Improved error bounds for floating-point products and Horner's scheme

Siegfried M. Rump, Florian Bünger, Claude-Pierre Jeannerod

Noname manuscript No.
(will be inserted by the editor)

Improved Error Bounds for Floating-Point Products and Horner's Scheme

Siegfried M. Rump · Florian Bünger · Claude-Pierre Jeannerod

Received: date / Accepted: date

Abstract Let $\mathbf{u}$ denote the relative rounding error of some floating-point format. Recently it has been shown that for a number of standard Wilkinson-type bounds the typical factors $\gamma_k := k\mathbf{u}/(1-k\mathbf{u})$ can be improved into $k\mathbf{u}$, and that the bounds are valid without restriction on k. Problems include summation, dot products and thus matrix multiplication, residual bounds for LU- and Cholesky-decomposition, and triangular system solving by substitution.

In this note we show a similar result for the product $\prod_{i=0}^{k} x_i$ of real and/or floating-point numbers x_i, for computation in any order, and for any base $\beta \geqslant 2$. The derived error bounds are valid under a mandatory restriction of k. Moreover, we prove a similar bound for Horner's polynomial evaluation scheme.

Keywords floating-point product · IEEE 754 standard · Wilkinson type error estimates · Horner scheme

CR Subject Classification 65G50 · 65F05

Siegfried M. Rump
Institute for Reliable Computing, Hamburg University of Technology, Schwarzenbergstraße 95, Hamburg 21071, Germany, and Faculty of Science and Engineering, Waseda University, 3-4-1 Okubo, Shinjuku-ku, Tokyo 169-8555, Japan (`rump@tuhh.de`).

Florian Bünger
Institute for Reliable Computing, Hamburg University of Technology, Schwarzenbergstraße 95, Hamburg 21071, Germany (`florian.buenger@tuhh.de`).

Claude-Pierre Jeannerod
Inria, Laboratoire LIP (CNRS, ENS de Lyon, Inria, UCBL), Université de Lyon, 46 allée d'Italie 69364 Lyon cedex 07, France (`claude-pierre.jeannerod@ens-lyon.fr`).

1 Introduction and notation

Denote by $\mathbb{F}$ a set of floating-point numbers with p digits precision in base β, and with operations according to IEEE 754 standard [3] in rounding to nearest with any tie breaking rule. Then, $\mathbf{u} := \frac{1}{2}\beta^{1-p}$ denotes the relative rounding error unit. Throughout the paper we assume that $\beta \geqslant 2$ and $p \geqslant 1$, and that neither overflow nor underflow occurs.

As usual, for $\circ \in \{+, -, \cdot, /\}$ and $a, b \in \mathbb{F}$, the floating-point result of an operation $a \circ b$ is defined to be $\mathrm{fl}(a \circ b)$ for a rounding to nearest $\mathrm{fl}\colon \mathbb{R} \to \mathbb{F}$. It follows [2, p. 38] that $|\mathrm{fl}(x) - x| \leqslant \mathbf{u}|x|$ for $x \in \mathbb{R}$, and in particular

$$|\mathrm{fl}(a \circ b) - (a \circ b)| \leqslant \mathbf{u}|a \circ b|. \tag{1.1}$$

For matrices $A \in \mathbb{F}^{m\times k}$ and $B \in \mathbb{F}^{k\times n}$, denote by $\widehat{C}$ the floating-point result of the exact product $C := AB$ computed using (blocked versions of) the classical algorithm, with any ordering for the inner products. A rounding error analysis *à la* Wilkinson then leads typically to $|\widehat{C} - C| \leqslant \gamma_k|A||B|$ with $\gamma_k := \frac{k\mathbf{u}}{1-k\mathbf{u}} = k\mathbf{u} + O(\mathbf{u}^2)$; see for example [2, p. 71]. This standard estimate has been improved in [4] into

$$|\widehat{C} - C| \leqslant k\mathbf{u}|A||B| \tag{1.2}$$

without restriction on the integer k and, in [9], similar improvements have been obtained for the residuals of the computed LU and Cholesky factors as well as for triangular system solutions.

A similar result was recently shown by Graillat, Lefèvre, and Muller [1] for binary arithmetic:

Theorem 1.1 *Assume $\beta = 2$ and let $x \in \mathbb{F}$ and $k \in \mathbb{N}$ be given. If the power x^{k+1} is computed by successive multiplications by x, then, in absence of underflow and overflow, the computed approximation $\widehat{r}$ satisfies*

$$\left|\widehat{r} - x^{k+1}\right| \leqslant k\mathbf{u}\left|x^{k+1}\right| \qquad \textit{if} \quad k + 1 \leqslant \sqrt{2^{1/3} - 1} \cdot \mathbf{u}^{-1/2}. \tag{1.3}$$

This improves the classical Wilkinson-type estimate $|\widehat{r} - x^{k+1}| \leqslant \gamma_k|x^{k+1}|$. They also note that for $k \approx \mathbf{u}^{-1}$ the relative error on $\widehat{r}$ can indeed be larger than $k\mathbf{u}$, thus suggesting that in the case of integer powers, the price to be paid for the refined constant $k\mathbf{u}$ is a necessary restriction on the range of k. This is in contrast with bounds like (1.2) and the results in [4,9], where restrictions on k can be avoided.

As Muller [8] mentioned, repeated multiplication may not be the method of choice to evaluate x^{k+1}. However, for better methods like binary exponentiation no improvement on the classical constant γ_k seems to be known.

In this note we generalize Theorem 1.1 to products of real and/or floating-point numbers, to any base, and to any evaluation scheme using k multiplications. Our restriction on k is weaker than the one in (1.3), though of the same order, and we show that it is essentially sharp.

Theorem 1.2 *Let $x_0, x_1, \ldots, x_k \in \mathbb{R}$ be given and suppose that ℓ of them are in $\mathbb{F}$. Let also*

$$K := 2k + 1 - \ell \qquad \text{and} \qquad \omega := \begin{cases} 1 & \text{if } \beta \text{ is odd,} \\ 2 & \text{if } \beta \text{ is even.} \end{cases} \tag{1.4}$$

Then, any order of evaluation of the product of $\prod_{i=0}^{k} \mathrm{fl}(x_i)$ produces an approximation $\widehat{r}$ such that, in absence of underflow and overflow,

$$\Big|\widehat{r} - \prod_{i=0}^{k} x_i\Big| \leqslant K\mathbf{u}\Big|\prod_{i=0}^{k} x_i\Big| \qquad \text{if} \quad K < \sqrt{\frac{\omega}{\beta}}\,\mathbf{u}^{-1/2}. \tag{1.5}$$

In particular, if $\beta = 2$ and all the x_i are in $\mathbb{F}$, then $(K, \omega) = (k, \beta)$ and (1.5) becomes

$$\Big|\widehat{r} - \prod_{i=0}^{k} x_i\Big| \leqslant k\mathbf{u}\Big|\prod_{i=0}^{k} x_i\Big| \qquad \text{if} \quad k < \mathbf{u}^{-1/2}. \tag{1.6}$$

For $\beta = 2$ and $p \geqslant 4$, the constraint in (1.6) cannot be replaced by $k < 12\mathbf{u}^{-1/2}$.

