MATHEMATICS OF COMPUTATION
VOLUME 59, NUMBER 200
OCTOBER 1992, PAGES 569-582

A STOCHASTIC ROUNDOFF ERROR ANALYSIS FOR THE CONVOLUTION

DANIELA CALVETTI

ABSTRACT. We study the accuracy of an algorithm which computes the convolution via Radix-2 fast Fourier transforms. Upper bounds are derived for the expected value and the variance of the accompanying linear forms in terms of the expected value and variance of the relative roundoff errors for the elementary operations of addition and multiplication. These results are compared with the corresponding ones for two algorithms computing the convolution directly, via Horner's sums and using cascade summation, respectively.

1. INTRODUCTION

In this paper we use a statistical model of error propagation to derive bounds on the first-order approximation of the absolute roundoff error of the algorithm that computes the convolution product of two vectors via Radix-2 fast Fourier transform (FFT). Circular convolution (CC) is a fundamental computational tool in many different fields where repeated computations of convolution products of large vectors are needed. One such example is the implementation of the class of digital filters which have an impulse response of finite duration. The output samples of such filters are obtained from the results of convolution products of the filter impulse response—the kernel—and sections of the input. Owing to the dimensions of the vectors involved and to the need for repeated computations, the direct evaluation of the convolution product is usually prohibitively expensive [8]. While some studies of the rounding error for the fast Fourier transform can be found in the literature (see [2, 6, 7, 9, 12, 14]), the issue of the numerical stability of circular convolution is only briefly addressed in [6].

In the present paper we compare the numerical stability of circular convolution using a unitary scaled Radix-2 FFT with the accuracy of two algorithms which compute the convolution directly.

The model of error propagation employed, based on the usual assumptions of floating-point arithmetic, is both linear and stochastic. The model is linear in the sense that the *absolute* global errors are approximated by the first-order terms of the Taylor expansion in local relative errors. It is stochastic in the sense that the local relative errors are regarded as random variables, independently

Received by the editor December 17, 1990 and, in revised form, May 20, 1991.
1991 *Mathematics Subject Classification.* Primary 44A35, 65G05.
Key words and phrases. Accompanying linear forms, floating-point arithmetic, Radix-2 fast Fourier transform, convolution, rounding errors, random variables.

0025-5718/92 $1.00 + $.25 per page

and identically distributed (i.i.d.) for each elementary operation in which they arise. This method of analysis makes it possible to measure the error in the final output by its statistical properties, i.e., its expected value and variance, rather than in worst-case terms, thus yielding more realistic bounds on the size of the error. In fact, the relative errors are random variables taking on values at most as large as the constant ε_M used in worst-case estimates. Therefore, their expected value is smaller than ε_M or, in any case, not larger. It is common practice in stochastic roundoff error analyses to assume that the relative errors have zero mean, leaving to the root mean square the task of measuring the size of the error. In this paper we do not adopt the zero mean assumption for the relative errors in order to take into consideration the case where the expected value of the relative errors for different operations are different and not equal to zero. The statistical properties of the final output will depend in turn on the distributional properties of the local errors arising in elementary operations.

The paper is organized as follows. In §2 we describe the model of roundoff error propagation used and the method of error analysis used. Section 3 contains the mathematical results which make it possible to use the FFT for the computation of the convolution. In §4 we describe the three algorithms and we derive their accompanying linear forms, estimating their expected value and variance. In §5 we discuss the results of some numerical experiments testing the validity of the bounds found in §4.

2. Assumptions and methodology

The floating-point representation of a real number $x \neq 0$ in the machine M is of the form

$$\tilde{x} = \pm m b^l,$$

where b is the base of the machine, l is an integer such that $-L \leq l \leq U$, and the mantissa m is a T-digit number in base b such that $b^{-1} \leq m < 1$. Let x and y be elements of R_M, the finite and discrete set of floating-point representations of the reals in the machine M, and let f be an elementary operation. The machine M will compute $f(x, y)^\Delta$, which is in R_M, where

$$f(x, y)^\Delta = f(x, y)(1 + \varepsilon)$$

with $|\varepsilon| \leq \varepsilon_M$, the constant ε_M depending only on M. The quantity ε is the local *relative error* associated with the operation f. Equivalently,

$$\varepsilon = \frac{f(x, y)^\Delta - f(x, y)}{f(x, y)}.$$

The algorithms that we are going to examine start with a set of data, the entries of the two vectors to be convoluted, and produce a set of intermediate results, $t_1, \ldots, t_k$, such that

$$t_k = f_k(t_1, \ldots, t_{k-1}; x_0, \ldots, x_{n-1}; y_0, \ldots, y_{n-1}),$$

where f_k is an elementary operation which operates on at most two of the values $t_1, \ldots, t_{k-1}; x_0, \ldots, x_{n-1}; y_0, \ldots, y_{n-1}$. Since in computation each data value x_j, y_j is replaced by its machine representation, $\tilde{x}_j, \tilde{y}_j$, and each t_j by its computed value t_j^Δ, if ε_k is the local relative error for the kth operation, then

$$t_k^\Delta = f_k(t_1^\Delta, \ldots, t_{k-1}^\Delta; \tilde{x}_0, \ldots, \tilde{x}_{n-1}; \tilde{y}_0, \ldots, \tilde{y}_{n-1})(1 + \varepsilon_k).$$

In the present paper we assume that the initial data are machine numbers, hence our roundoff error analysis does not take into account the effect of rounding the initial data on the accuracy of the output. We will also assume that in the course of the calculation we do not have any problems with underflow or overflow. Since the elementary operations in the algorithms to be considered are differentiable, we can write for each component of t_k^{Δ}

$$t_k^{\Delta} = t_k + \lambda_k(\varepsilon) + O(\|\varepsilon\|^2).$$

The λ_k's, homogeneous linear functions of the local errors $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_k)$, are known as the *accompanying linear forms* of the algorithm [10, 11]. The first-order absolute errors for the intermediate and final results of the algorithms to be considered are completely described by these forms. Since $|\varepsilon_k| \leq \varepsilon_M$, it follows that

$$|t_k^{\Delta} - t_k| \leq |\lambda_k(\varepsilon_M)| + O(\varepsilon_M^2),$$

from which worst-case type error bounds can be derived.

