Sharp estimates for perturbation errors in summations *

Marko Lange † Siegfried M. Rump ‡

Abstract

Standard Wilkinson-type error estimates of floating-point algorithms that are solely based on the first or second standard model typically involve a factor $\gamma_k := k\mathbf{u}/(1-k\mathbf{u})$, where $\mathbf{u}$ denotes the relative rounding error unit of a floating-point number system. Using specific properties of floating-point grids it was shown that often γ_k can be replaced by $k\mathbf{u}$, and the restriction on k can be removed. That is true for standard algorithms such as summation, dot product, matrix multiplication, LU- or Cholesky decomposition, et cetera.

Recently it was shown that, at least for summation and dot product, such results derive without any reference to a floating-point grid. In the current paper we further sharpen the error estimate for summation into $k\mathbf{u}/(1+k\mathbf{u})$, again without any reference to a floating-point grid. Furthermore, an estimate of type $h\mathbf{u}$ is shown for sums and dot products that are evaluated using a binary tree of height h. Both estimates require a mandatory restriction of size $1/\mathbf{u}$ on the number of summands and the height, respectively.

Finally, a different kind of error estimate is shown for recursive summation. The discussed bound is sharp, holds true for any number of summands, and is uniformly bounded by 1.

The novelty of our approach is two-fold. First, rather than using a rounding function, the discussed estimates are based on almost arbitrary perturbations of real operations without any reference to a floating-point grid. As a consequence, the corresponding floating-point error bounds in some base β for rounding to nearest, and partly also for directed roundings, follow as corollaries. Secondly, in regard to our weak assumptions, the new estimates are sharp. Our main result is sharp for actual realizations of grids floating-point arithmetics are based on. To be more specific, for any feasible problem size, for IEEE 754 binary32 as well as binary64 format, there are examples satisfying the given bound with equality.

*This research was partially supported by CREST, Japan Science and Technology Agency (JST).

†Waseda University, Faculty of Science and Engineering, 3–4–1 Okubo, Shinjuku-ku, Tokyo 169–8555, Japan (m.lange@aoni.waseda.jp).

‡Institute for Reliable Computing, Hamburg University of Technology, Am Schwarzenberg-Campus 1, Hamburg 21071, Germany, and Visiting Professor at Waseda University, Faculty of Science and Engineering, 3–4–1 Okubo, Shinjuku-ku, Tokyo 169–8555, Japan (rump@tuhh.de).

1 Introduction

For a set of floating-point numbers $\mathbb{F} \subseteq \mathbb{R}$, many floating-point error estimates are based on the first standard model [1]

$$a, b \in \mathbb{F}: \quad a \tilde{+} b = (a+b)(1+\varepsilon) \quad \text{for some} \quad |\varepsilon| \le \mathbf{u}, \tag{1}$$

where $0 < \mathbf{u} \in \mathbb{R}$ is a constant associated to $\mathbb{F}$, and $\tilde{+}$ denotes an addition on $\mathbb{F}$. Typically, $\mathbf{u}$ is referred to as the relative rounding error unit.

A direct consequence for the error of the result s_n of recursive summation

$$s_1 := x_1; \quad s_{i+1} = s_i \tilde{+} x_{i+1} = (s_i + x_{i+1})(1+\varepsilon_i) \quad \text{for} \;\; 1 \le i \le n-1 \tag{2}$$

of a vector $x \in \mathbb{F}^n$ is the standard error estimate [1]

$$\left| s_n - \sum_{i=1}^{n} x_i \right| \le \left((1+\mathbf{u})^{n-1} - 1 \right) \sum_{i=1}^{n} |x_i|. \tag{3}$$

That estimate is true in the absence of a floating-point grid. In fact, not even $a \tilde{+} b = c \tilde{+} d$ for $a + b = c + d$ is necessary. Moreover, the estimate is derived without any consideration of the rounding mode.

Without additional assumptions, the factor $(1+\mathbf{u})^{n-1} - 1$ in (3) cannot be replaced by $(n-1)\mathbf{u}$. As an example consider a logarithmic number system $\mathbb{F} := \{\pm c^k \colon k \in \mathbb{Z}\}$ for $1 < c \in \mathbb{R}$ with rounding upwards. Then $\mathbf{u} = \frac{c-1}{c+1}$ and, for sufficiently small $e \in \mathbb{F}$, $(1 \tilde{+} e) \tilde{+} e = c^2$ but $c^2 - (1+2e) > 2\frac{c-1}{c+1}(1+2e)$. The reason is that an arbitrary small summand e causes a relative error of almost size $\mathbf{u}$.