Remark. Note that for $\beta = 2$ and all the x_i in $\mathbb{F}$ the restriction $k < \mathbf{u}^{-1/2}$ improves on the restriction $k + 1 \leqslant \sqrt{2^{1/3} - 1} \cdot \mathbf{u}^{-1/2} = 0.509... \cdot \mathbf{u}^{-1/2}$ in (1.3).

The techniques to prove Theorem 1.2 can be used to obtain similar results for other evaluation schemes. As an example we show how to improve the classical factor γ_{2n} for Horner's scheme [2, p. 95].

Theorem 1.3 *Let $x, a_0, a_1, \ldots, a_n \in \mathbb{F}$ be given and let $\widehat{r}$ be the approximation to $\sum_{i=0}^{n} a_i x^i$ produced by Horner's scheme. Then, in absence of underflow and overflow,*

$$\Big|\widehat{r} - \sum_{i=0}^{n} a_i x^i\Big| \leqslant 2n\mathbf{u} \sum_{i=0}^{n} \big|a_i x^i\big| \qquad \text{if} \quad n < \frac{1}{2}\left(\sqrt{\frac{\omega}{\beta}}\,\mathbf{u}^{-1/2} - 1\right)$$

using ω defined in (1.4).

2 Products

We need some preliminaries to prove Theorem 1.2. If some x_i is zero, then $\widehat{r} = 0$ because no overflow occurs, and the results in Theorem 1.2 are trivial. If all the x_i are nonzero, then $\widehat{r} \neq 0$ because, by assumption, no underflow occurs. Furthermore, using $\mathbb{F} = -\mathbb{F}$ and $\mathrm{fl}(-x) = -\mathrm{fl}(x)$, we may henceforth assume without loss of generality that all the x_i are positive, so that all the $\widehat{r_i}$ are positive as well.

The standard estimate (1.1) can be improved in two ways. First, it is known that

$$x \in \mathbb{R}: \qquad |\mathrm{fl}(x) - x| \leqslant \frac{\mathbf{u}}{1 + \mathbf{u}}|x| \tag{2.1}$$

and that this bound is sharp; see for example [6, p. 232] and [5]. Second, we use the *unit in the first place* (ufp): a real number x being given, we set $\mathrm{ufp}(0) = 0$ and, if

$x \neq 0$, $\mathrm{ufp}(x) := \beta^{\lfloor \log_\beta |x| \rfloor}$. Thus, $\mathrm{ufp}(x)$ can be thought of as the weight of the first nonzero digit of x in its base-β representation. Then,

$$x \in \mathbb{R}: \qquad |\mathrm{fl}(x) - x| \leqslant \mathbf{u}\,\mathrm{ufp}(x). \tag{2.2}$$

This estimate is sharp as well; for more details, see [10]. Combining (2.1) and (2.2) yields the improved estimate

$$x \in \mathbb{R}\backslash\{0\}: \qquad \mathrm{fl}(x) = x(1+\varepsilon) \quad \text{with} \quad |\varepsilon| \leqslant \min\left[\frac{\mathbf{u}}{1+\mathbf{u}}, \mathbf{u}\frac{\mathrm{ufp}(x)}{|x|}\right]. \tag{2.3}$$

In the following we will use

$$x \in \mathbb{R}\backslash\{0\}: \qquad \mathrm{ufp}(x) \leqslant |x| < \beta\,\mathrm{ufp}(x), \tag{2.4}$$

as well as

$$\begin{aligned} f \in \mathbb{F} \cap [1,\beta] \quad &\Rightarrow \quad f = 1 + 2n\mathbf{u} \quad \text{with } n \in \mathbb{N}_0, \\ f \in \mathbb{F} \cap [\beta^{-1},1] \quad &\Rightarrow \quad f = 1 - \tfrac{2n}{\beta}\mathbf{u} \quad \text{with } n \in \mathbb{N}_0\,. \end{aligned} \tag{2.5}$$

Some notation is necessary to formalize the computation of the floating-point approximation $\widehat{r}$ in (1.5). The evaluation of $\prod_{i=0}^{k} \mathrm{fl}(x_i)$ in any given order by means of k floating-point multiplications is represented by a binary tree B whose $k+1$ leafs correspond to the $\mathrm{fl}(x_i)$ and whose k inner nodes correspond to the multiplications. Thus, B has $2k+1$ nodes N_i in total.

Since the order of evaluation is arbitrary, we may assume without loss of generality that $x_0, \ldots, x_L \in \mathbb{F}$ with $L := \ell - 1$. The numbering of the nodes shall be such that N_i corresponds to x_{i+L} for $i = -L, \ldots, k-L$, and $N_{k-L+1}, \ldots, N_K$ are the inner nodes. Moreover, N_K shall be the root of B.

Each node N_i is the root of a tree B_i and is identified with the floating-point value $\widehat{r_i} = \mathrm{fl}(r_i)$ computed by B_i. It follows in particular that $\widehat{r} = \widehat{r_K}$. More precisely, define $r_i := x_{i+L}$ for $i = -L, \ldots, k-L$ and, by means of a recursive definition, if an inner node N_i, $i \in \{k-L+1, \ldots, K\}$, has children N_{i_ν}, $1 \leqslant \nu \leqslant 2$, for which $\widehat{r_{i_1}}, \widehat{r_{i_2}}$ are already known, define $r_i := \widehat{r_{i_1}} \cdot \widehat{r_{i_2}}$. Since the x_i and $\widehat{r_i}$ have been assumed to be positive, the same holds for the r_i.