In the present work we assume that the local relative errors ε_j are random variables with given distributions. The distributional properties of the first-order approximation of the absolute global error can be estimated by computing the expected value and variance of the accompanying linear forms in terms of the corresponding parameters for the local errors.

The main results of this paper are the bounds on $E(\lambda a)$ and $E(\lambda m)$ for the output of the circular convolution in terms of the expected values of the relative roundoff errors for the elementary operations of addition and multiplication. In the course of the rounding error analysis of the circular convolution we utilize the results of a similar type of analysis for the Radix-2 fast Fourier transform [2]. The bounds for the expected value and variance of the contribution to the linear approximation of the absolute roundoff error coming from additions and multiplications for the Radix-2 fast Fourier transform are functions of the expected values and variances of the relative rounding errors for addition and multiplication, μ_a, σ_a^2, μ_m, σ_m^2, respectively.

3. The convolution

If $\mathbf{x} = \{x_k\}$ and $\mathbf{y} = \{y_k\}$ are two sequences in the space Π_n of sequences of complex numbers which are periodic with period n and infinite in both directions, their convolution product is defined to be the sequence $\mathbf{z} = \mathbf{x} * \mathbf{y}$ in Π_n such that

$$z_k = \frac{1}{n} \cdot \sum_{m=0}^{n-1} x_m y_{k-m}. \tag{1}$$

Each component of $\mathbf{z}$ is the sum of n products; since n such components need to be computed, the cost of direct calculation of the convolution product according to (1) amounts to n^2 complex multiplications.

The following theorem shows how the convolution product can be expressed in terms of Fourier transforms.

Theorem 1. *The discrete Fourier transform, $\hat{\mathbf{z}}$, of the convolution product of two n-dimensional vectors* $\mathbf{x}$ *and* $\mathbf{y}$ *is a scalar multiple of the componentwise product of their discrete Fourier transforms, that is,*

$$\hat{\mathbf{z}} = n(\hat{\mathbf{n}} \,\square\, \hat{\mathbf{y}}), \tag{2}$$

where $\square$ *indicates componentwise product and* $\hat{\mathbf{w}}$ *is the discrete Fourier transform of one period of* $\mathbf{w}$, *for* $\mathbf{w} = \mathbf{x}, \mathbf{y}, \mathbf{z}$, *defined as follows*:

$$\hat{w}_j = \frac{1}{\sqrt{n}} \sum_{k=0}^{n-1} w_j \exp\left(-\frac{2\pi i j k}{n}\right).$$

Proof. See [5].

Let $n = 2^l$ and W_j be an $n \times n$ block diagonal matrix of the form

$$W_j = \frac{1}{\sqrt{2}} \begin{bmatrix} I_j & D_j & & \\ I_j & -D_j & \ddots & \\ & \ddots & \ddots & I_j \quad D_j \\ & & & I_j \quad -D_j \end{bmatrix},$$

where I_j is the $2^{j-1} \times 2^{j-1}$ identity matrix and D_j is the $2^{j-1} \times 2^{j-1}$ diagonal matrix with entries

$$D_j(p, p) = \exp\left(-\frac{2\pi i (p-1) 2^{l-j}}{n}\right).$$

Corollary 2. *If* n *is a power of* 2, *then*

$$\mathbf{z} = \mathbf{x} * \mathbf{y} = \frac{1}{n}(\overline{F}_n)((F_n\mathbf{x}) \square (F_n\mathbf{y})), \tag{3}$$

where F_n *is the matrix* $F_n = W_l \cdots W_1$, $\overline{F}_n$ *is its conjugate transpose, and* $\square$ *is the componentwise product.*

Proof. If n is a power of 2, $n = 2^l$, the discrete Fourier transform of an n-dimensional vector $\mathbf{z}$ can be computed via the Radix-2 FFT algorithm, that is,

$$\hat{\mathbf{z}} = W_l \cdots W_1 \mathbf{z}.$$

From Theorem 1 it follows that

$$F_n\mathbf{z} = \frac{1}{n}((F_n\mathbf{x}) \square (F_n\mathbf{y})).$$

Since the matrix $F_n = W_l \cdots W_1$ is invertible and $(F_n)^{-1} = \overline{F}_n$, we have

$$\mathbf{z} = \mathbf{x} * \mathbf{y} = \frac{1}{n}\overline{F}_n(F_n\mathbf{x} \square F_n\mathbf{y}). \quad \square$$

The convolution product performed according to (3) requires three Fourier transforms and n complex multiplications. Since the computational cost of each Radix-2 FFT is only $\frac{1}{2} \cdot n \log_2 n$ complex multiplications, the computational cost of the circular convolution can be reduced to $(\frac{3}{2} \cdot \log_2 n + 1) \cdot n$ complex multiplications. The scalar multiplication by $\frac{1}{n}$ is not taken into account because it just amounts to a shift in the exponent in the binary representation of the number.

4. Accompanying linear forms

In this section we compute the accompanying linear forms for three algorithms which compute the convolution product, and we obtain bounds for the mean and the variance of these forms.

We start with the accompanying linear forms for the algorithm which computes the convolution directly using Horner's sums. To compute recursively

$$z_k = \frac{1}{n} \cdot \sum_{m=0}^{n-1} x_m y_{k-m}, \qquad k = 0, \ldots, n-1,$$

we define the intermediate components

$$z_{k,p} = \sum_{m=0}^{p} x_m y_{k-m}, \qquad k = 0, \ldots, n-1,$$

for $p = 0, \ldots, n-1$. Theoretically,

$$z_{k,0} = x_0 y_k, \qquad z_{k,p+1} = \sum_{m=0}^{p+1} x_m y_{k-m}.$$

Hence, for $p = 0, \ldots, n-2$ and $k = 0, \ldots, n-1$,

$$z_{k,p+1} = z_{k,p} + x_{p+1} y_{k-p-1}. \tag{4}$$

In computation, if we denote by $\tilde{z}$ the floating-point representation of z and by $(x \mathrm{op} y)^{\Delta}$ the computed value of $x \mathrm{op} y$, we have

$$\tilde{z}_{k,0} = (x_0 \cdot y_k)^{\Delta}, \qquad \tilde{z}_{k,p+1} = (\tilde{z}_{k,p} + (x_{p+1} y_{k-p-1})^{\Delta})^{\Delta}.$$

