Annales Univ. Sci. Budapest., Sect. Comp. 49 (2019) 259–277

ERROR ESTIMATES FOR MULTIPLICATION BASED ON FFT OVER THE COMPLEX NUMBERS

Antal Járai (Budapest, Hungary)

Communicated by Imre Kátai
(Received February 27, 2019; accepted May 12, 2019)

Abstract. We show that multiplication based on complex FFT is exact if usual rounding is used in a k-round FFT ($k \geq 2$) with floating point numbers having m-bit mantissa and we put ℓ bit digits into a floating point number, whenever the inequality

$$8.074(k-2) + 10.978 < 2^{m-2\ell-2k}$$

is satisfied.

1. Introduction

1.1. Fast multiplication. Fast multiplication of large numbers has a central role in computer algebra, primality testing, encryption, etc. Some operations reduced to fast multiplication:

-Reciprocal

-Quotient

-Logarithm: the series case

-Exponential: the series case

-Power: the series case

-Quotient and remainder

-Continued fraction from fraction

-Remainder tree

-Interpolation

Key words and phrases: Fast Fourier transform, multiplication of large numbers, error estimate, rounding error, IEEE754.
2010 Mathematics Subject Classification: 65G50, 65T50.

- -GCD
- -Coprime base
- -Matrix product
- -Product tree
- -Exponential: modular case
- -Exponential: general case
- -Small factors of a product
- -Fraction from continued fraction
- -Polynomial multiplication
- -Polynomial division
- -Polynomial GCD
- -Polynomial factorization
- -Multivariate polynomial operations
- -etc.

1.2. Multiplication with FFT. As was discovered by V. Strassen, fast multiplication of large numbers can be based on discrete Fourier transform, shortly DFT of a *real* sequence $f_0, f_1, \ldots, f_{2n-1}$ with $2n$ terms. For a detailed treatment of this, see Knuth [2], 4.3.3. By definition, the discrete Fourier transform is

$$\hat{f}_r = \sum_{s=0}^{2n-1} f_j \omega^{-rs}$$

for $j = 0, 1, \ldots, 2n-1$, where $\omega = e^{-\pi i/n} = e^{-2\pi i/(2n)}$ is a $2n$th root of unity.

If we take another sequence $g_0, g_1, \ldots, g_{2n-1}$, and choose $f_r = g_r = 0$ for $n \le r < 2n$, we may use the sequences $\hat{f}_r$ and $\hat{g}_r$ to compute the numbers

$$h_r = \sum_{s=0}^{r} f_s g_{r-s}$$

for $r = 0, 1, \ldots, 2n-2$. These numbers are the coefficients of the polynomial $\sum_{s=0}^{2n-2} h_s x^s$, and this polynomial is the product of the polynomials $\sum_{s=0}^{n-1} f_s x^s$ and $\sum_{s=0}^{n-1} g_s x^s$. They are hence very useful if we want to compute the product of two long numbers represented by $f_0, f_1, \ldots, f_{n-1}$ and $g_0, g_1, \ldots, g_{n-1}$ as digits in a number system with an arbitrary base x. Because the sequence h_r is the convolution of the sequences f_r and g_r, we may compute h_r as the inverse discrete Fourier transform of $\hat{h}_r = \hat{f}_r \hat{g}_r$.

Because f_r is real, we obtain

$$\overline{\hat{f}}_{2n-r} = \sum_{s=0}^{2n-1} f_s \omega^{(2n-r)s} = \hat{f}_r,$$

where indices are understood modulo $2n$. This means that half of the results is superfluous. This gives the idea to reduce the computation of the n independent complex $\hat{f}_r$ to the computation of the discrete Fourier transform of the complex sequence $F_r = f_{2r} + if_{2r+1}$, $r = 0, 1, \ldots, n-1$, because

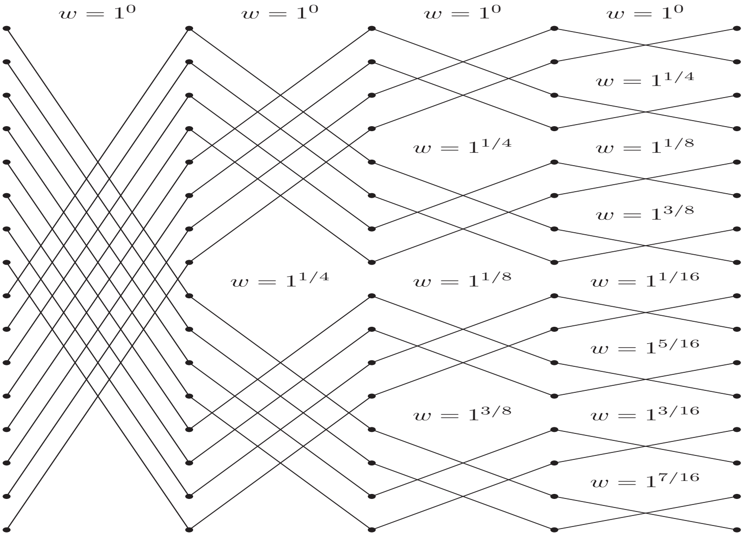
$$(\hat{F}_r + \overline{\hat{F}_{n-r}}) - i\omega^r(\hat{F}_r - \overline{\hat{F}_{n-r}}) = 2\hat{f}_r,$$

and

$$(\hat{h}_r + \overline{\hat{h}_{n-r}}) + i\omega^{-r}(\hat{h}_r - \overline{\hat{h}_{n-r}}) = 2\hat{H}_r;$$

indices understood modulo n.

If $n = 2^k$ is a two-power, the calculation of the discrete Fourier transform and its inverse may be done with $5n\log_2 n$ (real) operations using the *Fast Fourier Transform*, shortly FFT, algorithm. For a detailed treatment of the FFT see Knuth [1]. It is done in rounds. During one round, a sequence of butterfly operations is done in place. The inverse FFT is calculated by doing reverse butterfly operation in the reverse order of rounds. See the figure. For a detailed treatment of the ideas of this paragraph see Járai and Járai [1].