In a recent paper [4] we unveiled the necessary properties of floating-point schemes that allow for linearly growing a priori error estimates such as the ones for LU- and Cholesky factorizations in [7]. More specifically, we showed that the implication

$$a \tilde{+} b = a + b + \delta \quad \Longrightarrow \quad |\delta| \le \min(|a|, |b|) \tag{4}$$

is sufficient for the estimate

$$\left| s_n - \sum_{i=1}^{n} x_i \right| \le \sum_{i=1}^{n-1} |\varepsilon_i| \sum_{i=1}^{n} |x_i| \le (n-1)\mathbf{u} \sum_{i=1}^{n} |x_i|, \tag{5}$$

which is true without restriction on n and for any order of summation. It is straightforward to show that the assumption (4) is satisfied for any rounding to nearest, see (7) in the next section. In fact, the assumptions in [4] are even more general.

In the present paper, we define quantities ξ_i similar to ε_i and prove

$$\left| s_n - \sum_{i=1}^{n} x_i \right| \le \frac{\sum_{i=1}^{n-1} \xi_i}{1 + \sum_{i=1}^{n-1} \xi_i} \sum_{i=1}^{n} |x_i|.$$

Since these ξ_i are nonnegative and less than or equal to $\mathbf{u}$, this implies

$$\left| s_n - \sum_{i=1}^{n} x_i \right| \leq \frac{(n-1)\mathbf{u}}{1+(n-1)\mathbf{u}} \sum_{i=1}^{n} |x_i|,$$

which was first proved for recursive summation by Mascarenhas[1] [5], provided that $20(n-1)\mathbf{u} \leq 1$. We will show that our estimate is true under very weak assumptions on the perturbations of real sums. Moreover, it is true for any order of summation and no floating-point grid is necessary. For addition in rounding to nearest according to IEEE 754 the estimate is sharp.

Besides the improved and sharp bound, another target of the paper is to identify as weak as possible assumptions for its validity.

For binary summation in a tree of height h standard error analysis involves a factor $(1+\mathbf{u})^h - 1$. We show under very general assumptions that this factor can be replaced by $h\mathbf{u}$. A restriction of the height is mandatory, which, for floating-point number systems, is of the order $\mathbf{u}^{-\frac{1}{2}}$. Note that for IEEE 754 single precision (binary32) that imposes the restriction $n < 1.04 \cdot 10^{1233}$ on the number of summands.

Again we want to stress that besides the new estimate a major goal is to identify the necessary assumptions on an approximate addition $\tilde{+}$. In particular, the new estimate for binary summation is even true for IEEE 754 addition with directed rounding.

Finally we give another estimate for the error of summation in recursive order, being true without restriction on n. For increasing n, the corresponding factor in the error estimate converges to 1 from below. This property is peculiar for recursive summation and is not true for evaluations in arbitrary order. The corresponding assumptions are again very general, and the estimate is sharp.

The paper is organized as follows. For didactic reasons we first state in the next section the conclusions for floating-point arithmetic with given precision and general basis. That applies for instance to any arithmetic adhering to IEEE 754. We hope that motivates to go through the proofs of our new results presented in Sections 3 to 5 covering general summation, binary summation, and again a different estimate for recursive summation, respectively. We choose that because it may not be clear beforehand that the assumptions of our general theorems are indeed satisfied for floating-point number systems. Directly following the general results in each section, we prove the corresponding conclusions presented in Section 2.

Finally, we show some possible applications of the new bounds. The improved constants eventually lead to nice constants for the sum of products, for example when multiplying a vector by a Vandermonde matrix. Moreover, a new estimate for blocked summation is presented.

[1] In [5], the author introduces a new concept of using continuous mathematics to analyze floating-point arithmetic. However, the given proof is rather complicated and longish. Here we generalize his result using comparatively simple arguments.

2 Floating-point results

In this section we show implications of the subsequently presented, general perturbation estimates for the summation in some m-digit floating-point number system in base β following the IEEE 754 standard. The relative rounding error unit to such a system is $\mathbf{u} := 0.5\beta^{1-m}$.

In particular, we consider sums evaluated in *rounding to nearest* or in some *faithful rounding*. An addition $\tilde{+}$ on $\mathbb{F}$ is called a sum in rounding to nearest if

$$|(a+b) - (a \tilde{+} b)| = \inf\{|(a+b) - f| \colon f \in \mathbb{F}\}. \tag{6}$$

In this case (1) is true with replacing $\mathbf{u}$ by $\mathbf{u}/(1+\mathbf{u})$, see [3]. As mentioned before, summing in rounding to nearest implies (4) by

$$\begin{aligned} |(a \tilde{+} b) - (a+b)| &= \inf\left\{|f - (a+b)| \colon f \in \mathbb{F}\right\} \\ &\le \min\left\{|f - (a+b)| \colon f \in \{a, b\}\right\} \\ &= \min\left\{|a|, |b|\right\}. \end{aligned} \tag{7}$$

A result is rounded faithfully if, for all $a, b \in \mathbb{F}$, $a \tilde{+} b$ is the only floating-point number in the convex hull $(a+b) \underline{\cup} (a \tilde{+} b)$. The latter means that there is no other floating-point number between the actual and the computed result. In that case (1) is true with replacing $\mathbf{u}$ by $2\mathbf{u}$.