Therefore, for each k and p,

$$\begin{aligned} \tilde{z}_{k,p+1} &= (\tilde{z}_{k,p} + x_{p+1} y_{k-p-1}(1 + \pi_{k,p+1}))(1 + \alpha_{k,p+1}) \\ &\doteq \tilde{z}_{k,p} \cdot (1 + \alpha_{k,p+1}) \\ &\quad + (x_{p+1} y_{k-p-1} + x_{p+1} y_{k-p-1} \pi_{k,p+1}) \cdot (1 + \alpha_{k,p+1}). \end{aligned} \tag{5}$$

If we let

$$\lambda_{k,p+1} = \tilde{z}_{k,p+1} - z_{k,p+1},$$

then from (5) we have, to first-order terms in the local errors,

$$\begin{aligned} z_{k,p+1} + \lambda_{k,p+1} \doteq\ & z_{k,p} + \lambda_{k,p} + z_{k,p} \alpha_{k,p+1} + x_{p+1} y_{k-p-1} \\ & + x_{p+1} y_{k-p-1} \alpha_{k,p+1} + x_{p+1} y_{k-p-1} \pi_{k,p+1}. \end{aligned}$$

From (4) we obtain the following recurrence relation for the accompanying linear forms:

$$\lambda_{k,p+1} = \lambda_{k,p} + x_{p+1} y_{k-p-1} \pi_{k,p+1} + z_{k,p+1} \alpha_{k,p+1}. \tag{6}$$

Since the linear forms in (6) can be decomposed as

$$\lambda_{k,p} = (\lambda a)_{k,p} + (\lambda m)_{k,p}$$

to account separately for the contributions from additions and multiplications, we have

$$(\lambda a)_{k,p+1} = (\lambda a)_{k,p} + z_{k,p+1} \alpha_{k,p+1} \tag{7}$$

and

$$(\lambda m)_{k,p+1} = (\lambda m)_{k,p} + x_{p+1} y_{k-p-1} \pi_{k,p+1}. \tag{8}$$

The initial conditions for the difference equations (7) and (8) are

$$(\lambda a)_{k,0} = 0 \quad \text{and} \quad (\lambda m)_{k,0} = x_0 y_k \pi_0.$$

It immediately follows that

$$(\lambda a)_{k,n-1} = \frac{1}{n}\sum_{i=1}^{n-1} z_{k,i}\alpha_{k,i}$$

and

$$(\lambda m)_{k,n-1} = \frac{1}{n}\sum_{q=0}^{n-1} x_q \cdot y_{k-q} \cdot \pi_{k,q}\,.$$

Under the assumption that the $\alpha_{k,i}$'s are i.i.d. with mean μ_a and variance σ_a^2, we derive the following bounds on the expected value of the accompanying linear form for addition:

$$\begin{aligned}|E((\lambda a)_{k,n-1})| &\le \frac{1}{n}\mu_a \sum_{i=1}^{n-1} z_{k,i} = \frac{1}{n}\mu_a \sum_{i=1}^{n-1}\left(\sum_{q=0}^{i} x_q y_{k-q}\right) \\ &\le \mu_a \frac{1}{2}\left(n+1-\frac{2}{n}\right)\|\mathbf{x}\|_\infty \|\mathbf{y}\|_\infty\,.\end{aligned} \tag{9}$$

Under the assumption that the $\pi_{k,i}$'s are i.i.d. with mean μ_m and variance σ_m^2, we have the following bounds for the expected value of $(\lambda m)_{k,n-1}$:

$$|E((\lambda m)_{k,n-1})| \le \frac{1}{n}\mu_m \sum_{q=0}^{n-1} |x_q y_{k-q}| \le \mu_m \|\mathbf{x}\|_\infty \|\mathbf{y}\|_\infty\,. \tag{10}$$

The bounds on the variance of $(\lambda a)_{k,n-1}$ and $(\lambda m)_{k,n-1}$ are as follows:

$$\operatorname{var}((\lambda a)_{k,n-1}) \le \frac{1}{n^2}\sigma_a^2 \sum_{i=1}^{n-1}\left|\sum_{q=0}^{i} x_q y_{k-q}\right|^2 \le \left(n-\frac{1}{n}\right)\sigma_a^2 \|\mathbf{x}\|_\infty^2 \|\mathbf{y}\|_\infty^2 \tag{11}$$

and

$$\operatorname{var}((\lambda m)_{k,n-1}) \le \frac{1}{n^2}\sigma_m^2 \sum_{q=1}^{n-1} |x_q y_{k-q}|^2 \le \sigma_m^2 \|\mathbf{x}\|_\infty^2 \|\mathbf{y}\|_\infty^2\,. \tag{12}$$

An alternative method to compute the convolution directly is to perform a cascade summation. Let

$$z_{0,j} = x_j y_{k-j}$$

for $j = 0, \ldots, n-1$, and

$$z_{q,2^q j} = z_{q-1,2^q j} + z_{q-1,2^q j+2^{q-1}}$$

for $q = 1, \ldots, l$ and $j = 0, \ldots, 2^{l-q}-1$. In computation,

$$\begin{aligned} z_{q,2^q j} + \lambda_{q,2^q j} &= [(z_{q-1,2^q j} + \lambda_{q-1,2^q j}) \\ &\quad + (z_{q-1,2^q j+2^{q-1}} + \lambda_{q-1,2^q j+2^{q-1}})](1+\alpha_{q,j}) \\ &\doteq z_{q,2^q j} + \lambda_{q-1,2^q j} + \lambda_{q-1,2^q j+2^{q-1}} + z_{q,2^q j}\alpha_{q,j}\,. \end{aligned}$$

Therefore, the accompanying linear forms $\lambda_{q,2^q j}$ must satisfy the difference equation

$$\lambda_{q,2^q j} = \lambda_{q-1,2^q j} + \lambda_{q-1,2^q j+2^{q-1}} + z_{q,2^q j}\alpha_{q,j} \tag{13}$$

subject to the initial condition $\lambda_{0,j} = x_j y_{k-j} \pi_j$. If we separate the contributions to global errors from the operations of addition and multiplication, we find that

$$(\lambda m)_{q,2^q j} = \frac{1}{n} \sum_{i=0}^{2^q-1} x_{2^q j+i} y_{k-(2^q j+i)} \pi_{2^q j+i}.$$