FFT data movement

$$\begin{aligned} \text{in } \mathbb{C}: &\quad 1^\alpha = e^{-2\pi i\alpha}; \\ \text{butterfly}: &\quad (x,y) \leftarrow (x+wy, x-wy); \\ \text{inverse butterfly}: &\quad (x,y) \leftarrow \big((x+y)w, (x-y)w\big). \end{aligned}$$

1.3. Example. Let us consider IEEE 754 standard double precision floating point arithmetic where the number of bits in the mantissa is $m = 52$. Let us divide a number with less then $22 \cdot 2^{14}$ binary digit to 2^{14} parts (i. e., $k = 14$ and $n = 2^{14}$) each containing 22 bits. Complex Fourier transforms are calculated for a complex vector with 2^{14} terms. The terms of the resulting vector contain the sum of at most 2^{14} products each 44 bits at most. Hence roughly 58 bit of the resulting terms should have to be right. The situation is much better if we use 2^{22} as the base of the number system (i. e., $\ell = 22$) but signed digits between (inclusively) -2^{21} and 2^{21}. These signed digits can get easily. First set a carry to 0. Then cut 22 bits from the end, sign-extend it to a full word, store its sign bit as the new carry and add the old carry; repeat this step until all digits are obtained. In this case we have the sum of 2^{14} *signed* digits between -2^{42} and 2^{42}. With some probability we can reconstruct the product without error. Of course, in general $\ell \geq 2$ have to be satisfied.

1.4. Error estimate. To be sure that the result calculated as in the example is right we want to find sufficient condition under which multiplication based on complex FFT is exact after rounding if k-round FFT is used with floating point numbers having m-bit mantissa and we put a ℓ bits into a digit.

The only such condition which I know is the condition

$$(k-2) + 5.5 < 2^{m-2\ell-3k}$$

from Knuth [2], 4.3.3.(C). He use truncation toward zero.

We shall prove that using standard IEEE 754 floating point arithmetic with usual rounding,

$$8.074(k-2) + 10.978 < 2^{m-2\ell-2k}$$

is sufficient, if $k \geq 2$. This is a weaker condition as Knuth's for all $k > 2$.

2. Error estimates for elementary steps

During all the error estimates let a, b, c, d, ... complex numbers with approximate values a', b', c', d', ... having errors at most ε_a, ε_b, ε_c, ε_d, i.e., we have $|a - a'| \leq \varepsilon_a$. etc. We suppose that all the numbers has absolute value at most $B2^e$, where $1 \leq B \ll 2$. In practice, the bound B is usually $\sqrt{2}$ or 1. Now we *suppose* this but during the final estimate we will *prove* this.

All over the estimates we suppose that floating point numbers are represented using m bit mantissa. For example, in the IEEE 754 standard for simple precision numbers $m = 23$, for double precision numbers $m = 52$, for quadruple precision numbers $m = 112$. For a rounded value r' of the exact value r of a floating point machine operation we have that $|r - r'| \leq 2^{e_r-(m-1)}$, if we know that $|r|, |r'| \ll 2^{e_r+1} = 2 \cdot 2^{e_r}$; here $\ll 2 \cdot 2^{e_r}$ simply means that if the exponent is e_r, then the mantissa is far from 2, for example around $\sqrt{2}$ or less. For the approximate values we also suppose that they have absolute value at most $B'2^e$, where $1 \leq B \leq B' \ll 2$. (This can be proved by induction; see later.) We suppose an IEEE standard arithmetic, which gives back the exact rounded value of the exact result of the operation.

2.1. Error estimate for addition and subtraction. If $|a|, |b| \leq B2^e$, $1 \leq B \ll 2$, their approximations are a', b' with errors at most $\varepsilon_a, \varepsilon_b$ and $|a'|, |b'| \leq B'2^e$ where $1 \leq B \leq B' \ll 2$, $s = a' \pm b'$, then for the rounded value s' of s we have $|s'| \ll 2^{e+2}$ and

$$|a \pm b - s'| \leq |s - s'| + \bigl|a \pm b - (a' \pm b')\bigr|.$$

The second term is at most $\varepsilon_a + \varepsilon_b$, and the real and imaginary parts of the first term has absolute value at most 2^{e-m}, hence the error is at most

$$\varepsilon_a + \varepsilon_b + 2^{e-m+1/2}.$$

2.2. Error estimate for multiplication. Let $|a| \leq B_a 2^{e_a}$, $|b| \leq B_b 2^{e_b}$, where $1 \leq B_a < 2$, $1 \leq B_b < 2$, and let their approximations a', b' with errors at most $\varepsilon_a, \varepsilon_b$. Then $|ab| \leq B_a B_b 2^{e_a+e_b}$. We shall only investigate the case $1 \leq B_a B_b \ll 2$, because we shall use only this. Let $p = a'b'$ and p' be the rounded value of the product. Then we have

$$\begin{aligned}|ab - p'| \leq |ab - a'b'| + |p - p'| \leq |p - p'| + |a||b - b'| + |b'||a - a'| \leq \\ \leq |p - p'| + B_a 2^{e_a} \varepsilon_b + B_b' 2^{e_b} \varepsilon_a,\end{aligned}$$

where $1 \leq B_b \leq B_b' \ll 2$ such that $|b'| \leq B_b' 2^{e_b}$. If a_r, b_r and a_i, b_i are the real and the imaginary parts of a and b, respectively, then during the calculation of

$$p_r = a_r' b_r' - a_i' b_i' \qquad \text{and} \qquad p_i = a_r' b_i' + a_i' b_r'$$

the rounding errors of the products are at most $2^{e_a+e_b-m-1}$ and the rounding errors by the addition and subtraction are also bounded with this bound. Hence we obtain that the rounding error of the real and the imaginary parts are at most $3 \cdot 2^{e_a+e_b-m-1}$, hence $|p - p'|$ is at most $3 \cdot 2^{e_a+e_b-m-1/2}$.

2.3. Error estimate for weights. During FFT and inverse FFT only multiplications with weights, with $2^k = n$'th roots of the unity are used. Among these are ± 1 and $\pm i$; in these cases the representation as floating point number is exact. The real and imaginary parts of all other n'th roots of unity are irrational. This is clear for the eighth's roots $(\pm 1 \pm i)/\sqrt{2}$. Suppose that for some 2^j'th root of unity there are some with rational real or imaginary part, which is not a 2^{j-1}'th root of unity. Let us choose the minimal such j. Then some other would have rational real part. But because of $\cos(2\alpha) = 2\cos^2\alpha - 1$, we obtain that j can be only two.