Proposition 1. *Let s be the result of a floating-point summation of $a_1, \ldots, a_n \in \mathbb{F}$ in some nearest-addition in arbitrary order. If $n \le 1 + \frac{\beta-1}{2}\mathbf{u}^{-1}$, then*

$$\left| s - \sum_{j=1}^{n} a_j \right| \le \frac{(n-1)\mathbf{u}}{1 + (n-1)\mathbf{u}} \sum_{j=1}^{n} |a_j|. \tag{8}$$

Remark 2. The estimate in Proposition 1 is sharp. Using IEEE 754 in basis $\beta = 2$, rounding to nearest, and tie-breaking rule *round half to even*, the recursive summation of the summands $1, \mathbf{u}, \mathbf{u}, \ldots, \mathbf{u}$ satisfies (8) with equality. Moreover, the upper bound on n is mandatory. For the tie-breaking rule *round half away from zero* the upper bound on n cannot be replaced by the next larger integer. Similarly, for tie to even this restriction could be improved only by a relatively small number. This statement is shown by the following example for an even basis β:

$$s = 1 \tilde{+} \underbrace{t \tilde{+} \mathbf{u} \tilde{+} t \tilde{+} \mathbf{u} \tilde{+} \ldots \tilde{+} t \tilde{+} \mathbf{u}}_{\frac{\beta-1}{2\mathbf{u}} \text{ summands}} \tilde{+} \underbrace{\beta t \tilde{+} \beta\mathbf{u} + \beta t \tilde{+} \beta\mathbf{u} + \ldots}_{\frac{\beta}{2}+1 \text{ summands}},$$

where $t := \mathbf{u} + 2\mathbf{u}^2$. It is straightforward to check that for any mantissa length $m \ge 3$ the error in this recursive floating-point sum does not satisfy estimate (8). Hence, the tolerance to the restriction of n is not greater than $\frac{\beta}{2}$.

Proposition 3. *Let s be the result of a floating-point summation of $a_1, \dots, a_n \in \mathbb{F}$ in some nearest-addition in any order. If the height h of the corresponding binary summation tree satisfies*

$$h \le \begin{cases} \mathbf{u}^{-\frac{1}{2}} - 1 & \text{if } \beta = 2 \\ \sqrt{4 - 8\beta^{-1}}\mathbf{u}^{-\frac{1}{2}} - 1 & \text{otherwise,} \end{cases} \tag{9}$$

then

$$\left| s - \sum_{j=1}^{n} a_j \right| \le h\mathbf{u} \sum_{j=1}^{n} |a_j|. \tag{10}$$

Furthermore, if one substitutes $2\mathbf{u}$ for the error constant $\mathbf{u}$ in (9) as well as (10), the result remains valid for any faithful-addition.

Corollary 4. *Denote by s the result of a floating-point dot product of $a, b \in \mathbb{F}^n$ in some rounding to nearest. Let the height h of the corresponding binary evaluation tree satisfy (9). Then, barring underflow,*

$$\left| s - a^T b \right| \le h\mathbf{u} \sum_{i=1}^{n} |a_i b_i|. \tag{11}$$

For faithful rounding the result is true when replacing the error constant $\mathbf{u}$ by $2\mathbf{u}$.

3 Estimate for bounded number of summands

The main result of this section is presented by the following theorem.

Theorem 5. *Let a binary tree T with root r be given. Denote the set of inner nodes of the subtree with root j including j by N_j, and the set of its leaves by L_j. Let to each leaf i a real number x_i, and to each inner node j a real number δ_j be assigned. Moreover, let positive real numbers λ, μ satisfying*

$$\min\left\{ \lambda + \sum_{j \in N_r} |\delta_j|, \, \mu \right\} \le \sum_{i \in L_r} |x_i| \le \mu + \sum_{j \in N_r} |\delta_j| \tag{12}$$

be given. For all inner nodes k define

$$\sigma_k := \sum_{i \in L_k} |x_i| + \sum_{j \in N_k} |\delta_j| \quad \text{and} \quad \xi_k := \begin{cases} \frac{|\delta_k|}{\lambda} & \text{if } \sigma_{\mathrm{left}(k)} + \sigma_{\mathrm{right}(k)} - |\delta_k| < \mu \\ \frac{|\delta_k|}{\mu} & \text{otherwise,} \end{cases}$$

where $\mathrm{left}(k)$ and $\mathrm{right}(k)$ denote the left and right child of k, respectively. Suppose

$$\sigma_{\mathrm{left}(k)} + \sigma_{\mathrm{right}(k)} - |\delta_k| \ge \mu \quad \Longrightarrow \quad |\delta_k| \le \min\{\sigma_{\mathrm{left}(k)}, \sigma_{\mathrm{right}(k)}, \mu\} \tag{13}$$

as well as

$$\sum_{j\in N_r} \xi_j \le \frac{\mu-\lambda}{2\lambda}. \tag{14}$$

Then

$$\sum_{j\in N_r} |\delta_j| \le \frac{\sum_{j\in N_r}\xi_j}{1+\sum_{j\in N_r}\xi_j} \sum_{i\in L_r} |x_i|. \tag{15}$$