Therefore, the expected contribution to global roundoff error from multiplication in the kth entry of $\mathbf{x} * \mathbf{y}$ is

$$|E(\lambda m)| \le \frac{1}{n} \mu_m \sum_{i=0}^{2^q-1} |x_i y_{k-i}| \le \mu_m \|\mathbf{x}\|_\infty \|\mathbf{y}\|_\infty.$$

Consider now the contribution to global error from additions. From (13) it follows that

$$(\lambda a)_{q,2^q j} = \frac{1}{n} \sum_{i=1}^{q} \sum_{j=0}^{2^q-1} z_{i,2^q j+i} \alpha_{i,2^q j}.$$

Since

$$z_{i,2^q j} = \sum_{l=0}^{2^i-1} z_{0,2^q j+l},$$

we have

$$(\lambda a)_{q,2^q j} = \frac{1}{n} \sum_{i=1}^{q} \sum_{j=0}^{2^q-1} \left(\sum_{p=0}^{2^i-1} z_{0,2^q j+p} \right) \alpha_{i,2^q j}.$$

The expected value of the contribution to roundoff error from addition in each entry of the output is therefore bounded by

$$|E(\lambda a)| \le \frac{1}{n} \mu_a \sum_{i=1}^{q} \sum_{j=0}^{2^i-1} 2^i \|\mathbf{x}\|_\infty \|\mathbf{y}\|_\infty \le \log_2 n \mu_a \|\mathbf{x}\|_\infty \|\mathbf{y}\|_\infty.$$

Under the assumption that the relative errors introduced at each step are independent, we have the following bounds for the variance of the roundoff error in each entry of the output:

$$\operatorname{var}(\lambda a) = \frac{1}{n^2} \sigma_a^2 \sum_{k=1}^{l} \sum_{r=0}^{2^{l-k}-1} \left(\sum_{s=0}^{2^k-1} z_{0,2^k r+s} \right) \le \frac{1}{n} \log_2 n \sigma_a^2 \|\mathbf{x}\|_\infty^2 \|\mathbf{y}\|_\infty^2$$

and

$$\operatorname{var}(\lambda m) = \frac{1}{n^2} \sigma_m^2 \sum_{k=1}^{l} \sum_{r=0}^{2^{l-k}-1} \left(\sum_{s=0}^{2^k-1} z_{0,2^k(2r+1)+s} \right) \le \frac{n+1}{n^2} \sigma_m^2 \|\mathbf{x}\|_\infty^2 \|\mathbf{y}\|_\infty^2.$$

We now turn our attention to the accompanying linear forms for the algorithm that computes the convolution via Radix-2 FFT's. In computation,

$$\begin{aligned} \tilde{\mathbf{z}} &= \frac{1}{n} [[\overline{F}_n ((F_n \mathbf{x})^\Delta \square (F_n \mathbf{y})^\Delta)]^\Delta]^\Delta \\ &\doteq \frac{1}{n} [\overline{F}_n [(F_n \mathbf{x})^\Delta \square (F_n \mathbf{y})^\Delta \square (\mathbf{1} + \boldsymbol{\pi}_n)]]^\Delta. \end{aligned}$$

Since $(F_n\mathbf{x})^{\Delta} \doteq F_n\mathbf{x} + {}_F\boldsymbol{\lambda}_x$ and $(F_n\mathbf{y})^{\Delta} \doteq F_n\mathbf{y} + {}_F\boldsymbol{\lambda}_y$, where ${}_F\boldsymbol{\lambda}_w$ is the accompanying linear form for the Radix-2 FFT of $\mathbf{w}$, and

$$(\overline{F}_n(\mathbf{w}))^{\Delta} \doteq \overline{F}_n(\mathbf{w}) + {}_{\overline{F}}\boldsymbol{\lambda}_w\,,$$

where ${}_{\overline{F}}\boldsymbol{\lambda}_w$ is the accompanying linear form for the inverse Radix-2 FFT of $\mathbf{w}$, it follows that

$$\begin{aligned}\tilde{\mathbf{z}} &\doteq \frac{1}{n}\cdot\overline{F}_n[(\hat{\mathbf{x}} + {}_F\boldsymbol{\lambda}_x)\,\square\,(\hat{\mathbf{y}} + {}_F\boldsymbol{\lambda}_y)\,\square\,(\mathbf{1} + \boldsymbol{\pi}_n)] + \frac{1}{n}\cdot{}_{\overline{F}}\boldsymbol{\lambda}_{(\hat{\mathbf{x}}\,\square\,\hat{\mathbf{y}})}\\ &= \frac{1}{n}\cdot\overline{F}_n(\hat{\mathbf{x}}\,\square\,\hat{\mathbf{y}}) + \frac{1}{n}\cdot\overline{F}_n[(\hat{\mathbf{x}}\,\square\,{}_F\boldsymbol{\lambda}_y + \hat{\mathbf{y}}\,\square\,{}_F\boldsymbol{\lambda}_x + \hat{\mathbf{x}}\,\square\,\hat{\mathbf{y}}\,\square\,\boldsymbol{\pi}_n)] + \frac{1}{n}\cdot{}_{\overline{F}}\boldsymbol{\lambda}_{(\hat{\mathbf{x}}\,\square\,\hat{\mathbf{y}})}\end{aligned}$$

by the linearity of the FFT. Therefore, since the accompanying linear forms for the convolution via Radix-2 FFT are defined by

$${}_c\boldsymbol{\lambda}_z = \tilde{\mathbf{z}} - \mathbf{z}\,,$$

it follows that

$${}_c\boldsymbol{\lambda}_z = \frac{1}{n}\cdot\overline{F}_n[(\hat{\mathbf{x}}\,\square\,{}_F\boldsymbol{\lambda}_y + \hat{\mathbf{y}}\,\square\,{}_F\boldsymbol{\lambda}_x + \hat{\mathbf{x}}\,\square\,\hat{\mathbf{y}}\,\square\,\boldsymbol{\pi}_n)] + \frac{1}{n}\cdot{}_{\overline{F}}\boldsymbol{\lambda}_{(\hat{\mathbf{x}}\,\square\,\hat{\mathbf{y}})}\,.$$