This means that high precision interval arithmetic calculation can give the truly rounded values of the real and imaginary parts of any n'th root with error at most $2^{-(m+2)}$ in the real and imaginary parts, too. This means that the error of these complex values is at most $\sqrt{2} \cdot 2^{-(m+2)}$. The estimate for the approximation w' of the weight is

$$|w'| \le |w| + |w - w'| = (1 + \sqrt{2} \cdot 2^{-(m+2)}) =: f.$$

This factor f is fairly common in the calculations below, so we shall fix this value for f in what follows. By our experiments some hundred binary digit precision is enough to obtain a fairly large table of roots: see [1].

2.4. Error estimate for multiplication with a weight. Let a and w be the exact values with $|a| \le B2^e$, $|a'| \le B'2^e$, where $1 \le B \le B' \ll 2$. Let $p = a'w'$ with rounded value p'. Then

$$|aw - p'| \le |p - p'| + |aw - a'w'| = |p - p'| + |a||w - w'| + |a - a'||w'|.$$

If $|a - a'| \le \varepsilon$, then the last term is at most εf. The term before is at most $B2^{e-(m+3/2)}$. To estimate the first term on the right hand side consider the real and imaginary parts $p_r = a_r w_r - a_i w_i$ and $p_i = a_i w_r + a_r w_i$. The rounding errors of the real products are at most $2^{e-(m+1)}$. The rounding error by the addition or subtraction also the same, because $|a'w'| \ll 2^{e+1}$, hence this is true for the real and imaginary parts, too. Hence the real and imaginary parts of $p - p'$ has absolute value at most $3 \cdot 2^{e-(m+1)}$. Note that by $w = \pm 1$ and $w = \pm i$ the multiplication is exact.

2.5. Sharper error estimate for multiplication with a weight. Let us consider a special case: if $B = 1$ and hence $B' \le 1 + \varepsilon 2^{-e}$ can be chosen, moreover for each weight which is not ± 1 and not $\pm i$ we have

$$B'\Big(1 - \big|\Re(w)\big|\Big) < \big(1 - 2^{-(m+2)}\big) \quad \text{and} \quad B'\Big(1 - \big|\Im(w)\big|\Big) < \big(1 - 2^{-(m+2)}\big),$$

a better estimate can be given. The conditions means, roughly speaking, that the weights are not very close to ± 1 and $\pm i$ and it is clear that equivalent to

$$(2^e + \varepsilon)\Big(1 - \big|\Re(w)\big|\Big) < 2^{e-1}\big(2 - 2^{-(m+1)}\big)$$

and

$$(2^e + \varepsilon)\Big(1 - \big|\Im(w)\big|\Big) < 2^{e-1}\big(2 - 2^{-(m+1)}\big).$$

In this case the rounding errors of the real products are at most $2^{e-(m+2)}$ and the rounding errors by the addition and subtraction are at most $2^{e-(m+1)}$. Hence the real and imaginary part of $p - p'$ has absolute value at most $2 \cdot 2^{e-(m+1)}$.

Note that multiplication and addition/subtraction considered as separate operation. Today several processor capable the do fused multiplication and addition/subtraction. (Earlier this was typical only for IBM processors, for example for the Power family and predecessors.) In this case somewhat better error estimate can be given. Of course, which instructions and in which order are done is depends on the assembly code. Hence such estimates are valid only for given assembly code, and not a high level language code, where the corresponding assembly code depends on the compiler.

2.6. Error estimate for butterfly operation. Let a, b be the exact values with $|a|, |b| \le B2^e$, $|a'|, |b'| \le B'2^e$, where $1 \le B \le B' \ll 2$. Let w be a weight. We want to calculate the results of the butterfly operation $a \pm bw$. Let $p = b'w'$ the product with rounded value p', and let $r = a' \pm p'$. Then

$$\begin{aligned} |a \pm bw - r'| &\le \big|a \pm bw - (a' \pm b'w')\big| + +\big|r' - (a' \pm b'w')\big| \le \\ &\le |a - a'| + |bw - b'w'| + \big|r' - (a' \pm b'w')\big| \le \\ &\le \varepsilon_a + |b||w - w'| + |b - b'||w'| + \big|r' - (a' \pm b'w')\big| \le \\ &\le \varepsilon_a + B2^{e-(m+3/2)} + \varepsilon_b f + \big|r' - (a' \pm b'w')\big|. \end{aligned}$$

As we explained by the estimation of the error of the product with weight, the real and imaginary part of the product has rounding error at most $3 \cdot 2^{e-(m+1)}$. During addition or subtraction a rounding error at most $2 \cdot 2^{e-(m+1)}$ is added in the real and imaginary parts, too. Hence the complete rounding error is at most

$$\big|r' - (a' \pm b'w')\big| \le 5 \cdot 2^{e-(m+1/2)}.$$

Hence the total error is at most

$$\varepsilon_a + B2^{e-(m+3/2)} + \varepsilon_b f + 5 \cdot 2^{e-(m+1/2)}.$$

2.7. Sharper error estimate for the butterfly operation. On the same way as by the multiplication with a weight, in the case $B = 1$ a sharper error estimate can be given, if for each weight which is not ±1 and not $\pm i$ we have

$$(1 + \varepsilon_b 2^{-e})\Big(1 - \big|\Re(w)\big|\Big) < 1 - 2^{-(m+2)}$$

and

$$(1 + \varepsilon_b 2^{-e})\Big(1 - \big|\Im(w)\big|\Big) < 1 - 2^{-(m+2)}.$$

Then, as we have seen, the rounding errors of the real and imaginary parts of the complex product are at most $2 \cdot 2^{e-(m+1)}$. During the addition or subtraction the rounding errors are at most $2 \cdot 2^{e-(m+1)}$ in the real and imaginary parts, too. Hence the complete rounding error is at most

$$\left|r' - (a' \pm b'w')\right| \le 4 \cdot 2^{e-(m+1/2)},$$

and the total error is at most

$$\varepsilon_a + B2^{e-(m+3/2)} + \varepsilon_b f + 4 \cdot 2^{e-(m+1/2)}.$$

If we, additionally, know that the result has absolute value $\ll 2^e$, then the rounding error during the addition or subtraction is at most $2^{e-(m+2)}$ in the real and the imaginary part, hence the complete rounding error is