By assumption, $\widehat{r_i} = \mathrm{fl}(r_i) = x_{i+L}$ for $i = -L, \ldots, 0$. Moreover, for $i = 1, \ldots, K$ we have

$$\widehat{r_i} = \mathrm{fl}(r_i) =: (1+\varepsilon_i) r_i \qquad \text{with} \quad |\varepsilon_i| \leqslant \min\left[\frac{\mathbf{u}}{1+\mathbf{u}}, \mathbf{u}\frac{\mathrm{ufp}(r_i)}{r_i}\right] < \mathbf{u}. \tag{2.6}$$

For $i \in \{1, \ldots, k-L\}$, the relative errors ε_i correspond to the rounding of x_{i+L} into $\mathrm{fl}(x_{i+L})$, while for the remaining indices $i \in \{k-L+1, \ldots, K\}$ they correspond to the k multiplications. This implies $\prod_{i=0}^{k} \mathrm{fl}(x_i) = \prod_{i=1}^{k-L} (1+\varepsilon_i) \cdot \prod_{i=0}^{k} x_i$, and therefore

$$\widehat{r_K} - \prod_{i=0}^{k} x_i = \Big(\prod_{i=1}^{K} (1+\varepsilon_i) - 1 \Big) \cdot \prod_{i=0}^{k} x_i. \tag{2.7}$$

Since all factors x_i are positive, (1.5) is equivalent to $\left|\prod_{i=1}^{K}(1+\varepsilon_i)-1\right| \leqslant K\mathbf{u}$, and because $\prod_{i=1}^{K}(1+\varepsilon_i) \geqslant (1-\mathbf{u})^K \geqslant 1-K\mathbf{u}$ it suffices to prove

$$\prod_{i=1}^{K}(1+\varepsilon_i) \leqslant 1+K\mathbf{u}. \tag{2.8}$$

Hence, we need only upper bounds on the ε_i for the proof of Theorem 1.2.

Furthermore, the lemma below shows that, under weaker assumptions on the maximum K, the estimate (1.5) in Theorem 1.2 is true if a single ε_i is not positive, that is, if any of the $k-L$ real x_i or any single intermediate product is not rounded upwards. A similar observation was already made in [1, Lemma 3].

Lemma 2.1 *With the notation above, in particular (2.6), assume* $K \leqslant \sqrt{2}\,\mathbf{u}^{-1/2}$. *If there exists an index* $i \in \{1,\dots,K\}$ *with* $\varepsilon_i \leqslant 0$*, then (1.5) holds true.*

Proof. By (2.6) and (2.8), it suffices to show $Z := (1+\mathbf{u})^{K-1} \leqslant 1+K\mathbf{u}$. Using $K^2\mathbf{u} \leqslant 2$ gives

$$\ln(Z) = (K-1)\ln(1+\mathbf{u}) \leqslant (K-1)\mathbf{u} \leqslant K\mathbf{u} - \tfrac{1}{2}K^2\mathbf{u}^2 \leqslant \ln(1+K\mathbf{u}). \qquad \square$$

Proof of Theorem 1.2. With the notation above, in particular using (2.6), we have to prove (2.8). For $K \in \{0,1\}$ the assertion is trivial so that henceforth we assume $K \geqslant 2$. By Lemma 2.1 we can also assume that

$$\varepsilon_i > 0 \quad \text{for all } i \in \{1,\dots,K\}. \tag{2.9}$$

Let $\varphi \in \mathbb{N}$ be the largest integer satisfying

$$\varphi < \sqrt{\frac{\omega}{\beta}}\,\mathbf{u}^{-1/2}. \tag{2.10}$$

Note that $\varphi \geqslant 2$ because $2 \leqslant K < \sqrt{\omega/\beta}\,\mathbf{u}^{-1/2} \leqslant \varphi+1$. Define $I \subseteq \{1,\dots,K\}$ to be the index set with

$$i \in I \qquad :\Leftrightarrow \qquad \varepsilon_i > \frac{\mathbf{u}}{1+\varphi\mathbf{u}}. \tag{2.11}$$

The following two properties will be proved for distinct $i, j \in I$:

a) The nodes N_i and N_j are not adjacent in the tree B. (2.12)

b) The nodes N_i and N_j do not have the same parent node in B. (2.13)

Proof of (2.12). In order to derive a contradiction suppose that N_i is a child of N_j. It follows that $r_j = \widehat{r_i}\,\widehat{q}$, where $\widehat{q} \in \mathbb{F}$ is a (rounded) x_i or some intermediate result. If $\mathrm{ufp}(\widehat{r_i}) = \widehat{r_i}$, then $\widehat{r_i}$ is a power of β and $\varepsilon_j = 0$ contradicting (2.9), so that (2.6) and $i \in I$ imply

$$\mathrm{ufp}(r_i) = \mathrm{ufp}(\widehat{r_i}) < r_i < (1+\varphi\mathbf{u})\mathrm{ufp}(r_i) \qquad \text{for} \quad i \in I. \tag{2.14}$$

Since the second inequality is strict and $1+\varphi\mathbf{u} < 1+\sqrt{\mathbf{u}} < \beta$, it follows by (2.5), no matter whether φ is odd or even, that

$$\mathrm{ufp}(r_i) = \mathrm{ufp}(\widehat{r_i}) < \widehat{r_i} \leqslant (1+\varphi\mathbf{u})\mathrm{ufp}(r_i) \qquad \text{for} \quad i \in I. \tag{2.15}$$

By (2.15) and (2.5) we have

$$\widehat{r_i} = \mathrm{ufp}(\widehat{r_i})(1 + m\mathbf{u}) \quad \text{for even } m \in \mathbb{N} \ \text{ with } \ 2 \leqslant m \leqslant \varphi. \tag{2.16}$$

Hence, $r_j = \widehat{r_i}\,\widehat{q}$, (2.4), (2.16), $j \in I$, and (2.14) imply

$$\frac{R}{1 + m\mathbf{u}} \leqslant \frac{r_j}{\widehat{r_i}} = \widehat{q} \leqslant (1 + \varphi\mathbf{u})R \qquad \text{abbreviating } \ R := \frac{\mathrm{ufp}(r_j)}{\mathrm{ufp}(\widehat{r_i})}. \tag{2.17}$$

Since $\widehat{q} \in \mathbb{F}$, R is a power of β, and $R/(1 + m\mathbf{u}) > R(1 - m\mathbf{u}) \in \mathbb{F}$, (2.5) implies that there exists $\nu \in \mathbb{Q}$ such that

$$\widehat{q} = R(1 + \nu\mathbf{u}) \qquad \text{and} \qquad -m < \nu \leqslant \varphi. \tag{2.18}$$