Since the accompanying linear forms can be decomposed as

$$\boldsymbol{\lambda} = \boldsymbol{\lambda}\mathbf{a} + \boldsymbol{\lambda}\mathbf{m}\,,$$

we have that

$$(14)\qquad {}_c(\boldsymbol{\lambda}\mathbf{a})_z = \frac{1}{n}\cdot\overline{F}_n[\hat{\mathbf{x}}\,\square\,{}_F(\boldsymbol{\lambda}\mathbf{a})_y + \hat{\mathbf{y}}\,\square\,{}_F(\boldsymbol{\lambda}\mathbf{a})_x] + \frac{1}{n}\cdot{}_{\overline{F}}\,(\boldsymbol{\lambda}\mathbf{a})_{(\hat{\mathbf{x}}\,\square\,\hat{\mathbf{y}})}$$

and

$$(15)\quad {}_c(\boldsymbol{\lambda}\mathbf{m})_z = \frac{1}{n}\cdot\overline{F}_n[(\hat{\mathbf{x}}\,\square\,{}_F(\boldsymbol{\lambda}\mathbf{m})_y + \hat{\mathbf{y}}\,\square\,{}_F(\boldsymbol{\lambda}\mathbf{m})_x + \hat{\mathbf{x}}\,\square\,\hat{\mathbf{y}}\,\square\,\boldsymbol{\pi}_n)] + \frac{1}{n}\cdot{}_{\overline{F}}(\boldsymbol{\lambda}\mathbf{m})_{(\hat{\mathbf{x}}\,\square\,\hat{\mathbf{y}})}\,.$$

The derivation of the accompanying linear forms for the Radix-2 FFT with scalar factor $\frac{1}{n}$ instead of $\frac{1}{\sqrt{n}}$ can be found in [2]. Although the derivation of λa and λm is not affected by rescaling, the infinity norm of the intermediate results changes, therefore the estimates for the expected value and variance of the global error change. For each $k = 1, \ldots, l$ we have, with the scalar $\frac{1}{\sqrt{n}}$,

$$\|\mathbf{z}_k\|_\infty \le \sqrt{2}\|\mathbf{z}_{k-1}\|_\infty\,,$$

where $\mathbf{z}_k$ is the intermediate result at the kth step. From [2] it follows that

$${}_F\lambda a_l = \sum_{k=1}^{l-2} W_l\cdots W_{k+1}(\mathbf{z}_k\,\square\,\boldsymbol{\alpha}_k) + W_l(\mathbf{z}_{l-1}\,\square\,\boldsymbol{\alpha}_{l-1}) + (\mathbf{z}_l\,\square\,\boldsymbol{\alpha}_l)\,.$$

Since each element of the matrix W has exactly two nonzero entries, each one of absolute value $\frac{1}{\sqrt{2}}$, for each entry of $|E({}_F\lambda a_l)|$ we have

$$\begin{aligned}(16)\qquad |E({}_F\boldsymbol{\lambda}\mathbf{a}_l)| &\le \left(\sum_{k=1}^{l-2} 2^{(l-k)/2}\cdot 2^{(k+1)/2} + \sqrt{2}\cdot 2^{(l-1)/2} + 2^{l/2}\right)\mu_a\|\mathbf{x}\|_\infty\\ &= ((2-2\sqrt{2})\sqrt{n} + \sqrt{2n}\cdot\log_2 n)\mu_a\|\mathbf{x}\|_\infty\,.\end{aligned}$$

Similarly, from a simple modification of the results of [2] it follows that

$$
\begin{aligned}
{}_F\boldsymbol{\lambda}\mathbf{m}_l = &\sum_{k=1}^{l-2} W_l \cdots W_{k+1}(\mathbf{b}_k \square \mathbf{z}_{k-1} \square \boldsymbol{\pi}_k) \\
&+ W_l(\mathbf{b}_{l-1} \square \mathbf{z}_{l-2} \square \boldsymbol{\pi}_{l-1}) + (\mathbf{b}_l \square \mathbf{z}_{l-1} \square \boldsymbol{\pi}_l).
\end{aligned}
$$

Therefore, for each entry of $|E({}_F\boldsymbol{\lambda}\mathbf{m}_l)|$ we have

$$
\begin{aligned}
|E({}_F\lambda m_l)| &\le \frac{1}{2}\left(\sum_{k=0}^{l-3} 2^{(l-k)/2} \cdot 2^{k/2} + 2^{1/2} \cdot 2^{(l-2)/2} + 2^{(l-1)/2}\right)\mu_m \|\mathbf{x}\|_\infty \\
&= \frac{1}{2}\left(\sqrt{n} \cdot (\log_2 n - 2 + \sqrt{2})\right)\mu_m \|\mathbf{x}\|_\infty .
\end{aligned} \tag{17}
$$

The bounds for the entries of the covariance matrix of the global error for the particular Radix-2 FFT considered here can be found by utilizing the results of [2]. In particular, since

$$
\operatorname{cov}{}_F\boldsymbol{\lambda}\mathbf{a}_l = \sigma_a^2\left(\sum_{k=1}^{l-2} W_l \cdots W_{k+1} Z_k^2 W_{k+1}' \cdots W_l' + W_l Z_{l-1}^2 W_l' + Z_l^2\right),
$$

where Z_j^2 is the $n \times n$ diagonal matrix such that

$$
Z_j^2(d, d) = (z_j(d))^2,
$$

we have for each entry of the matrix $\operatorname{cov}{}_F\lambda a_l$

$$
\begin{aligned}
&\operatorname{cov}({}_F\lambda a_l)_{ij} \\
&\le \left(\sum_{k=1}^{l-2} 2^{(l-k)/2} \cdot 2^{2\cdot(k+1)/2} \cdot 2^{(l-k)/2} + 2^{1/2} \cdot 2^{(l-1)/2} \cdot 2^{1/2} + 2^l\right)\sigma_a^2 \|\mathbf{x}\|_\infty^2 \\
&= (2n \log_2 n - 3n + n^{1/2})\sigma_a^2 \|\mathbf{x}\|_\infty^2 .
\end{aligned} \tag{18}
$$

Similarly, since

$$
\operatorname{cov}{}_F\boldsymbol{\lambda}\mathbf{m}_l = \frac{1}{4}\sigma_m^2\left(\sum_{k=1}^{l-2} W_l \cdots W_{k+1} Z_{k-1}^2 W_{k+1}' \cdots W_l' + W_l Z_{l-2}^2 W_l' + Z_{l-1}^2\right),
$$

it follows that, for each entry of the matrix $\operatorname{cov}({}_F\lambda m_l)$,

$$
|\operatorname{cov}({}_F\lambda m_l)|_{ij} \le \tfrac{1}{4} \cdot (2n \log_2 n - 3n + \sqrt{2n})\sigma_m^2 \|\mathbf{x}\|_\infty^2 . \tag{19}
$$