$$\left|r' - (a' \pm b'w')\right| \le 5 \cdot 2^{e-(m+3/2)},$$

and the total error is at most

$$\varepsilon_a + B2^{e-(m+3/2)} + \varepsilon_b f + 5 \cdot 2^{e-(m+3/2)}.$$

2.8. Error estimate for double round butterfly operation. If we do two rounds, then supposing that $\varepsilon_a = \varepsilon_b = \varepsilon_c = \varepsilon_d = \varepsilon$ for the four numbers and taking into consideration that during the second round we have to substitute $e_1 = e + 1$ to the place of e, we obtain the error bound

$$(1+f)\varepsilon_1 + B2^{e_1-(m+3/2)} + 5 \cdot 2^{e_1-(m+1/2)},$$

where

$$\varepsilon_1 = (1+f)\varepsilon + B2^{e-(m+3/2)} + 5 \cdot 2^{e-(m+1/2)}.$$

This results the total error bound

$$\begin{aligned}&(1+f)^2\varepsilon + (1+f)B2^{e-(m+3/2)} + (1+f)5 \cdot 2^{e-(m+1/2)} + \\ &\quad + 2B2^{e-(m+3/2)} + 10 \cdot 2^{e-(m+1/2)} = \\ &\quad = (1+f)^2\varepsilon + (3+f)B2^{e-(m+3/2)} + (3+f)5 \cdot 2^{e-(m+1/2)}.\end{aligned}$$

Let us estimate the error by two-round butterfly operations using less arithmetic operations (see [1]). By these the exact results are calculated as

$$(a \pm cw_2) \pm i(bw_1 + dw_3),$$

where w_1, w_2, w_3 are appropriate weights. Let r' be the rounded result. Then

$$\begin{aligned}&\left|(a \pm cw_2) \pm i(bw_1 + dw_3) - r'\right| \le \\ &\quad \le \left|(a \pm cw_2) \pm i(bw_1 + dw_3) - \big((a' \pm c'w_2') \pm i(b'w_1' + d'w_3')\big)\right| + \\ &\quad + \left|r' - \big((a' \pm c'w_2') \pm i(b'w_1' + d'w_3')\big)\right|.\end{aligned}$$

For the first term on the right hand side, similarly as by the butterfly operation, for the absolute value we obtain the upper bound

$$\varepsilon(1+3f)+3B2^{e-(m+3/2)}.$$

The second term is the rounding error. During the calculation of the products the rounding error in the real and in the imaginary parts are at most $3{\cdot}2^{e-(m+1)}$. By addition or subtraction in the small parenthesizes we get errors at most $2\cdot 2^{e-(m+1)}$ and by the final addition or subtraction at most $4\cdot 2^{e-(m+1)}$ in the real and imaginary parts, too. Hence the error of the real and imaginary parts is at most $17\cdot 2^{e-(m+1)}$ and the total rounding error is at most $17\cdot 2^{e-(m+1/2)}$. Hence the total error is at most

$$\varepsilon(1+3f)+3B2^{e-(m+3/2)}+17\cdot 2^{e-(m+1/2)}.$$

This is somewhat less, than the error of two separated rounds.

2.9. Error estimate for triple round butterfly operation. About these see [1]. Here the last operation is a butterfly $a\pm wb$, where $w=1,i,(1+i)/\sqrt{2},(1-i)/\sqrt{2}$. The value a is obtained from 4 starting value and 3 weights on the same way as above, hence we have

$$|a-a'|\le\varepsilon(1+3f)+3B2^{e-(m+3/2)}+17\cdot 2^{e-(m+1/2)}.$$

The value b is obtained similarly, but all the four starting value is multiplied by a weight. Hence similar calculation as above shows that

$$|b-b'|\le 4\varepsilon f+4B2^{e-(m+3/2)}+20\cdot 2^{e-(m+1/2)}.$$

Because the last operation is a butterfly operation, but with $e_2=e+2$ instead of e, the error of the result is at most

$$\begin{aligned}
&\varepsilon_a+B2^{e_2-(m+3/2)}+\varepsilon_b f+5\cdot 2^{e_2-(m+1/2)}=\\
&\quad=\varepsilon(1+3f)+3B2^{e-(m+3/2)}+17\cdot 2^{e-(m+1/2)}+\\
&\qquad+4\varepsilon f^2+4Bf2^{e-(m+3/2)}+20f\cdot 2^{e-(m+1/2)}+\\
&\qquad+4B2^{e-(m+3/2)}+20\cdot 2^{e-(m+1/2)}=\\
&\quad=\varepsilon(1+3f+4f^2)+B(7+4f)2^{e-(m+3/2)}+(37+20f)2^{e-(m+1/2)}.
\end{aligned}$$

This is somewhat better as the estimate for three separate round:

$$\begin{aligned}
&(1+f)^3\varepsilon+(1+f)(3+f)B2^{e-(m+3/2)}+(1+f)(3+f)5\cdot 2^{e-(m+1/2)}+\\
&\quad+4B2^{e-(m+3/2)}+20\cdot 2^{e-(m+1/2)}=\\
&\quad=(1+f)^3\varepsilon+B(7+4f+f^2)2^{e-(m+3/2)}+\\
&\quad+(35+20f+5f^2)\cdot 2^{e-(m+1/2)}.
\end{aligned}$$

3. Error estimate for the complete FFT

We shall consider only the basic case: high-precision multiplication using FFT. We only consider the case of single round butterfly operations.

3.1. Error estimate for the FFT. We suppose that by starting the $n/2 = 2^{k-1}$ complex numbers in the lower part of the array has absolute value of the real and imaginary parts at most $2^{\ell-1}$, other elements of the array are zero. After one round all the array contains complex numbers having absolute value at most $2^{\ell-1/2}$, and all they are exact. After the second round they are still exact if $\ell \leq m$ and they have absolute value at most $2^{\ell+1/2}$, i.e., $B2^{e_2}$, where $B = sqrt2$ and $e_2 = l$. After $j \geq 2$ rounds the absolute value of the exact values is at most $B2^{e_j}$, where $B = \sqrt{2}$, $e_j = \ell + j - 2$. For the error estimates we know that $\varepsilon_0 = \varepsilon_1 = \varepsilon_2 = 0$. Using the error estimate of the butterfly operation for $j > 2$ we have

$$\begin{aligned}\varepsilon_{j+1} &= (1+f)\varepsilon_j + B2^{e_j-(m+3/2)} + 5\cdot 2^{e_j-(m+1/2)} = \\ &= (1+f)\varepsilon_j + (5\sqrt{2}+1)2^{l+j-m-3}.\end{aligned}$$