Moreover, if ν is non-negative, then ν is a non-negative even integer by (2.5). From (2.18) and (2.16) we get $|\nu| \leqslant \varphi$. Now $r_j = \widehat{r_i}\,\widehat{q}$, (2.18), and (2.16) give

$$\mathrm{ufp}(r_j) \leqslant r_j = \mathrm{ufp}(r_j)(1 + (m + \nu)\mathbf{u} + m\nu\mathbf{u}^2), \tag{2.19}$$

and (2.14) together with $j \in I$ yields

$$0 \leqslant (m + \nu)\mathbf{u} + m\nu\mathbf{u}^2 \leqslant \varphi\mathbf{u}. \tag{2.20}$$

First, assume that ν is an even integer. Then, $m + \nu > 0$ is also even by (2.16), so that $1 + (m + \nu)\mathbf{u} \in \mathbb{F}$ and $|m\nu\mathbf{u}^2| \leqslant \varphi^2\mathbf{u}^2 < \mathbf{u}$ imply $\widehat{r_j} = \mathrm{ufp}(r_j)(1 + (m + \nu)\mathbf{u})$ and

$$\varepsilon_j = \frac{\widehat{r_j} - r_j}{r_j} = -\frac{\mathrm{ufp}(r_j)m\nu\mathbf{u}^2}{r_j} \leqslant \varphi|\nu|\mathbf{u}^2. \tag{2.21}$$

If $\nu \geqslant 0$, then $\varepsilon_j \leqslant 0$, a contradiction. Otherwise, (2.18) and $-\nu \in \mathbb{N}$ give $|\nu| = -\nu \leqslant m - 1 \leqslant \varphi - 1$, so that $\varphi < \mathbf{u}^{-1/2}$ implies

$$\varphi|\nu|\mathbf{u}^2(1 + \varphi\mathbf{u}) < \frac{1}{\sqrt{\mathbf{u}}}\left(\frac{1}{\sqrt{\mathbf{u}}} - 1\right)\mathbf{u}^2(1 + \sqrt{\mathbf{u}}) = (1 - \sqrt{\mathbf{u}})\mathbf{u}(1 + \sqrt{\mathbf{u}}) \leqslant \mathbf{u}.$$

Hence, $\varepsilon_j < \frac{\mathbf{u}}{1+\varphi\mathbf{u}}$ by (2.21), again a contradiction to $j \in I$ by (2.11).

Second, assume that ν is not an even integer. Then, (2.18) and (2.5) give $\nu < 0$. Write $\nu = 2n/\beta =: s + r/\beta$ with $n, s, r \in \mathbb{Z}_{\leqslant 0}$ with $|r| := (2|n|) \bmod \beta$. Since $2n$ is even, necessarily

$$|r| \leqslant \begin{cases} \beta - 2 & \text{if } \beta \text{ is even,} \\ \beta - 1 & \text{if } \beta \text{ is odd,} \end{cases} \qquad \Rightarrow \qquad \left|\frac{r}{\beta}\right| \leqslant 1 - \frac{\omega}{\beta} \tag{2.22}$$

using ω as in (1.4). In particular, for $\beta = 2$ this means $r = 0$. Now, (2.19) becomes

$$r_j = \mathrm{ufp}(r_j)(1 + (m + s + \delta)\mathbf{u}) \quad \text{with} \quad \delta := \frac{r}{\beta} + m\nu\mathbf{u} < 0 \tag{2.23}$$

because $r \leqslant 0$ and $-m \leqslant \nu < 0$. Using (2.22) and (2.10) we obtain

$$|\delta|\mathbf{u} \leqslant \left(1 - \frac{\omega}{\beta} + \varphi^2\mathbf{u}\right)\mathbf{u} < \mathbf{u}. \tag{2.24}$$

If s is odd, then $\delta < 0$ and (2.24) yield $\widehat{r_j} = \mathrm{ufp}(r_j)(1 + (m + s - 1)u)$ and $\varepsilon_j < 0$, a contradiction. If s is even, then $\widehat{r_j} = \mathrm{ufp}(r_j)(1 + (m + s)\mathbf{u})$ and

$$\varepsilon_j = -\delta\mathbf{u}\frac{\mathrm{ufp}(r_j)}{r_j} \leqslant |\delta|\mathbf{u}. \tag{2.25}$$

Note that s even implies $r \neq 0$ as ν is not an even integer.[1] By (2.18) we have $-m < \nu = s + r/\beta$. Since m, s are even integers and $r/\beta < 0$, it follows $-m+2 \leqslant s = \nu - r/\beta$, so that (2.22) yields

$$|\nu| = -\nu \leqslant m - 2 - \frac{r}{\beta} \leqslant \varphi - 1 - \frac{\omega}{\beta}. \tag{2.26}$$

From (2.25), (2.23), (2.22), (2.16), (2.26), and (2.10) we deduce the final contradiction to $j \in I$ and (2.11):

$$\frac{\varepsilon_j}{\mathbf{u}} \leqslant |\delta| \leqslant 1 - \frac{\omega}{\beta} + \varphi\left(\varphi - 1 - \frac{\omega}{\beta}\right)\mathbf{u} < 1 - \varphi\mathbf{u} - \frac{\omega}{\beta} + \varphi^2\mathbf{u} < 1 - \varphi\mathbf{u} < \frac{1}{1+\varphi\mathbf{u}}.$$

This finishes the proof of (2.12).

Proof of (2.13). Again, in order to derive a contradiction, assume that N_i and N_j are the left and right children of an inner node N_a, $a \in \{k-L+1, \ldots, K\}$, that is, $r_a = \widehat{r_i}\widehat{r_j}$ and $\widehat{r_a} = \mathrm{fl}(r_a)$. Then, like in the proof of (2.12), $i, j \in I$ implies

$$\begin{aligned}
\mathrm{ufp}(r_i) = \mathrm{ufp}(\widehat{r_i}) < r_i < \widehat{r_i} &= (1 + m\mathbf{u})\mathrm{ufp}(r_i) \leqslant (1 + \varphi\mathbf{u})\mathrm{ufp}(r_i),\\
\mathrm{ufp}(r_j) = \mathrm{ufp}(\widehat{r_j}) < r_j < \widehat{r_j} &= (1 + n\mathbf{u})\mathrm{ufp}(r_j) \leqslant (1 + \varphi\mathbf{u})\mathrm{ufp}(r_j)
\end{aligned}$$

with even $m, n \in \mathbb{N}_{\leqslant\varphi}$. Thus,

$$r_a = (1 + (m+n)\mathbf{u} + mn\mathbf{u}^2)\mathrm{ufp}(r_i)\mathrm{ufp}(r_j), \tag{2.27}$$