Furthermore,

$$\forall k \in N_r\colon \quad \sigma_k \le \sum_{i\in L_r} |x_i| + \sum_{j\in N_r} |\delta_j| \le \frac{\mu^2}{\lambda}. \tag{16}$$

Theorem 5 applies to summations in some m-digit floating-point number system with base β. The tree T describes the order in which the respective sum is evaluated. By choosing the numbers λ and μ to be consecutive powers of β that satisfy the condition (12), it is possible to use the relative error constant $\mathbf{u}$ as a bound for the relative errors ξ_j. Proposition 1 then follows as a subcase of Theorem 5. The corresponding proof is given at the end of this section.

Nevertheless, here we want to stress the fact that the above result is more general than Proposition 1. The inequality (15) not only regards a bound for the local relative errors with respect to the intermediate sums but gives a tighter estimate in correspondence to the relative errors defined with respect to the maximally possible sum of absolute values of the x_i and absolute values of the perturbations δ_i in the respective subtree. In the same manner, the core condition (14) does not bound the number of summands directly but describes a bound on the sum of the actual relative errors ξ_i. Moreover, the nearest-addition property $|(a \mathbin{\tilde{+}} b) - (a+b)| \le \min\{|a|,|b|\}$ is required solely for the intermediate values whose absolute values are greater than or equal to μ, see (13).

It is straightforward to prove that the estimate (15) is sharp, i.e., to any given height h it is possible to construct a tree T such that (15) is satisfied with equality. Additionally, the restriction on the sum of relative errors is mandatory, and it is sharp in the sense that the upper bound cannot be replaced by any larger value. We skip the arguments for these statements here, since the sharpness has already been shown for the subcase treated in Proposition 1.

In order to avoid a mess up of two considerably different lines of arguments, we will first show the following auxiliary result.

Lemma 6. *Let a positive real number μ as well as a binary tree T with root r be given. Furthermore, let to each leaf i a real number x_i, and to each inner node j a real number δ_j be assigned. Denote the set of inner nodes of the subtree with root j including j by N_j, and the set of its leaves by L_j. For all inner nodes k define*

$$\sigma_k := \sum_{i\in L_k} |x_i| + \sum_{j\in N_k} |\delta_j|$$

as well as

$$\Omega_k := \{j \in N_k \colon \sigma_{\mathrm{left}(j)} + \sigma_{\mathrm{right}(j)} - |\delta_j| \ge \mu\} \quad \text{and} \quad \mho_k := N_k \setminus \Omega_k, \tag{17}$$

where $\operatorname{left}(j)$ *and* $\operatorname{right}(j)$ *denote the left and right child of* j*, respectively. Suppose*

$$\forall j \in \Omega_r\colon \quad |\delta_j| \leq \min\{\sigma_{\operatorname{left}(j)}, \sigma_{\operatorname{right}(j)}, \mu\}. \tag{18}$$

Then

$$\Omega_r \neq \emptyset \quad \implies \quad \sum_{i \in L_r} |x_i| \geq \mu + \sum_{j \in \Omega_r} |\delta_j| - \sum_{j \in \mho_r} |\delta_j|. \tag{19}$$

Proof. The following proof is by induction on the height h of the tree, whereby for $h = 1$ the validity of the implication (19) is evident. Let a tree T with height h and root r be given, and suppose (19) is true for trees with height up to $h-1$. Denote the children of the root r by p and q. After possible renaming, we henceforth assume without loss of generality that $\Omega_q \neq \emptyset$ implies $\Omega_p \neq \emptyset$.

We distinguish three cases. First, suppose that $\Omega_q \neq \emptyset$. The induction hypothesis implies

$$\begin{aligned}
\sum_{i \in L_r} |x_i| &= \sum_{i \in L_p} |x_i| + \sum_{i \in L_q} |x_i| \\
&\geq \mu + \sum_{j \in \Omega_p} |\delta_j| - \sum_{j \in \mho_p} |\delta_j| + \mu + \sum_{j \in \Omega_q} |\delta_j| - \sum_{j \in \mho_q} |\delta_j| \\
&= \mu + \sum_{j \in \Omega_r \setminus \{r\}} |\delta_j| + \mu - \sum_{j \in \mho_r \setminus \{r\}} |\delta_j| \\
&\geq \mu + \sum_{j \in \Omega_r} |\delta_j| - \sum_{j \in \mho_r} |\delta_j|,
\end{aligned}$$

where the last inequality is evident if $r \notin \Omega_r$, and otherwise follows from $|\delta_r| \leq \mu$ implied by (18). Secondly, assume that $\Omega_q = \emptyset \neq \Omega_p$. By the induction hypothesis, $N_q = \mho_q$, and $r \in \Omega_r \implies |\delta_r| \leq \sigma_q$, we derive