Hence

$$\begin{aligned}\varepsilon_3 &= (5\sqrt{2}+1)2^{l-m}, \\ \varepsilon_4 &= (1+f)(5\sqrt{2}+1)2^{l-m} + (5\sqrt{2}+1)2^{l-m+1} = \\ &\quad = (3+f)(5\sqrt{2}+1)2^{l-m}, \\ \varepsilon_5 &= (5\sqrt{2}+1)\big((1+f)^2 + 2(1+f) + 4\big)2^{l-m}.\end{aligned}$$

By induction we have

$$\varepsilon_j = (5\sqrt{2}+1)2^{l-m}\sum_{t=0}^{j-3}(1+f)^t 2^{j-3-t}.$$

From this

$$\begin{aligned}\varepsilon_j &\leq (5\sqrt{2}+1)2^{l-m}(j-2)(1+f)^{j-3} = \\ &= \left(\frac{1+f}{2}\right)^{j-3} 2^{l+j-m}\frac{5\sqrt{2}+1}{8}(j-2),\end{aligned}$$

whenever $j > 2$.

3.2. Bound for exact results of the FFT. By the previous calculation we have $|\hat{F}_j| \leq 2^{l+k-3/2}$. Hence we have $2|\hat{f}_j| \leq 2^{l+k+1/2}$. A sharper bound can be obtained if we consider what would happen during the FFT of the array f_j. We would start with 2^k real numbers having absolute value at most 2^{l-1} and with 2^k zero. After the first round we would obtain 2^{k+1} real number with absolute value at most $2^{\ell-1}$, after the second round we would obtain complex numbers having absolute value at most 2^{ℓ}, etc. Finally after $k+1$ rounds we may get the numbers $\hat{f}_j$ having absolute value at most 2^{l+k-1}, i.e., $2|\hat{f}_j| \leq 2^{\ell+k}$.

This results the bounds $4|\hat{h}_j| \leq 2^{2l+2k}$. From this we may obtain the bound $8|\hat{H}_j| \leq 2^{2l+2k+2}$, but this later is not optimal. In the next point we shall obtain better bound, considering the properties of the array h.

3.3. Bound for all exact partial results during inverse FFT. After the inverse FFT we obtain the convolution sums $h_j = \sum_{s=0}^{j} f_s g_{j-s}$. Because $|f_s|, |g_s| \leq 2^{l-1}$ we obtain that

$$|h_j| \leq \begin{cases} (j+1)2^{2\ell-2}, & \text{if } 0 \leq j < n, \\ (2n-1-j)2^{2\ell-2}, & \text{if } n \leq j < 2n. \end{cases}$$

From the complex numbers $H_j = h_{2j} + ih_{2j+1}$, $0 \leq j < n$ after one round of FFT we would obtain complex numbers having real and imaginary parts with absolute value at most $2^{2\ell+k-2}$; indeed the absolute value of the real part of $H_j \pm H_{j+n/2}$ is at most

$$(2j+1)2^{2\ell-2} + \big(2n-1-2(j+n/2)\big)2^{2\ell-2} = n2^{2\ell-2},$$

and the same is true for the imaginary part. This means that the absolute value of these complex numbers is at most $2^{2\ell+k-3/2}$. Using this, during the FFT, after j rounds we obtain that the absolute value of the numbers is at most $2^{2\ell+k+j-5/2}$. Hence $|8\hat{H}_j| \leq 2^{2\ell+2k+1/2}$. Moreover, during the inverse FFT, this is true after each round, except the last one, because the inverse FFT rounds are the reverse of the FFT rounds, except we does not divide by 2. The same is true for the results of the two, three, etc. rounds inverse FFT steps, including all results off addition/subtraction, because each weight has absolute value at most 1.

4. Error estimate for further elementary operations

4.1. Error estimate for $\hat{F} \to \hat{f}$ conversion. $\hat{F} \to \hat{f}$ denotes the conversion of the DFT of the complex sequence F_j to the DFT of the real sequence f_j. Let us suppose that $|\hat{F}_j| \leq B2^e$, where $1 \leq B \ll 2$ and the error of the approximation of each $\hat{F}_j$ is at most ε. We should like to obtain an error estimate ε_r for $|2\hat{f}_j - 2\hat{f}'_j|$. The error of the approximation of $\hat{F}_j \pm \overline{\hat{F}}_{n-j}$ is at most $2\varepsilon + 2^{e-m+1/2}$. Using the error estimate for the butterfly operation we obtain that

$$\varepsilon_r \leq (1+f)(2\varepsilon + 2^{e-m+1/2}) + B2^{e-m+1/2} + 5 \cdot 2^{e-m+1/2}.$$

Because $B = \sqrt{2}$ and $e = \ell + k - 2$, we have

$$\begin{aligned} \varepsilon_r &\leq (1+f)(2\varepsilon + 2^{l+k-m-3/2}) + 2^{\ell+k-m-3/2} + 2^{\ell+k-m-1} + 5 \cdot 2^{l+k-m-3/2} \leq \\ &\leq \frac{1+f}{2}\bigl(4\varepsilon + 2^{\ell+k-m-2}(7\sqrt{2}+2)\bigr). \end{aligned}$$

If $k = 2$ and $\ell < m$, then the calculation of $\hat{F}_j \pm \overline{\hat{F}}_{n-j}$ is exact, hence we have

$$\varepsilon_r \leq 2^{\ell+k-m-2}(5\sqrt{2}+2).$$

4.2. Error estimate for digit-by-digit multiplication. Let us denote by ε_x the error bound for $4\hat{h}'_j$. Using the error estimate for complex multiplication we obtain

$$\varepsilon_x \leq 3 \cdot 2^{2\ell+2k-m-1/2} + 2^{\ell+k}\varepsilon_r + (2^{\ell+k} + \varepsilon_r)\varepsilon_r.$$