and $(m+n)\mathbf{u} \leqslant 2\varphi\mathbf{u} < 2\sqrt{\omega/\beta}\,\mathbf{u}^{1/2} \leqslant \frac{2\omega}{K\beta} \leqslant \frac{2}{K} \leqslant 1$ because $K \geqslant 2$. Moreover, $m+n$ is even and $mn\mathbf{u}^2 \leqslant \varphi^2\mathbf{u}^2 < \mathbf{u}$. Thus (2.27) yields $\mathrm{ufp}(r_a) = \mathrm{ufp}(\widehat{r_a}) = \mathrm{ufp}(r_i)\mathrm{ufp}(r_j)$, $\widehat{r_a} = (1 + (m+n)\mathbf{u})\mathrm{ufp}(\widehat{r_a})$, and $\varepsilon_a = -mn\mathbf{u}^2\mathrm{ufp}(\widehat{r_a})/r_a < 0$ contradicting (2.9). This finishes the proof of (2.13).

For I consisting of k' indices, (2.6) and (2.11) give

$$\prod_{i=1}^{K}(1+\varepsilon_i) \leqslant \left(1 + \frac{\mathbf{u}}{1+\mathbf{u}}\right)^{k'}\left(1 + \frac{\mathbf{u}}{1+\varphi\mathbf{u}}\right)^{K-k'}. \tag{2.28}$$

Using (2.12) and (2.13) we will show by Lemma 2.2 in Subsection 2.1 that $k' \leqslant \left\lfloor \frac{K+1}{2} \right\rfloor$. This implies

$$\prod_{i=1}^{K}(1+\varepsilon_i) \leqslant \left(1 + \frac{\mathbf{u}}{1+\mathbf{u}}\right)^{\left\lfloor \frac{K+1}{2} \right\rfloor}\left(1 + \frac{\mathbf{u}}{1+\varphi\mathbf{u}}\right)^{\left\lceil \frac{K-1}{2} \right\rceil}. \tag{2.29}$$

[1] Thus, for the classical case $\beta = 2$ a contradiction to $\{i, j\} \subseteq I$ is already obtained.

Hence, according to (2.8) and using $\frac{\mathbf{u}}{1+\mathbf{u}} \geqslant \frac{\mathbf{u}}{1+\varphi\mathbf{u}}$, the proof is finished if we show

$$F(K) := \left(1 + \frac{\mathbf{u}}{1+\mathbf{u}}\right)^{\frac{K+1}{2}} \left(1 + \frac{\mathbf{u}}{1+\varphi\mathbf{u}}\right)^{\frac{K-1}{2}} \leqslant 1 + K\mathbf{u}. \tag{2.30}$$

For later use, we do this by proving for real ψ the following stronger statement

$$G(\psi) := \left(1 + \frac{\mathbf{u}}{1+\mathbf{u}}\right)^{\frac{\psi+1}{2}} \left(1 + \frac{\mathbf{u}}{1+\psi\mathbf{u}}\right)^{\frac{\psi-3}{2}} \leqslant 1 + (\psi-1)\mathbf{u} \tag{2.31}$$

provided that $1 \leqslant \psi \leqslant \sqrt{\frac{\omega}{\beta}}\,\mathbf{u}^{-1/2}$. If this is true, then for $1 \leqslant K \leqslant \varphi$ we obtain

$$F(K) \leqslant G(K)\left(1 + \frac{\mathbf{u}}{1+\varphi\mathbf{u}}\right) \leqslant (1 + (K-1)\mathbf{u})\left(1 + \frac{\mathbf{u}}{1+(K-1)\mathbf{u}}\right) = 1 + K\mathbf{u}$$

which is (2.30). A computation yields the Taylor expansion

$$G(\psi) = 1 + (\psi-1)\mathbf{u} + \frac{1}{2}G''(\xi)\xi^2 \qquad \text{with} \quad G''(\xi) =: \alpha N(\xi)$$

for some $0 < \xi < \mathbf{u}$ and

$$\alpha := -\frac{(\psi-1)\left(\frac{1+(\psi+1)\xi}{1+\psi\xi}\right)^{\frac{\psi-1}{2}}\left(1+\frac{\xi}{1+\xi}\right)^{\frac{\psi-1}{2}}}{4\,(1+2\xi)\,(1+\xi)^3\,(1+(\psi+1)\xi)^3\,(1+\psi\xi)} < 0.$$

It suffices to show $N(\xi) \geqslant 0$ for $0 < \xi < \mathbf{u}$. Now $N(\xi) = \sum_{\nu=0}^{5} c_\nu$ and $\psi^2\mathbf{u} \leqslant 1$ with

$$\begin{aligned}
c_0 &= 60\,\xi^4 + 160\,\xi^3 + 144\,\xi^2 + 48\,\xi + 4 &&> 4 + 48\,\xi \\
c_1 &= \left(48\,\xi^5 + 192\,\xi^4 + 248\,\xi^3 + 124\,\xi^2 + 20\,\xi\right)\psi &&> 20\,\psi\,\xi + 124\,\psi\,\xi^2 \\
c_2 &= \left(72\,\xi^5 + 187\,\xi^4 + 140\,\xi^3 + 24\,\xi^2 - 4\xi\right)\psi^2 &&> -4 \\
c_3 &= \left(32\,\xi^5 + 41\,\xi^4 - 8\,\xi^2\right)\psi^3 &&> -8\,\psi\,\xi \\
c_4 &= \left(8\,\xi^5 + \xi^4 - 4\,\xi^3\right)\psi^4 &&> -4\,\xi \\
c_5 &= -\,\xi^4\,\psi^5 &&> -\psi\,\xi^2.
\end{aligned}$$

The series expansions were computed by the Symbolic Math Toolbox of MATLAB [7]. It follows $N(\xi) > 0$ for $0 < \xi < \mathbf{u}$, and this proves (1.5) and (1.6).

The assertion on possible constraints of k is deferred to the appendix. This finishes the proof of Theorem 1.2. □

In the proof of Theorem 1.2 we defined φ to be the largest integer less than $\sqrt{\omega/\beta}\,\mathbf{u}^{-1/2}$, which reduces to $\varphi < \mathbf{u}^{-1/2}$ for binary arithmetic. Switching from binary arithmetic to another basis requires indeed an adapted definition of φ. Consider $p := 5$ decimal digits, that is, $\mathbf{u} = 0.5 \cdot 10^{-4}$. Then, $\widehat{r_i} := \mathrm{fl}(1.3033 \cdot 0.7697) = 1.0032$ and $\widehat{q} := 0.99696$ yield $\widehat{r_j} = 1.0002$. Moreover, $\varphi = 63$ whilst the largest integer less than $\mathbf{u}^{-1/2}$ is $\varphi' = 141$. However, both ε_i and ε_j would satisfy (2.11) if φ was replaced by φ', and indices of adjacent nodes would belong to I.