$$\begin{aligned}
\sum_{i \in L_r} |x_i| &= \sum_{i \in L_p} |x_i| + \sum_{i \in L_q} |x_i| + \sum_{j \in N_q} |\delta_j| - \sum_{j \in \mho_q} |\delta_j| \\
&\geq \mu + \sum_{j \in \Omega_p} |\delta_j| - \sum_{j \in \mho_p} |\delta_j| + \sum_{i \in L_q} |x_i| + \sum_{j \in N_q} |\delta_j| - \sum_{j \in \mho_q} |\delta_j| \\
&= \mu + \sum_{j \in \Omega_r \setminus \{r\}} |\delta_j| + \sigma_q - \sum_{j \in \mho_r \setminus \{r\}} |\delta_j| \\
&\geq \mu + \sum_{j \in \Omega_r} |\delta_j| - \sum_{j \in \mho_r} |\delta_j|.
\end{aligned}$$

Finally, suppose that $\Omega_q = \emptyset = \Omega_p$. Then $\Omega_r \neq \emptyset$ implies $\Omega_r = \{r\}$ and therefore

$\sigma_p + \sigma_q - |\delta_r| \geq \mu$. Thus, using $N_p = \mho_p$ and $N_q = \mho_q$, we obtain

$$\begin{aligned}\sum_{i\in L_r} |x_i| &= \sum_{i\in L_p} |x_i| + \sum_{j\in N_p} |\delta_j| - \sum_{j\in \mho_p} |\delta_j| + \sum_{i\in L_q} |x_i| + \sum_{j\in N_q} |\delta_j| - \sum_{j\in \mho_q} |\delta_j| \\ &= \sigma_p + \sigma_q - |\delta_r| + |\delta_r| - \sum_{j\in \mho_r} |\delta_j| \\ &\geq \mu + \sum_{j\in \Omega_r} |\delta_j| - \sum_{j\in \mho_r} |\delta_j|,\end{aligned}$$

which completes the proof. □

Using Lemma 6 the proof of the main result is rather straightforward and can be done without the use of an induction argument.

Proof of Theorem 5. Let Ω_r and $\mho_r$ be defined as in Lemma 6. Since the implications (13) and (18) are equivalent, the assumption of Lemma 6 is satisfied, and the result therefore applicable. We begin by proving inequality (15), for which we distinguish two cases. First, assume that $\Omega_r = \emptyset$ and $\lambda + \sum_{j\in N_r} |\delta_j| \leq \sum_{i\in L_r} |x_i|$. Then $\xi_j = \frac{|\delta_j|}{\lambda}$, and

$$\sum_{j\in N_r} |\delta_j| = \frac{\sum_{j\in N_r} \frac{|\delta_j|}{\lambda}}{1 + \sum_{j\in N_r} \frac{|\delta_j|}{\lambda}} \left(\lambda + \sum_{j\in N_r} |\delta_j| \right) \leq \frac{\sum_{j\in N_r} \xi_j}{1 + \sum_{j\in N_r} \xi_j} \sum_{i\in L_r} |x_i|$$

shows the validity of (15) for the first case.

Secondly, suppose the opposite. Then (12) implies that either $\Omega_r \neq \emptyset$ or that $\mu \leq \sum_{i\in L_r} |x_i|$. For the former case Lemma 6 gives

$$\sum_{i\in L_r} |x_i| \geq \mu + \sum_{j\in \Omega_r} |\delta_j| - \sum_{j\in \mho_r} |\delta_j|.$$

For the latter we may assume $\Omega_r = \emptyset$ and derive the same inequality via

$$\sum_{i\in L_r} |x_i| \geq \mu \geq \mu - \sum_{j\in \mho_r} |\delta_j| = \mu + \sum_{j\in \Omega_r} |\delta_j| - \sum_{j\in \mho_r} |\delta_j|. \tag{20}$$

The assumption (14) implies

$$\frac{\mu}{\lambda} - \sum_{j\in N_r} \xi_j = \frac{\mu - \lambda}{\lambda} + 1 - \sum_{j\in N_r} \xi_j \geq 2 \sum_{j\in N_r} \xi_j + 1 - \sum_{j\in N_r} \xi_j = 1 + \sum_{j\in N_r} \xi_j,$$

such that $\frac{\frac{\mu}{\lambda} - \sum_{j\in N_r} \xi_j}{1 + \sum_{j\in N_r} \xi_j} \geq 1$. Together with $N_r = \Omega_r \cup \mho_r$ and the inequality in