4.3. Error estimate for $\hat{h} \to \hat{H}$ conversion. The exact value of $4\hat{h}_j \pm 4\overline{\hat{h}}_{n-j}$ is bounded by $2^{2\ell+2k+1}$ and the error by

$$2\varepsilon_x + 2^{2\ell+2k-m+1/2}.$$

Hence using the error estimate for the butterfly operation we obtain for the error bound ε_c of $8\hat{H}'_j$ that

$$\begin{aligned} \varepsilon_c &\leq (1+f)(2\varepsilon_x + 2^{2\ell+2k-m+1/2}) + 2^{2\ell+2k-m-1/2} + 5 \cdot 2^{2\ell+2k-m+1/2} \leq \\ &\leq \frac{1+f}{2}\bigl(4\varepsilon_x + 2^{2\ell+2k-m-2}(8\sqrt{2} + 2\sqrt{2} + 20\sqrt{2})\bigr) \leq \\ &\leq \frac{1+f}{2}(4\varepsilon_x + 15\sqrt{2} \cdot 2^{2\ell+2k-m-1}). \end{aligned}$$

If we prove that the sharper estimate can be used by the butterfly operation than we have that

$$\begin{aligned} \varepsilon_c &\leq (1+f)(2\varepsilon_x + 2^{2\ell+2k-m+1/2}) + 2^{2\ell+2k-m-1/2} + 5 \cdot 2^{2\ell+2k-m-1/2} \leq \\ &\leq \frac{1+f}{2}(4\varepsilon_x + 5\sqrt{2} \cdot 2^{2\ell+2k-m}). \end{aligned}$$

4.4. Error estimate for inverse butterflies. Let $|a|, |b| \leq B2^e$, where $1 \leq B \ll 2$ and suppose that $|a \pm b| \leq B2^e$ also satisfied. Let $s = a' + b'$ and

$p = s'w'$. Then

$$\begin{aligned}\bigl|(a \pm b)w - p'\bigr| &\le \bigl|(a \pm b)w - (a' \pm b')w'\bigr| + \bigl|p' - (a' \pm b')w'\bigr| \le \\ &\le |a - a'||w'| + |b - b'||w'| + |a||w - w'| + |b||w - w'| + \\ &\quad + \bigl|p' - (a' \pm b')w'\bigr| \le \\ &\le (\varepsilon_a + \varepsilon_b)f + B2^{e-m-1/2} + \bigl|p' - (a' \pm b')w'\bigr|.\end{aligned}$$

Here

$$\bigl|p' - (a' \pm b')w'\bigr| \le |p - p'| + |s'w' - sw'| \le |w'||s - s'| + |p - p'|.$$

Because $|p - p'|$ is the rounding error of the multiplication, it is at most $3 \cdot 2^{e-m-1/2}$. The other term is at most $f2^{e-m-1/2}$. Hence the complete error is at most

$$(\varepsilon_a + \varepsilon_b)f + B2^{e-m-1/2} + 3 \cdot 2^{e-m-1/2} + f2^{e-m-1/2}.$$

Using that $B = \sqrt{2}$, $e = 2\ell + 2k$ and $\varepsilon_a = \varepsilon_b = \varepsilon$, the total error is at most

$$2f\varepsilon + 2^{2\ell+2k-m-1}\bigl(2 + (3 + f)\sqrt{2}\bigr).$$

In the last round $|a \pm b| \le B2^e$ not satisfied, but there is no multiplication. Hence the same estimate remains valid.

5. Sufficient conditions from the estimates

5.1. Sufficient condition from the error estimate for the inverse FFT. Let ε denote the error bound by starting the inverse FFT, and let

$$\delta = 2^{2\ell+2k-m-1}\bigl(2 + (3 + f)\sqrt{2}\bigr).$$

After one round of inverse FFT the error bound is $2f\varepsilon + \delta$, after two rounds of inverse FFT the error bound is

$$2f(2f\varepsilon + \delta) + \delta = 4f^2\varepsilon^2 + 2f\delta + \delta,$$

etc., after k rounds of inverse FFT at most

$$2^k f^k \varepsilon + \delta(1 + 2f + 4f^2 + \cdots + 2^{k-1}f^{k-1}) \le (2f)^k(\varepsilon + \delta).$$

Because after the inverse FFT we divide by 2^{k+3}, and the error of the result have to be less then $1/2$, the condition $(2f)^k(\varepsilon + \delta) < 2^{k+2}$ have to be satisfied, which is equivalent to $f^k(\varepsilon + \delta) < 4$. To obtain an upper estimate for f^k we shall use Bernoulli's inequality $(1+x)^j \geq 1+jx$, if $x>-1$, $j=0,1,\dots$. From this inequality we obtain

$$\Big(\frac{1}{f}\Big)^j \geq 1+j\Big(\frac{1}{f}-1\Big) = 1-j\Big(1-\frac{1}{f}\Big),$$

whence

$$f^j \leq \frac{1}{1-j\big(1-\frac{1}{f}\big)} = \frac{1}{1-j\big(\frac{f-1}{f}\big)} \leq \frac{1}{1-j2^{-(m+3/2)}}.$$

Hence the condition $f^k(\varepsilon+\delta)<4$ is certainly satisfied if

$$\frac{\varepsilon+\delta}{1-k2^{-(m+3/2)}} < 4,$$

i.e. if

$$\varepsilon+\delta < 4-k2^{-m+1/2}.$$

This results for ε the condition

$$\varepsilon < 4-k2^{-m+1/2}-2^{2\ell+2k-m-1}\big(2+(3+f)\sqrt{2}\big).$$

5.2. Sufficient condition for the number of rounds. Summarizing our estimates we obtain that

$$\varepsilon_r \leq \Big(\frac{1+f}{2}\Big)^{k-2} 2^{\ell+k-m-1}\big((5\sqrt{2}+1)(k-2)+7/\sqrt{2}+1\big),$$

and in the case $k=2$ this is true without the factor $(1+f)/2$. From this

$$\begin{aligned}\varepsilon_x &\leq 2^{\ell+k+1}\varepsilon_r+\varepsilon_r^2+3\cdot 2^{2\ell+2k-m-1/2} \leq \\ &\leq \Big(\frac{1+f}{2}\Big)^{k-2} 2^{2\ell+2k-m-2}\cdot\Bigg(4(5\sqrt{2}+1)(k-2)+20\sqrt{2}+4+ \\ &\quad +2^{-m}\Big(\frac{1+f}{2}\Big)^{k-2}\big((5\sqrt{2}+1)(k-2)+7/\sqrt{2}+1\big)^2\Bigg).\end{aligned}$$