2.1 A result on colored trees

In (2.29) in the proof of Theorem 1.2 we used the upper bound $\lfloor \frac{K+1}{2} \rfloor$ for the number k' of nodes in the index set I. This bound is a consequence of the following lemma.

Lemma 2.2 *Let T be a tree with M nodes, each having at most two children. Assume that C nodes of T are colored according to the following rules:*

(i) *colored nodes are not adjacent;*
(ii) *each node has at most one colored child.*

Then,

$$C \leqslant \begin{cases} \left\lfloor \frac{M+1}{2} \right\rfloor & \text{if the root of } T \text{ is colored,} \\ \left\lfloor \frac{M}{2} \right\rfloor & \text{otherwise.} \end{cases}$$

Furthermore, these inequalities are sharp for all M.

Proof The result is trivial for $M = 1$, so assume $M \geqslant 2$ and that the result is true up to $M - 1$. The root R of T is then connected to a tree T_1 and, possibly, also to another tree T_2 disjoint from T_1. Let T_1 have M_1 nodes, C_1 of which being colored. Define M_2 and C_2 similarly if T_2 exists, and let $C_2 = M_2 = 0$ otherwise. Clearly, $M = M_1 + M_2 + 1$ and $0 \leqslant C_i \leqslant M_i \leqslant M - 1$ for $i = 1, 2$.

If R is colored, then $C = C_1 + C_2 + 1$ and (i) implies the root of T_1 is not colored. Hence, by induction, $C_1 \leqslant \lfloor M_1/2 \rfloor \leqslant M_1/2$. Similarly, $C_2 \leqslant M_2/2$, so that

$$C \leqslant \frac{M_1}{2} + \frac{M_2}{2} + 1 = \frac{M+1}{2}.$$

If R is not colored, then $C = C_1 + C_2$ and (ii) implies that R has at most one colored child. Hence, for M_2 either zero or nonzero,

$$C \leqslant \frac{M_1}{2} + \frac{M_2}{2} + \frac{1}{2} = \frac{M}{2}.$$

Since C is an integer, the claimed bounds follow for $M \geqslant 2$. Finally, trees with all internal nodes having exactly one child ("linked lists") and whose colored and uncolored nodes alternate show that the bound is attained for any M. □

Now, the upper bound for k' in the proof of Theorem 1.2 is obtained as follows. First, we construct a tree T by removing from the binary tree B the leafs $N_{-L}, \ldots, N_0$ associated with the ℓ operands x_i already in $\mathbb{F}$. The nodes of T are the nodes $N_1, \ldots, N_K$ of B, and the nodes N_i with $i \in I$ are considered as colored. Then, (2.12) and (2.13) imply that T follows the rules (i) and (ii) of Lemma 2.2, so that $|I| = k' \leqslant \left\lfloor \frac{K+1}{2} \right\rfloor$.

Optimality of the bounds in Lemma 2.2 is established by linked lists which represent recursive multiplication of floating-point numbers. We note that optimal bounds are attained for other evaluation schemes as well. Examples for all M for trees with colored root are sketched in Figure 2.1; examples with uncolored root follow similarly.

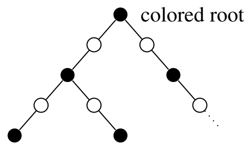

Fig. 2.1 Trees attaining the bound $C = \left\lfloor \frac{M+1}{2} \right\rfloor$ for colored root.

3 Horner scheme

Using the techniques of the previous section we prove Theorem 1.3. For $n = 0$ the assertion is trivial so that we may assume $n \geqslant 1$. The Horner scheme computes

$$\widehat{r}_0 := \mathrm{fl}(a_n x); \qquad \widehat{r}_i := \mathrm{fl}(\mathrm{fl}(\widehat{r}_{i-1} + a_{n-i})x), \quad i = 1, \ldots, n-1; \qquad \widehat{r} = \widehat{r}_n := \mathrm{fl}(\widehat{r}_{n-1} + a_0).$$

For $i = 1, \ldots, n$, let the relative error of the i-th addition and multiplication be denoted by ε_i and ε'_{i-1}, respectively. Then,

$$\begin{aligned} \widehat{r}_0 &= a_n x(1 + \varepsilon'_0), \\ \widehat{r}_i &= (\widehat{r}_{i-1} + a_{n-i})x(1 + \varepsilon_i)(1 + \varepsilon'_i), \quad i = 1, \ldots, n-1, \\ \widehat{r} &= (\widehat{r}_{n-1} + a_0)(1 + \varepsilon_n). \end{aligned} \tag{3.1}$$

For each $i \in \{1, \ldots, n-1\}$ we apply Theorem 1.2 to the product $x_0 x_1$ with $x_0 := \widehat{r}_{i-1} + a_{n-i} \in \mathbb{R}$ and $x_1 := x \in \mathbb{F}$. Then, $k = 1$, $\ell = 1$ and therefore $K = 2$, so that (2.29) with the constant φ defined in (2.10) yields[2]

$$(1 + \varepsilon_i)(1 + \varepsilon'_i) \leqslant \left(1 + \frac{\mathbf{u}}{1 + \mathbf{u}}\right)\left(1 + \frac{\mathbf{u}}{1 + \varphi\mathbf{u}}\right), \quad i = 1, \ldots, n-1. \tag{3.2}$$

Furthermore, (2.6) gives

$$(1 + \varepsilon'_0)(1 + \varepsilon_n) \leqslant \left(1 + \frac{\mathbf{u}}{1 + \mathbf{u}}\right)^2. \tag{3.3}$$

From the equalities in (3.1) we deduce that $\widehat{r} = \sum_{i=0}^{n} a_i(1 + \alpha_i)x^i$, where

$$\begin{aligned} 1 + \alpha_n &= (1 + \varepsilon'_0) \cdot \textstyle\prod_{j=1}^{n-1}(1 + \varepsilon_j)(1 + \varepsilon'_j) \cdot (1 + \varepsilon_n), \\ 1 + \alpha_i &= \textstyle\prod_{j=n-i}^{n-1}(1 + \varepsilon_j)(1 + \varepsilon'_j) \cdot (1 + \varepsilon_n), \quad i = 1, \ldots, n-1, \\ 1 + \alpha_0 &= 1 + \varepsilon_n. \end{aligned}$$

Hence, (1.1), (3.2) and (3.3) imply

$$(1 - \mathbf{u})^{2n} \leqslant 1 + \alpha_n \leqslant \left(1 + \frac{\mathbf{u}}{1 + \mathbf{u}}\right)^{n+1} \left(1 + \frac{\mathbf{u}}{1 + \varphi\mathbf{u}}\right)^{n-1} =: H_n$$

[2] In fact, (2.29) is applied to $|x_0|, |x_1|$ because the proof of Theorem 1.2 assumes positive factors.