We now compute the bounds for the expected value and variance of the accompanying linear forms for the circular convolution. From (14) it follows that, for each entry of $E({}_c(\boldsymbol{\lambda}\mathbf{a})_z)$, we have

$$
\begin{aligned}
|E({}_c(\boldsymbol{\lambda}\mathbf{a})_z)| &\le \frac{1}{n}[2\sqrt{n}(\sqrt{2n} \cdot \log_2 n + (2 - 2\sqrt{2})\sqrt{n})]\mu_a \|\mathbf{x}\|_\infty \|\mathbf{y}\|_\infty \\
&\quad + \frac{1}{n}[\sqrt{2n} \cdot \log_2 n + (2 - 2\sqrt{2})\sqrt{n}]n\mu_a \|\mathbf{x}\|_\infty \|\mathbf{y}\|_\infty \\
&= 3[\sqrt{2n}(\log_2 n + (2 - 2\sqrt{2})\sqrt{n})]\mu_a \|\mathbf{x}\|_\infty \|\mathbf{y}\|_\infty .
\end{aligned} \tag{20}
$$

Similarly, from (15) it follows that, for each element of $E({}_c(\boldsymbol{\lambda}\mathbf{m})_z)$, we have

$$
|E({}_c(\boldsymbol{\lambda}\mathbf{m})_z)| \le \sqrt{n}[\tfrac{3}{2}(\log_2 n + \sqrt{2} - 2) + 1]\mu_m \|\mathbf{x}\|_\infty \|\mathbf{y}\|_\infty . \tag{21}
$$

In order to find bounds for the variance of the components of the vectors ${}_c(\boldsymbol{\lambda}\mathbf{a})_z$ and ${}_c(\boldsymbol{\lambda}\mathbf{m})_z$, that is, for the diagonal entries of their covariance matrices, we express the accompanying linear forms in matrix notation:

$$
{}_c(\boldsymbol{\lambda}\mathbf{a})_z = \frac{1}{n}(\overline{F}_n \quad I_n)\begin{pmatrix} D\hat{x} & D\hat{y} & 0 \\ 0 & 0 & I_n \end{pmatrix}\begin{pmatrix} {}_F(\boldsymbol{\lambda}\mathbf{a})_y \\ {}_F(\boldsymbol{\lambda}\mathbf{a})_x \\ \overline{F}(\boldsymbol{\lambda}\mathbf{a})_{\hat{x}\,\square\,\hat{y}} \end{pmatrix},
$$

$$
{}_c(\boldsymbol{\lambda}\mathbf{m})_z = \frac{1}{n}(\overline{F}_n \quad I_n)\begin{pmatrix} D\hat{x} & D\hat{y} & D(\hat{x}\,\square\,\hat{y}) & 0 \\ 0 & 0 & 0 & I_n \end{pmatrix}\begin{bmatrix} {}_F(\boldsymbol{\lambda}\mathbf{m})_y \\ {}_F(\boldsymbol{\lambda}\mathbf{m})_x \\ \boldsymbol{\pi}_n \\ \overline{F}(\boldsymbol{\lambda}\mathbf{m})_{\hat{x}\,\square\,\hat{y}} \end{bmatrix},
$$

where Dw is the $n \times n$ diagonal matrix with the entries of the vector $\mathbf{w} = \hat{\mathbf{x}}, \hat{\mathbf{y}}, \hat{\mathbf{x}}\,\square\,\hat{\mathbf{y}}$ on the main diagonal. Then

$$
\operatorname{cov}({}_c(\boldsymbol{\lambda}\mathbf{a})_z) = \frac{1}{n^2}(\overline{F}_n \quad I_n)\begin{pmatrix} D\hat{x} & D\hat{y} & 0 \\ 0 & 0 & I_n \end{pmatrix} \times \operatorname{cov}\begin{bmatrix} {}_F(\boldsymbol{\lambda}\mathbf{a})_y \\ {}_F(\boldsymbol{\lambda}\mathbf{a})_x \\ \overline{F}(\boldsymbol{\lambda}\mathbf{a})_{\hat{x}\,\square\,\hat{y}} \end{bmatrix}\begin{bmatrix} D\hat{x} & 0 \\ D\hat{y} & 0 \\ 0 & I_n \end{bmatrix}\begin{bmatrix} \overline{F}'_n \\ I_n \end{bmatrix}
$$

and

$$
\operatorname{cov}({}_c(\boldsymbol{\lambda}\mathbf{m})_z) = \frac{1}{n^2}(\overline{F}_n \quad I_n)\begin{pmatrix} D\hat{x} & D\hat{y} & D\hat{x}\,\square\,\hat{y} & 0 \\ 0 & 0 & 0 & I_n \end{pmatrix} \times \begin{bmatrix} {}_F(\boldsymbol{\lambda}\mathbf{m})_y \\ {}_F(\boldsymbol{\lambda}\mathbf{m})_x \\ \boldsymbol{\pi}_n \\ \overline{F}(\boldsymbol{\lambda}\mathbf{m})_{\hat{x}\,\square\,\hat{y}} \end{bmatrix}\begin{bmatrix} D\hat{x} & 0 \\ D\hat{y} & 0 \\ D(\hat{x}\,\square\,\hat{y}) & 0 \\ 0 & I_n \end{bmatrix}\begin{bmatrix} \overline{F}'_n \\ I_n \end{bmatrix}.
$$