The last term in the large parenthesis is very small. To obtain an upper bound we use again Bernoulli's inequality:

$$\Big(\frac{2}{1+f}\Big)^j \geq 1+j\Big(\frac{2}{1+f}-1\Big) = 1-j\Big(1-\frac{2}{1+f}\Big),$$

whence

$$\Big(\frac{1+f}{2}\Big)^j \leq \frac{1}{1-j\big(1-\frac{2}{1+f}\big)} = \frac{1}{1-j\big(\frac{f-1}{f+1}\big)} \leq \frac{1}{1-j2^{-(m+5/2)}}.$$

From this if $k \leq m$ then the last term in the large parenthesis is at most

$$0 \leq c(k,m) := \frac{2^{-m}\big((5\sqrt{2}+1)(k-2)+7/\sqrt{2}+1\big)^2}{1-(k-2)2^{-(m+5/2)}}. \tag{1}$$

Hence finally we obtain the estimate

$$\varepsilon_x \leq \Big(\frac{1+f}{2}\Big)^{k-2} 2^{2\ell+2k-m-2}\Big(4(5\sqrt{2}+1)k-20\sqrt{2}-4+c(k,m)\Big).$$

From this

$$\varepsilon_c \leq \Big(\frac{1+f}{2}\Big)^{k-1} 2^{2\ell+2k-m-2}\Big(16(5\sqrt{2}+1)k-50\sqrt{2}-16+4c(k,m)\Big).$$

So it is enough if the following condition is satisfied:

$$\Big(\frac{1+f}{2}\Big)^{k-1} 2^{2\ell+2k-m}\Big(4(5\sqrt{2}+1)k-25/\sqrt{2}-4+c(k,m)\Big)< \\ < 4-k2^{-m+1/2}-2^{2\ell+2k-m}\big(1+(3+f)/\sqrt{2}\big).$$

This certainly satisfied if $k \geq 2$ and

$$\Big(\frac{1+f}{2}\Big)^{k-1} 2^{2\ell+2k-m}\big(4(5\sqrt{2}+1)k-22/\sqrt{2}+f/\sqrt{2}-3+c(k,m)\big)< 4-k2^{-m+1/2}.$$

Using again the estimate from Bernoulli's inequality we obtain that for this it is enough if

$$2^{2\ell+2k-m}\big(4(5\sqrt{2}+1)k-22/\sqrt{2}+f/\sqrt{2}-3+c(k,m)\big)< \\ < (4-k2^{-m+1/2})\big(1-(k-1)2^{-(m+5/2)}\big),$$

i. e., if

$$2^{2\ell+2k-m}\big(4(5\sqrt{2}+1)k-22/\sqrt{2}+f/\sqrt{2}-3+c(k,m)\big)< \\ < 4-k2^{-m+1/2}-(k-1)2^{-m-1/2}+k(k-1)2^{-2m-2}.$$

To this it is clearly enough if

$$2^{2\ell+2k-m}\big(4(5\sqrt{2}+1)k-22/\sqrt{2}+f/\sqrt{2}-3+c(k,m)\big)< \\ < 4-k2^{-m+1/2}-(k-1)2^{-m-1/2}.$$

Rearranging this inequality we obtain

$$(5\sqrt{2}+1+3\cdot 2^{-2\ell-2k-5/2})k-2^{-2\ell-2k-5/2}-\frac{11}{4}\sqrt{2}+\frac{f}{8}\sqrt{2}-\frac{3}{4}+\frac{c(k,m)}{4} < 2^{m-2\ell-2k}.$$

Because $\ell \geq 2$ supposing that $k \geq 2$ we obtain the semifinal form of our sufficient condition:
(2)

$$(5\sqrt{2}+1+3\cdot 2^{-17/2})(k-2)+3\cdot 2^{-19/2}+\frac{29}{4}\sqrt{2}+\frac{f}{8}\sqrt{2}+\frac{5}{4}+\frac{c(k,m)}{4} < 2^{m-2\ell-2k}.$$

5.3. Sharper error estimates and weaker sufficient condition for the number of rounds. We may obtain a sharper estimate if we use the sharper estimate of the butterfly operation during the computation of $8\hat{H}_j$, using that $4|\hat{h}_j| \leq 2^{2\ell+2k}$, and hence the absolute value of (exact) incoming data of the butterfly are at most $2^{2\ell+2k+1}$ and that $|8\hat{H}_j| \leq 2^{2\ell+2k+3/2}$. But the sharper estimate of the butterfly operation can be used only if the condition

$$(1+\xi)(1-\eta) < 1-\zeta$$

is satisfied, where $\xi = \varepsilon_b 2^{-2\ell-2k-1}$, $1-\eta$ is supremum of the absolute values of all real and imaginary parts of the weights $w = e^{2\pi ij/2^{k+1}}$ which are different from ± 1 and $\pm i$, i.e., $\cos(\pi/2^k)$, and $\zeta = 2^{-(m+2)}$. Because ξ, η and ζ are positive quantities less then 1, the inequality is satisfied if $\xi + \zeta \leq \eta$.

Let us substitute η with a lower estimate. Because

$$\cos x = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \cdots,$$

we obtain that

$$\begin{aligned}\eta = 1 - \cos\frac{\pi}{2^k} &\geq \frac{\pi^2}{2^{2k+1}}\Big(1 - \frac{\pi^2}{12\cdot 2^{2k}}\Big) \geq \\ &\geq \frac{\pi^2}{2^{2k+1}}\Big(1 - \frac{\pi^2}{12\cdot 2^4}\Big) \geq \frac{\pi^2}{2^{2k+1}}\Big(1 - \frac{10}{192}\Big) \geq \\ &\geq \frac{91\pi^2}{192}2^{-2k} > 4\cdot 2^{-2k}.\end{aligned}$$

Using that the error of $4\hat{h}_j \pm 4\hat{h}_{n-j}$ is at most

$$2\varepsilon_x + 2^{2\ell+2k-m+1/2},$$

the sharper error estimate for the multiplication can be used if

$$\frac{2\varepsilon_x + 2^{2\ell+2k-m+1/2}}{2^{2\ell+2k+1}} + 2^{-(m+2)} \leq 4\cdot 2^{-2k},$$

i. e., if

$$\Big(\frac{1+f}{2}\Big)^{k-2} 2^{-m}\big((5\sqrt{2}+1)k - 4\sqrt{2} - 1 + c(k,m)/4\big) + 2^{-m-2} \leq 4\cdot 2^{-2k}.$$