(20), we then derive

$$\begin{aligned}
\sum_{j\in N_r} |\delta_j| &\leq \sum_{j\in\Omega_r} |\delta_j| + \frac{\frac{\mu}{\lambda} - \sum_{j\in N_r} \xi_j}{1+\sum_{j\in N_r} \xi_j} \sum_{j\in\mho_r} |\delta_j| \\
&= \frac{\sum_{j\in\Omega_r} |\delta_j| + \sum_{j\in N_r} \xi_j \sum_{j\in\Omega_r} |\delta_j| + \mu \sum_{j\in\mho_r} \xi_i - \sum_{j\in N_r} \xi_j \sum_{j\in\mho_r} |\delta_j|}{1+\sum_{j\in N_r} \xi_j} \\
&= \frac{\mu \sum_{j\in N_r} \xi_j + \sum_{j\in N_r} \xi_j \sum_{j\in\Omega_r} |\delta_j| - \sum_{j\in N_r} \xi_j \sum_{j\in\mho_r} |\delta_j|}{1+\sum_{j\in N_r} \xi_j} \\
&= \frac{\sum_{j\in N_r} \xi_j}{1+\sum_{j\in N_r} \xi_j} \left(\mu + \sum_{j\in\Omega_r} |\delta_j| - \sum_{j\in\mho_r} |\delta_i|\right) \\
&\leq \frac{\sum_{j\in N_r} \xi_j}{1+\sum_{j\in N_r} \xi_j} \sum_{i=1}^{n} |x_i|
\end{aligned}$$

and validate (15).

For the validation of (16) we use (12) to show that

$$\sum_{i\in L_r} |x_i| + \sum_{j\in N_r} |\delta_j| \leq \mu + 2 \sum_{j\in N_r} |\delta_j| = \mu + \frac{\mu+\lambda}{\lambda} \sum_{j\in N_r} |\delta_j| - \frac{\mu-\lambda}{\lambda} \sum_{j\in N_r} |\delta_j|.$$

Together with (15), (14), and the implication $0 \leq a \leq b \implies \frac{a}{1+a} \leq \frac{b}{1+b}$, we derive

$$\begin{aligned}
\sum_{i\in L_r} |x_i| + \sum_{j\in N_r} |\delta_j| &\leq \mu + \frac{\mu+\lambda}{\lambda} \frac{\sum_{j\in N_r} \xi_j}{1+\sum_{j\in N_r} \xi_j} \sum_{i\in L_r} |x_i| - \frac{\mu-\lambda}{\lambda} \sum_{j\in N_r} |\delta_j| \\
&\leq \mu + \frac{\mu+\lambda}{\lambda} \frac{\frac{\mu-\lambda}{2\lambda}}{1+\frac{\mu-\lambda}{2\lambda}} \left(\mu + \sum_{j\in N_r} |\delta_j|\right) - \frac{\mu-\lambda}{\lambda} \sum_{j\in N_r} |\delta_j| \\
&= \mu + \frac{\mu-\lambda}{\lambda} \left(\mu + \sum_{j\in N_r} |\delta_j|\right) - \frac{\mu-\lambda}{\lambda} \sum_{j\in N_r} |\delta_j| \\
&= \mu + \frac{\mu-\lambda}{\lambda} \mu = \frac{\mu^2}{\lambda},
\end{aligned}$$

which completes the proof. □

We close this section by giving the argument for Proposition 1.

Proof of Proposition 1. In the following we exploit the same notation as in Lemma 6 and Theorem 5, where we assume that the δ_k correspond to the rounding errors introduced by the floating-point operations. In addition we denote the intermediate sums by

$$s_k = \sum_{i\in L_k} x_i + \sum_{j\in N_k} \delta_j = s_{\text{left}(k)} + s_{\text{right}(k)} + \delta_k = s_{\text{left}(k)} \tilde{+} s_{\text{right}(k)}.$$

Since floating-point additions in the underflow range do not cause rounding errors, we may henceforth assume without loss of generality that $\sum_{i\in L_r} |x_i|$ lies in the range of normalized numbers. Moreover, without loss of generality, we may also reduce the following argument taking only these inner nodes k into account for which $s_{\text{left}(k)} + s_{\text{right}(k)}$ lies in the normalized range. Denote by $\text{ufp}(s_k)$ the largest power of β being less than or equal to $|s_k|$. Define $\tau := \text{ufp}(\sum_{i\in L_r} |x_i|)$,

$$\lambda := \begin{cases} \beta^{-1}\tau & \text{if } \sum_{i\in L_r} |x_i| < \tau + \sum_{j\in N_r} |\delta_j| \\ \tau & \text{otherwise,} \end{cases}$$

and $\mu := \beta\lambda$. Evidently, this definition complies with the assumption (12).