Under the assumption that the components of the vectors of the local relative errors coming from addition and multiplication at each step of the computation of $\hat{\mathbf{x}}$, $\hat{\mathbf{y}}$, and $\hat{\mathbf{x}}\,\square\,\hat{\mathbf{y}}$ are independent, the matrices

$$
\operatorname{cov}\begin{bmatrix} {}_F(\boldsymbol{\lambda}\mathbf{a})_y \\ {}_F(\boldsymbol{\lambda}\mathbf{a})_x \\ \overline{F}(\boldsymbol{\lambda}\mathbf{a})_{\hat{x}\,\square\,\hat{y}} \end{bmatrix} \quad \text{and} \quad \operatorname{cov}\begin{bmatrix} {}_F(\boldsymbol{\lambda}\mathbf{m})_y \\ {}_F(\boldsymbol{\lambda}\mathbf{m})_x \\ \boldsymbol{\pi}_n \\ \overline{F}(\boldsymbol{\lambda}\mathbf{m})_{\hat{x}\,\square\,\hat{y}} \end{bmatrix}
$$

are block diagonal of the form

$$
\operatorname{cov}\begin{bmatrix} {}_F(\boldsymbol{\lambda}\mathbf{a})_y \\ {}_F(\boldsymbol{\lambda}\mathbf{a})_x \\ \overline{F}(\boldsymbol{\lambda}\mathbf{a})_{\hat{x}\,\square\,\hat{y}} \end{bmatrix} = \begin{bmatrix} \operatorname{cov}({}_F(\boldsymbol{\lambda}\mathbf{a})_y) & & \\ & \operatorname{cov}({}_F(\boldsymbol{\lambda}\mathbf{a})_x) & \\ & & \operatorname{cov}(\overline{F}(\boldsymbol{\lambda}\mathbf{a})_{\hat{x}\,\square\,\hat{y}}) \end{bmatrix}
$$

and

$$
\operatorname{cov}\begin{bmatrix} {}_F(\boldsymbol{\lambda}\mathbf{m})_y \\ {}_F(\boldsymbol{\lambda}\mathbf{m})_x \\ \boldsymbol{\pi}_n \\ \overline{F}(\boldsymbol{\lambda}\mathbf{m})_{\hat{x}\,\square\,\hat{y}} \end{bmatrix} = \begin{bmatrix} \operatorname{cov}({}_F(\boldsymbol{\lambda}\mathbf{m})_y) & & & \\ & \operatorname{cov}({}_F(\boldsymbol{\lambda}\mathbf{m})_x) & & \\ & & \sigma_m^2 I_n & \\ & & & \operatorname{cov}(\overline{F}(\boldsymbol{\lambda}\mathbf{m})_{\hat{x}\,\square\,\hat{y}}) \end{bmatrix},
$$

respectively. Therefore,

$$
\begin{aligned}
\operatorname{cov}({}_c(\boldsymbol{\lambda}\mathbf{a})_z) = \frac{1}{n^2}\{&\overline{F}_n D\hat{x}\operatorname{cov}({}_F(\boldsymbol{\lambda}\mathbf{a})_y)D\hat{x}\overline{F}'_n \\ &+ \overline{F}_n D\hat{y}\operatorname{cov}({}_F(\boldsymbol{\lambda}\mathbf{a})_x)D\hat{y}\overline{F}'_n + \operatorname{cov}(\overline{F}(\boldsymbol{\lambda}\mathbf{a})_{\hat{x}\,\square\,\hat{y}})\}
\end{aligned} \tag{22}
$$

and

$$\text{cov}({}_c(\boldsymbol{\lambda}\mathbf{m})_z) = \frac{1}{n^2}\{\overline{F}_n D\hat{x}\,\text{cov}({}_F(\boldsymbol{\lambda}\mathbf{m})_y) D\hat{x}\overline{F}_n' + \overline{F}_n D\hat{y}\,\text{cov}({}_F(\boldsymbol{\lambda}\mathbf{m})_x) D\hat{y}\overline{F}_n' + \sigma_m^2(\overline{F}_n D(\hat{x}\,\square\,\hat{y}) D(\hat{x}\,\square\,\hat{y})\overline{F}_n') + \text{cov}(\overline{{}_F(\boldsymbol{\lambda}\mathbf{m})}_{\hat{x}\,\square\,\hat{y}})\}. \tag{23}$$

Since each entry of $\text{cov}({}_F\lambda a_x)$ is bounded in absolute value by

$$2n(\log_2 n - 3n + \sqrt{n})\sigma_a^2\|\mathbf{x}\|_\infty^2,$$

we have for each entry of the matrix $\text{cov}({}_c(\boldsymbol{\lambda}\mathbf{a})_z)$

$$|\text{cov}({}_c(\boldsymbol{\lambda}\mathbf{a})_z)_{ij}| \le 3(2n\log_2 n - 3n + \sqrt{n})\sigma_a^2\|\mathbf{x}\|_\infty^2\|\mathbf{y}\|_\infty^2. \tag{24}$$

Similarly, from (20) it follows that, for each entry of $\text{cov}({}_c(\boldsymbol{\lambda}\mathbf{m})_z)$, we have

$$\begin{aligned} |\text{cov}({}_c(\boldsymbol{\lambda}\mathbf{m})_z)_{ij}| &\le \frac{1}{n^2}\left[\frac{1}{4}\cdot 2(2n\log_2 n - 3n + \sqrt{2n})n^2\right]\sigma_m^2\|\mathbf{x}\|_\infty^2\|\mathbf{y}\|_\infty^2 \\ &\quad + \frac{1}{n^2}n^2\sigma_m^2\|\mathbf{x}\|_\infty^2\|\mathbf{y}\|_\infty^2 \\ &\quad + \frac{1}{n^2}\left[\frac{1}{4}(2n\log_2 n - 3n + \sqrt{2n})\right]\sigma_m^2\|\hat{\mathbf{x}}\,\square\,\hat{\mathbf{y}}\|_\infty^2 \\ &\le \left\{\frac{3}{4}(2n\log_2 n - 3n + \sqrt{2n}) + 1\right\}\sigma_m^2\|\mathbf{x}\|_\infty^2\|\mathbf{y}\|_\infty^2. \end{aligned} \tag{25}$$

5. Numerical experiments

In order to test numerically the consistency of the results of our roundoff error analysis for the circular convolution using a unitary Radix-2 FFT with the errors actually observed in computations, we generated, for each $l = 5, 6, \ldots, 10$, a total of 3000 pairs of n-dimensional vectors, with $n = 2^l$, with the real and imaginary parts independent random variables from the uniform $[0, 1]$ distribution. Each component was then rounded to eight significant digits and the circular convolution was computed twice, once with the results of each operation rounded to eight digits, and once with all operations performed in double precision. The output vectors were compared componentwise, and the absolute error in each component was calculated. The infinity norms of the sample mean and sample variance for the different values of n are listed in Table 1 (see next page).