Using the upper estimate to power of $(1+f)/2$ from Bernoulli's inequality, this is certainly satisfied if

$$\frac{(5\sqrt{2}+1)k-4\sqrt{2}-1+c(k,m)/4}{1-(k-2)2^{-(m+5/2)}}+\frac{1}{4}\leq 4\cdot 2^{m-2k}$$

i. e., if

$$(5\sqrt{2}+1)k-4\sqrt{2}-1+\frac{c(k,m)}{4}+\frac{1}{4}-\frac{k-2}{4}2^{-(m+5/2)}\leq$$
$$\leq 4\cdot 2^{m-2k}-4(k-2)2^{-2k-5/2}.$$

This is certainly satisfied if

$$\frac{5\sqrt{2}+1}{4}k-\sqrt{2}-\frac{3}{16}+\frac{c(k,m)}{16}+(k-2)2^{-2k-5/2}\leq 2^{m-2k}.$$

If $k\geq 2$ then $k-2\leq 2^{k-3}$, from which

$$(k-2)2^{-2k-5/2}\leq 2^{(k-3)-2k-5/2}=2^{-k-11/2}\leq 2^{-15/2}.$$

Hence our condition is satisfied if

$$\text{(3)}\qquad \frac{5\sqrt{2}+1}{4}(k-2)+\frac{3}{2}\sqrt{2}+\frac{5}{16}+\frac{c(k,m)}{16}+2^{-15/2}\leq 2^{m-2k}.$$

If this condition is satisfied then the sharper estimate can be used for the $\hat{h}\to\hat{H}$ conversion, and we obtain

$$\varepsilon_c\leq\frac{1+f}{2}(4\varepsilon_x+5\sqrt{2}\cdot 2^{2\ell+2k-m}).$$

Using the estimate for ε_x from the previous paragraph we obtain that

$$\varepsilon_c\leq\Big(\frac{1+f}{2}\Big)^{k-1}2^{2\ell+2k-m}\Big(4(5\sqrt{2}+1)k-15\sqrt{2}-4+c(k,m)\Big).$$

Similar calculations as in the previous paragraph result the sufficient condition (4)

$$(5\sqrt{2}+1+3\cdot 2^{-17/2})(k-2)+5\cdot 2^{-19/2}+\frac{27}{4}\sqrt{2}+\frac{f}{8}\sqrt{2}+\frac{5}{4}+\frac{c(k,m)}{4}<2^{m-2\ell-2k}$$

if $k\geq 2$. Because $\ell\geq 2$, this condition is much stronger as condition (3), so we have to consider only (4). On the left hand side $c(k,m)$ and $(f-1)/\sqrt{2}$ are small quantities. To make the condition nicer let us estimate their sum. Because $(f-1)/\sqrt{2}=2^{-(m+2)}$, we obtain

$$\frac{f-1}{\sqrt{2}}+c(k,m)\leq 2^{-m}\frac{\big(5\sqrt{2}+1)(k-2)+7/\sqrt{2}+1\big)^2+1/4}{1-(k-2)2^{-(m+5/2)}}.$$

From (4) we obtain that $6\sqrt{2} < 2^{m-2\ell-2k}$. Hence the integer $m-2\ell-2k$ is at least 4, and because of $\ell \geq 2$ we obtain $2k \leq m-8$ and by $k \geq 2$ also that $m \geq 12$. The left hand side is at most

$$\frac{\left((5\sqrt{2}+1)(k-2)+7/\sqrt{2}+3/2\right)^2}{2^m-(m/2-6)2^{-5/2}} < \frac{(5\sqrt{2}+1)^2k^2}{2^m-(m/2-6)2^{-5/2}} \leq$$

$$\leq \frac{\left(\frac{5\sqrt{2}+1}{2}\right)^2(m-8)^2}{2^m-(m/2-6)2^{-5/2}} < \frac{4.1^2(m-8)^2}{2^m-(m/2-6)2^{-5/2}}.$$

We shall investigate the function g defined by the right hand side and prove by induction that it is strictly monotonic for $m \geq 12$. Indeed,

$$g(m+1) = \frac{4.1^2(m-7)^2}{2^{m+1}-\left(\frac{m}{2}-\frac{11}{2}\right)2^{-5/2}} < \frac{2\cdot 4.1^2(m-8)^2}{2^{m+1}-(m-12)2^{-5/2}} = g(m)$$

is proved if we prove that the nominator is larger and the denominator is smaller on the right hand side. But

$$2(m-8)^2-(m-7)^2 = m^2-18m+79 = (m-9)^2-2,$$

which is positive if $m \geq 11$. Similarly

$$2^{m+1}-\left(\frac{m}{2}-\frac{11}{2}\right)2^{-5/2} > 2^{m+1}-(m-12)2^{-5/2},$$

if $m > 11$. The final upper estimate is given by $g(12)$ which is less then 0.066.

6. Final results

Substituting the numerical values in (4) and rounding up we obtain the condition

$$8.074(k-2)+10.978 < 2^{m-2\ell-2k}$$

for the case $k \geq 2$. Although not important, we remark that in the case $k=0$ the computation is exact if $2\ell \leq m+3$ and in the case $k=1$ the computation is exact if $2\ell \leq m+2$.

Probably the only processor today having quadruple precision floating point arithmetic in hardware is the Power9 of IBM. On this, choosing $\ell = 16$ we may choose $k = 35$ and multiply numbers shorter than 2^{39} bit, or we may choose $\ell = 8$ and $k = 43$ and multiply numbers shorter than 2^{46} bit. But using only the high speed of double precision floating point SIMD operations we may do exact multiplication of shorter numbers on this or on more common processors.

Acknowledgement. The author thanks for the referee correcting several misprints and two miscalculations.

References

[1] **Járai, A. and Z. Járai,** *Fast Arbitrary Precision Package: Natural number routines*, ELTE Informatikai Kar, 2012.

[2] **Knuth, Donald E.,** *The Art of Computer Programming, second edition, Volume 2: Seminumerical Algorithms*, Addison-Wesley, 1981.

A. Járai
Eötvös Loránd University
Pázmány Péter sétány 1/C
H-1117 Budapest
Hungary
ajarai@moon.inf.elte.hu