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

$$(1-\mathbf{u})^{2i+1} \leqslant 1+\alpha_i \leqslant \left(1+\frac{\mathbf{u}}{1+\mathbf{u}}\right)^{i+1}\left(1+\frac{\mathbf{u}}{1+\varphi\mathbf{u}}\right)^{i}.$$

Then, using $1-2n\mathbf{u} < (1-\mathbf{u})^{2n}$, we see that $1-2n\mathbf{u} \leqslant 1+\alpha_i \leqslant H_n$ for all $i = 0, 1, \ldots, n$. The assumption $n < \frac{1}{2}\left(\sqrt{\frac{\omega}{\beta}}\,\mathbf{u}^{-1/2} - 1\right)$ implies $2n+1 \leqslant \varphi$. Thus, (2.31) proves $H_n \leqslant G(2n+1) \leqslant 1+2n\mathbf{u}$. □

We close this note with an application of Theorem 1.2.

Corollary 3.1 (Evaluation of a polynomial given by its roots)
Given $z, z_1, \ldots, z_n, a_n \in \mathbb{F}$, let $\widehat{r} \in \mathbb{F}$ be a floating-point approximation to

$$r = a_n \prod_{i=1}^{n} (z - z_i)$$

obtained by first evaluating the n differences and then, in any order, a product of $n+1$ terms. If $n < \frac{1}{2}\sqrt{\frac{\omega}{\beta}}\,\mathbf{u}^{-1/2}$ then, in absence of underflow and overflow,

$$|\widehat{r} - r| \leqslant 2n\mathbf{u}|r|.$$

Proof Define $x_0 := a_n \in \mathbb{F}$ and $x_i := z - z_i \in \mathbb{R}$ for $i = 1, \ldots, n$. Then, Theorem 1.2 with $k = n$, $\ell = 1$, $K = 2k+1-\ell = 2n < \sqrt{\frac{\omega}{\beta}}\,\mathbf{u}^{-1/2}$ yields the assertion. □

4 Appendix

The goal of this appendix is to prove that for $\beta = 2$ and $p \geqslant 4$ the constraint $k < \mathbf{u}^{-1/2}$ in Theorem 1.2 cannot be replaced by $k < 12\mathbf{u}^{-1/2}$. To do that[3] we construct $x_0, x_1, x_2 \in \mathbb{F}$ for given precision p such that $x_1x_2 < 1$ and $\mathrm{fl}(\mathrm{fl}(x_0x_1)x_2) = x_0$. Subsequent multiplications by x_1x_2 produce an exponential growth of the rounding error, eventually exceeding $k\mathbf{u}$.

Define $s := \lfloor \mathbf{u}^{-1/2} \rfloor \in \mathbb{N}$, so that $s = \mathbf{u}^{-1/2} - \delta$ with $0 \leqslant \delta < 1$. We henceforth assume $p \geqslant 15$ and treat the case $p \leqslant 14$ later. Note that $\beta = 2$ and $p \geqslant 15$ imply $s \geqslant 181$. We distinguish two cases.

First, assume s is odd. Set

$$x_0 := 1 + (2s+8)\mathbf{u}, \quad x_1 := 1 - (s-4)\mathbf{u}, \quad \text{and} \quad x_2 := 1 + (s-5)\mathbf{u},$$

so that $x_i \in \mathbb{F}$. Then, $x_0x_1 = 1 + (s+10)\mathbf{u} + \mu_1\mathbf{u}$ with $\mu_1 := 4\delta\sqrt{\mathbf{u}} + (32 - 2\delta^2)\mathbf{u}$, so that $0 < \mu_1 < 1$ and s odd imply $\mathrm{fl}(x_0x_1) = 1 + (s+11)\mathbf{u}$. Moreover, $\mathrm{fl}(x_0x_1)x_2 = 1 + (2s+7)\mathbf{u} + \mu_2\mathbf{u}$ with

$$\mu_2 := \sqrt{\mathbf{u}}\left(6 - 55\sqrt{\mathbf{u}} + \Phi\delta\right) \quad \text{with} \quad \Phi := (\delta - 6)\sqrt{\mathbf{u}} - 2.$$

[3] In [1] long sequences $x_i \in \mathbb{F}$ with $\mathrm{fl}((\ldots(\mathrm{fl}(x_0x_1)x_2)\ldots)x_k) = x_0$ are constructed for some precisions.

Now $\Phi < 0$ for any value of δ, so that $0 < 4\sqrt{\mathbf{u}} - 60\mathbf{u} \leqslant \mu_2 \leqslant 6\sqrt{\mathbf{u}} - 55\mathbf{u} < 1$. Thus,

$$\mathrm{fl}(\mathrm{fl}(x_0 x_1)x_2) = x_0. \tag{4.1}$$

Define a vector $X := [x_0\ x\ x \ldots x] \in \mathbb{F}^{2m+1}$ with m times repeating the row vector $x = [x_1\ x_2] \in \mathbb{F}^2$. Denoting $\widehat{r_0} := x_0$ and $\widehat{r_i} := \mathrm{fl}(\widehat{r_{i-1}}X_i)$ for $i \geqslant 1$ yields $\widehat{r_2} = v_0$. Then, abbreviating $\pi := x_1 x_2$ and using $\widehat{r_{2m}} = \widehat{r_2} = x_0$ gives

$$\widehat{r_{2m}} - \prod_{i=0}^{2m} X_i = x_0 - x_0\pi^m = (\pi^{-m} - 1)\prod_{i=0}^{2m} X_i \qquad \text{for} \quad 1 \leqslant m \in \mathbb{N}. \tag{4.2}$$

