* On local roundoff errors in floating-point arithmetic. J. Assoc. Comput. Mach. 20 (1973), 391–398.

[12] *G. U. Ramos:* Roundoff error analysis of the fast Fourier transform. Math. Comp. 25 (1971), 757–768.

[13] *P. H. Sterbenz:* Floating-Point Computation. Prentice-Hall, Englewood Cliffs, N. J., 1974.
[14] System/360 Scientific Subroutine Package. IBM Corporation, White Plains, N. Y., 1970.
[15] *T. Thong, B. Liu:* Accumulation of roundoff errors in floating point FFT. IEEE Trans. Circuits and Systems 24 (1977), 132—143.
[16] *T. Thong, B. Liu:* Floating point fast Fourier transform computation using double precision floating point accumulators. ACM Trans. Math. Software 3 (1977), 54—59.
[17] *J. H. Wilkinson:* Rounding Errors in Algebraic Processes. HMSO, London 1963.

Souhrn

ZAOKROUHLOVACÍ CHYBY PŘI RYCHLÉM VÝPOČTU DISKRÉTNÍCH KONVOLUCÍ

KAREL SEGETH

Úloha vyčíslit efektivně hodnoty diskrétní konvoluce (1.1) se v praxi převádí na opakované vyčíslení hodnot konvoluce typu (1.2) pomocí rychlé Fourierovy transformace při optimální volbě parametru N. Článek je věnován analýze zaokrouhlovacích chyb při rychlém výpočtu konvoluce (1.2) (věta 3.3); pro srovnání se analyzují zaokrouhlovací chyby i při obvyklém (přímém) výpočtu této konvoluce (věta 3.2). Používá se stochastický model šíření zaokrouhlovacích chyb. Teoretické výsledky jsou porovnány se skutečnými zaokrouhlovacími chybami, které se projeví při vyčíslení diskrétní konvoluce pro jistou jednoduchou volbu posloupností (4.9) a (4.10).

Author's address: RNDr. *Karel Segeth*, CSc., Matematický ústav ČSAV, Žitná 25, 115 67 Praha 1.