For any inner node $k \in \mho_r$ we have

$$|s_{\text{left}(k)} + s_{\text{right}(k)}| - |\delta_k| \leq \sigma_{\text{left}(k)} + \sigma_{\text{right}(k)} - |\delta_k| < \mu.$$

By nearest-addition and $\mu \in \mathbb{F}$ this implies $|s_{\text{left}(k)} + s_{\text{right}(k)}| < \mu$. Thus, we have $\text{ufp}(s_{\text{left}(k)} + s_{\text{right}(k)}) \leq \beta^{-1}\mu = \lambda$ and

$$\xi_k = \frac{|\delta_k|}{\lambda} \leq \frac{|\delta_k|}{\text{ufp}(s_{\text{left}(k)} + s_{\text{right}(k)})} \leq \mathbf{u} \qquad \text{for} \quad k \in \mho_r.$$

Furthermore, (16) gives $|s_{\text{left}(k)}| + |s_{\text{right}(k)}| \leq \sigma_k \leq \frac{\mu^2}{\lambda} = \beta\mu$ for all inner nodes k, such that

$$\xi_k = \frac{|\delta_k|}{\mu} \leq \frac{|\delta_k|}{\text{ufp}(s_{\text{left}(k)} + s_{\text{right}(k)})} \leq \mathbf{u} \qquad \text{for} \quad k \in \Omega_r.$$

Hence, all ξ_k are bounded by $\mathbf{u}$. The assumption (13), in turn, follows by the nearest-addition property (4)

$$|\delta_k| \leq \min\{|s_{\text{left}(k)}|, |s_{\text{right}(k)}|\}$$

and the fact that, for any m-digit floating point number system with $m \geq 1$, the difference between two successive numbers is never greater than the ufp of any of these numbers, i.e.

$$|\delta_k| \leq \frac{\text{ufp}(s_{\text{left}(k)} + s_{\text{right}(k)})}{2} \leq \frac{\mu}{2} < \mu.$$

Finally, the limit on n and $\xi_k \leq \mathbf{u}$ imply

$$\sum_{j\in N_r} \xi_j \leq (n-1)\mathbf{u} \leq \frac{\beta - 1}{2} = \frac{\mu - \lambda}{2\lambda}.$$

Since all assumptions in Theorem 5 are satisfied, (15) implies (8). □

4 Estimate depending on the height of the evaluation tree

The objective of this section is the proof of a new error estimate for binary floating-point summation depending on the height of the summation tree rather than on the number of summands. As before, we will first show a much more general result for perturbations of sums of real numbers from which the validity of the corresponding floating-point result stated in Section 2 follows. Having said that, treating arbitrary α-ary trees in Theorem 7 instead of just binary trees is rather due to technical reasons than in favor of the generality. In fact, the consideration of α-ary trees actually simplifies the notation.

Theorem 7. *Let an α-ary tree T with root r and height h be given. For an inner node j of T, denote the set of leaves of the corresponding subtree by L_j and the set of all its inner nodes including j by N_j. To each leaf i of T associate a real number x_i. Moreover, let positive real numbers b, ε as well as $\beta \geq \alpha$ be given, and let two numbers*

$$\delta_j \in \mathbb{R} \qquad \textit{and} \qquad b_j \in \{0\} \cup \{\beta^m b \mid m \in \mathbb{Z}\} \tag{21}$$

be assigned to each inner node j of T. Suppose that for each inner node j

$$|\delta_j| \leq b_j \leq \varepsilon \left(\sum_{i \in L_j} |x_i| + \sum_{i \in N_j \setminus \{j\}} |\delta_i| \right). \tag{22}$$

If h is restricted by

$$h \leq 2\sqrt{c_h \varepsilon^{-1}} - 1 \qquad \textit{with} \quad c_h := \begin{cases} \beta^{-1} - \beta^{-2} & \textit{if } \alpha = \beta \\ 1 - \alpha\beta^{-1} & \textit{otherwise,} \end{cases} \tag{23}$$

then

$$\sum_{i \in N_r} |\delta_i| \leq h\varepsilon \sum_{i \in L_r} |x_i|. \tag{24}$$

Proof. The proof is by induction on the height h. However, we replace the right inequality of (24) by

$$\sum_{i \in N_r} |\delta_i| \leq h\varepsilon \sum_{i \in L_r} |x_i| - (\eta - h) \max\left\{ b_r - \varepsilon \sum_{i \in L_r} |x_i|, 0 \right\}, \tag{25}$$

where $\eta := 2\sqrt{c_h \varepsilon^{-1}} - 1$. This induction hypothesis is stronger than (24) because $\eta \geq h$ by (23). For ease of notation we specify $\delta_k := 0$ for all leaves $k \in L_r$.