The l_∞ norm of the mean and variance of the global error is a somewhat conservative measure of the error affecting the output, in the sense that it measures the largest sample mean and sample variance of the global error affecting each individual component of the output. In order to provide, in addition, an overall measure of the mean and variance of the global error in the output, we list in Table 2 (see next page) the l_2 norms of the mean and variance of the global error. Notice that the l_2 norms are much smaller than the corresponding l_∞ norms, because instead of looking for the largest entry we average all entries over the full vector.

The slow growth of the infinity norm of the variance and expected value of the global error predicted by the analysis are confirmed by the numerical experiment, at least when the data has real and imaginary parts uniformly distributed

TABLE 1. l_∞ norm of the mean and variance of the global error for circular convolution. Sample size = 3000.

n	mean	variance
32	9.17E-8	8.78E-17
64	8.60E-8	5.00E-17
128	1.02E-7	1.78E-17
256	2.92E-7	1.49E-17
512	5.48E-7	2.31E-16
1024	4.05E-7	1.48E-16

TABLE 2. l_2 norm of the mean and variance of the global error for circular convolution. Sample size = 3000.

n	mean	variance
32	1.68E-8	1.88E-17
64	1.13E-8	1.48E-17
128	9.27E-9	1.92E-18
256	2.01E-8	1.30E-17
512	3.79E-8	1.23E-17
1024	1.33E-8	2.54E-18

in the interval [0, 1]. More specifically, the ratio of the bounds for the expected error for two successive values of l is of the order of $\sqrt{2}(l+1)/l$, not far from the ratios of the corresponding sample means. The overall trend observed in the numerical experiments is that of a slow growth of both mean and variance of the global error. The occasional reduction of either parameter as l increases can be attributed to the characteristics of the particular sample selected.

Theoretical and numerical estimates of the expected value for the relative error for addition and multiplication of floating-point numbers uniformly distributed in [0, 1] suggested that their values are approximately b^{-t-1} and b^{-2t}, respectively. In view of this observation, the value of the sample mean for the error in circular convolution at $n = 32$ is very close to the theoretical bound, which is approximately $1.02\mu_a\|\mathbf{x}\|_\infty\|\mathbf{y}\|_\infty$.

In order to compare the accuracy of the circular convolution with the accuracy of an algorithm which evaluates the convolution directly, we computed the convolution of the same pairs of vectors used to test the accuracy of circular

TABLE 3. l_∞ norm of the mean and variance of the global error for direct convolution. Sample size = 3000.

n	mean	variance
32	4.67E-8	2.85E-17
64	6.94E-8	4.66E-17
128	2.81E-7	1.77E-16
256	4.36E-7	2.86E-17
512	1.45E-6	2.27E-15

TABLE 4. l_2 norm of the mean and variance of the global error for direct convolution. Sample size = 3000.

n	mean	variance
32	8.97E-9	7.25E-18
64	9.35E-9	1.06E-17
128	2.69E-8	1.98E-17
256	2.98E-8	2.31E-17
512	3.69E-7	1.29E-16

convolution by means of a cascade summation. The l_∞ and l_2 norms of the sample mean and sample variance of the global error, with $l = 5, 6, 7, 8, 9$, are listed in Tables 3 and 4.

On first sight the entries in Table 3 seem to disagree with the theoretical bound for the expected error, which is approximately $\log_2 n\mu_a\|\mathbf{x}\|_\infty\|\mathbf{y}\|_\infty$, since the l_∞ norm of the sample mean grows at a rate larger than $(l+1)/l$. The value of μ_a, however, is a function of the numbers being added together, hence as the size of the vector increases the interval containing all terms added together becomes larger, thus yielding a larger value for μ_a.

A comparison of the values of the expected value and variance of the global error listed in Tables 1 and 3 suggests that the faster algorithm is as accurate as the one computing the convolution via cascade summations. We conclude by pointing out that the amount of work required by the two algorithms, $O(n^2)$ flops for the direct calculation versus $O(n \log_2 n)$ for the circular convolution, together with the results of our error analysis point to the circular convolution using a unitary Radix-2 FFT as the algorithm of choice.

ACKNOWLEDGMENTS

We thank Jon Tolle for his valuable suggestions and the referee and the editor for their useful comments which significantly improved the presentation of the paper.

BIBLIOGRAPHY

1. P. Bloomfield, *Fourier analysis of time series*, Wiley, New York, 1976.
2. D. Calvetti, *A stochastic roundoff error analysis for the fast Fourier transform*, Math. Comp. **56** (1991), 755–774.
3. J. W. R. Cooley and J. W. Tukey, *An algorithm for the machine calculation of complex Fourier series*, Math. Comp. **19** (1965), 297–301.
4. P. Henrici, *Error propagation for difference methods*, Wiley, New York, 1963.
5. ____, *Essentials of numerical analysis*, Wiley, New York, 1982.
6. ____, *Applied and computational complex analysis*, Vol. 3, Wiley, New York, 1986.
7. T. Kaneko and B. Liu, *Accumulation of roundoff error in fast Fourier transforms*, J. Assoc. Comput. Mach. **17** (1970), 637–654.
8. A. Oppenheim and C. Weinstein, *A bound on the output of a circular convolution with application to digital filtering*, Trans. Audio Electroacoust. **17** (1969), 120–124.
9. G. U. Ramos, *Roundoff error analysis of the fast Fourier transform*, Math. Comp. **25** (1971), 757–768.
10. F. Stummel, *Perturbation theory for evaluation algorithms of arithmetic expressions*, Math. Comp. **37** (1981), 435–473.
11. F. Stummel and K. Heiner, *Praktische Mathematik*, 2nd ed., Teubner, Stuttgart, 1982.
12. C. J. Weinstein, *Roundoff noise in floating point fast Fourier transform computation*, IEEE Trans. Audio Electroacoust. **17** (1969), 209–215.
13. P. D. Welsh, *A fixed-point fast Fourier transform error analysis*, IEEE Trans. Audio Electroacoust. **17** (1969), 151–157.
14. J. H. Wilkinson, *Rounding errors in algebraic processes*, Prentice-Hall, Englewood Cliffs, NJ, 1963.

DEPARTMENT OF PURE AND APPLIED MATHEMATICS, STEVENS INSTITUTE OF TECHNOLOGY, CASTLE POINT ON THE HUDSON, HOBOKEN, NEW JERSEY 07030
E-mail address: roe_dcalvetti@sitvax.stevens-tech.edu