Now,

$$\pi = 1 - \left(2 - (9 + 2\delta)\sqrt{\mathbf{u}}\right)\mathbf{u} - (20 + 9\delta + \delta^2)\mathbf{u}^2 < 1 - \left(2 - 11\sqrt{\mathbf{u}}\right)\mathbf{u} =: 1 - \gamma\mathbf{u},$$

and for $m \in \mathbb{N}$,

$$\pi^{-m} > 1 + m\gamma\mathbf{u} + \frac{m(m-1)}{2}\gamma^2\mathbf{u}^2 = 1 + 2m\mathbf{u} + \frac{m\mathbf{u}\sqrt{\mathbf{u}}}{2}\left[(m-1)\gamma^2\sqrt{\mathbf{u}} - 22\right].$$

The assumption $p \geqslant 15$ implies

$$(6 - 2\sqrt{\mathbf{u}})\gamma^2 - 22 = 2 - 272\sqrt{\mathbf{u}} + (814 - 242\sqrt{\mathbf{u}})\mathbf{u} > 2 - 272\sqrt{\mathbf{u}} > 0,$$

and therefore

$$m \geqslant 6\mathbf{u}^{-1/2} - 1 \quad \Rightarrow \quad \pi^{-m} > 1 + 2m\mathbf{u}. \tag{4.3}$$

Combining this with (4.2) shows that the error bound in (1.6) is not satisfied for $k = 2\left\lceil 6\mathbf{u}^{-1/2} - 1\right\rceil < 12\mathbf{u}^{-1/2}$, and that finishes the first part.

Second, assume s is even and define as before

$$y_0 := 1 + (2s + 6)\mathbf{u}, \quad y_1 := 1 - (s - 3)\mathbf{u}, \quad \text{and} \quad y_2 := 1 + (s - 4)\mathbf{u}. \tag{4.4}$$

Then, $y_i \in \mathbb{F}$. Furthermore, $y_0 y_1 = 1 + (s + 7)\mathbf{u} + \mu_1\mathbf{u}$ with $\mu_1 := 4\delta\sqrt{\mathbf{u}} + (18 - 2\delta^2)\mathbf{u}$, so that $0 < \mu_1 < 1$ and s even imply $\mathrm{fl}(y_0 y_1) = 1 + (s + 8)\mathbf{u}$. Moreover, $\mathrm{fl}(y_0 y_1)y_2 = 1 + (2s + 5)\mathbf{u} + \mu_2\mathbf{u}$ with

$$\mu_2 := \sqrt{\mathbf{u}}\left(4 - 32\sqrt{\mathbf{u}} + \Phi\delta\right) \quad \text{with} \quad \Phi := (\delta - 4)\sqrt{\mathbf{u}} - 2.$$

As before, $\Phi < 0$ for any value of δ. Thus, $0 < 2\sqrt{\mathbf{u}} - 35\mathbf{u} \leqslant \mu_2 \leqslant 4\sqrt{\mathbf{u}} - 32\mathbf{u} < 1$. Hence, similar to (4.1), $\mathrm{fl}(\mathrm{fl}(y_0 y_1)y_2) = y_0$ is again true. Now for the values y_1, y_2 in (4.4) we obtain

$$y_1 y_2 = (1 - (s - 3)\mathbf{u})(1 + (s - 4)\mathbf{u}) < x_1 x_2,$$

and the result follows as before. Finally, for the cases $4 \leqslant p \leqslant 14$, consider

p	m_0	m_1	m_2	F
4	2	-4	4	9.6
5	20	-3	2	8.9
6	32	-14	16	5.8
7	28	-9	8	6.8
8	52	-39	44	5.8
9	48	-21	20	4.6
10	140	-117	130	5.2
11	94	-43	42	5.8
12	186	-154	158	4.0
13	184	-89	88	4.1
14	262	-125	124	7.2

For precision p define $x_i := 1 + m_i\mathbf{u}$. Then, (4.1) is satisfied, and the error bound in (1.6) is not true for $k < F\mathbf{u}^{-1/2}$. This finishes the proof. □

We finally mention that it is easy to see that, if $1 \leqslant p \leqslant 2$, then the error bound in (1.6) is satisfied for all $k \in \mathbb{N}$, and if $p = 3$, then the minimum value of k for which it is not satisfied is $k = 72 \approx 25\mathbf{u}^{-1/2}$.

5 Summary

In previous papers, the factor γ_k has been replaced by $k\mathbf{u}$ in a number of classical error estimates in numerical analysis together with removing the restriction on k. We proved that $k\mathbf{u}$ can be used for general products and for the Horner scheme, however, with a mandatory restriction on k. So, as by Theorem 1.2, a general principle to replace γ_k by $k\mathbf{u}$ is necessarily restricted to $k \lesssim \mathbf{u}^{-1/2}$.

6 Acknowledgement

The authors wish to thank Marko Lange, Vincent Lefèvre, and Jean-Michel Muller for their fruitful and constructive comments on a preliminary version of this note.

References

1. S. Graillat, V. Lefèvre, and J.-M. Muller. On the maximum relative error when computing integer powers by iterated multiplications in floating-point arithmetic. Research report ensl-00945033, September 2014. Available from `https://hal-ens-lyon.archives-ouvertes.fr/ensl-00945033v2`.
2. N.J. Higham. *Accuracy and Stability of Numerical Algorithms*. SIAM Publications, Philadelphia, 2nd edition, 2002.
3. IEEE, New York. *ANSI/IEEE 754-2008: IEEE Standard for Floating-Point Arithmetic*, 2008.
4. C.-P. Jeannerod and S.M. Rump. Improved error bounds for inner products in floating-point artihmetic. *SIAM. J. Matrix Anal. & Appl. (SIMAX)*, 34(2):338–344, 2013.
5. C.-P. Jeannerod and S.M. Rump. On relative errors of floating-point operations: optimal bounds and applications. Preprint, January 2014.
6. D.E. Knuth. *The Art of Computer Programming, Volume 2, Seminumerical Algorithms.* Third edition, Addison-Wesley, Reading, Massachusetts, 1998.
7. MATLAB. User's Guide, Version 2013b, the MathWorks Inc., 2013.

8. J.M. Muller. On the maximum relative error when computing iterated integer powers in floating-point arithmetic. INVA conference Tokyo, 2014.
9. S.M. Rump and C.-P. Jeannerod. Improved error bounds for LU and Cholesky factorizations. *SIAM. J. Matrix Anal. & Appl. (SIMAX)*, 35(2):699–724, 2014.
10. S.M. Rump, T. Ogita, and S. Oishi. Accurate floating-point summation, Part I: Faithful rounding. *SIAM J. Sci. Comput.*, 31(1):189–224, 2008.