For $h = 1$, condition (22) implies $|\delta_r| \leq b_r \leq \varepsilon \sum_{i \in L_r} |x_i|$ which is (25). Suppose that (25) is true for trees with height up to $\hbar := h - 1 \geq 1$, and denote by C_j the set of children of an inner node j. The induction hypothesis, (22), and $\eta - \hbar > 0$ imply

$$\begin{aligned}
\sum_{i \in N_r} |\delta_i| &= |\delta_r| + \sum_{j \in C_r} \sum_{i \in N_j} |\delta_i| \\
&\le b_r + \sum_{j \in C_r} \left(\hbar\varepsilon \sum_{i \in L_j} |x_i| - (\eta - \hbar) \max\left\{ b_j - \varepsilon \sum_{i \in L_j} |x_i|, 0 \right\} \right) \\
&= b_r + \hbar\varepsilon \sum_{i \in L_r} |x_i| - (\eta - \hbar) \max\left\{ \sum_{j \in C_r} \max\left\{ b_j - \varepsilon \sum_{i \in L_j} |x_i|, 0 \right\}, 0 \right\} \\
&\le b_r + \hbar\varepsilon \sum_{i \in L_r} |x_i| - (\eta - \hbar) \max\left\{ \sum_{j \in C_r} \left(b_j - \varepsilon \sum_{i \in L_j} |x_i| \right), 0 \right\} \\
&= b_r + \hbar\varepsilon \sum_{i \in L_r} |x_i| - (\eta - \hbar) \max\left\{ \sum_{j \in C_r} b_j - \varepsilon \sum_{i \in L_r} |x_i|, 0 \right\}.
\end{aligned}$$

We proceed by case distinction. First, suppose $b_r \le \varepsilon \sum_{i \in L_r} |x_i|$. Then the max expression in (25) is zero, and

$$\sum_{i \in N_r} |\delta_i| \le b_r + \hbar\varepsilon \sum_{i \in L_r} |x_i| - (\eta - \hbar) \max\{ \sum_{j \in C_r} b_j - \varepsilon \sum_{i \in L_r} |x_i|, 0 \} \le h\varepsilon \sum_{i \in L_r} |x_i|$$

implies (25). Secondly, suppose $b_r > \varepsilon \sum_{i \in L_r} |x_i|$ and $b_r \le \sum_{j \in C_r} b_j$. Then

$$\begin{aligned}
\sum_{i \in N_r} |\delta_i| &\le b_r + \hbar\varepsilon \sum_{i \in L_r} |x_i| - (\eta - \hbar) \max\left\{ b_r - \varepsilon \sum_{i \in L_r} |x_i|, 0 \right\} \\
&= b_r + \hbar\varepsilon \sum_{i \in L_r} |x_i| - (\eta - \hbar) \left(b_r - \varepsilon \sum_{i \in L_r} |x_i| \right) \\
&= h\varepsilon \sum_{i \in L_r} |x_i| - (\eta - h) \left(b_r - \varepsilon \sum_{i \in L_r} |x_i| \right)
\end{aligned}$$

proves (25) for the second case. Finally, it remains to show the validity of (25) for

$$b_r > \varepsilon \sum_{i \in L_r} |x_i| \qquad \text{and} \qquad b_r > \sum_{j \in C_r} b_j. \tag{26}$$

Due to (21), all summands b_j on the right-hand side of the right inequality have to be less than or equal to $\beta^{-1} b_r$. Moreover, since equality is not allowed, $\alpha = \beta$ and $|C_r| \le \alpha$ require at least one summand to be less or equal to $\beta^{-2} b_r$. Therefore,

$$\sum_{k \in C_r} |\delta_k| \le \sum_{k \in C_r} b_k \le \begin{cases} (\alpha - 1)\beta^{-1} b_r + \beta^{-2} b_r & \text{if } \alpha = \beta \\ \alpha\beta^{-1} b_r & \text{otherwise.} \end{cases}$$

By definition of c_h both right most terms are equal to $(1-c_h)b_r$, so that

$$\sum_{k\in C_r} |\delta_k| \le (1-c_h)b_r. \tag{27}$$

Furthermore, (22) and the induction hypothesis give

$$b_r - \varepsilon \sum_{i\in L_r} |x_i| \le \varepsilon \sum_{i\in N_r\setminus\{r\}} |\delta_i| = \varepsilon \sum_{j\in C_r}\sum_{i\in N_j} |\delta_i| \le (h-1)\varepsilon^2 \sum_{i\in L_r} |x_i|. \tag{28}$$

Then the induction hypothesis, (27), (26), and (28) yield

$$\begin{aligned}
\sum_{i\in N_r} |\delta_i| &\le b_r + \sum_{k\in C_r}\left(|\delta_k| + \sum_{j\in C_k}\sum_{i\in N_j} |\delta_i|\right)\\
&\le b_r + \sum_{k\in C_r}\left(|\delta_k| + \sum_{j\in C_k}\left((h-2)\varepsilon \sum_{i\in L_j} |x_i|\right)\right)\\
&= b_r + \sum_{k\in C_r} |\delta_k| + (h-2)\varepsilon \sum_{i\in L_r} |x_i|\\
&\le (2-c_h)b_r + (h-2)\varepsilon \sum_{i\in L_r} |x_i|\\
&= h\varepsilon \sum_{i\in L_r} |x_i| - (\eta-h)\left(b_r - \varepsilon \sum_{i\in L_r} |x_i|\right)\\
&\qquad + (\eta-h+2)\left(b_r - \varepsilon \sum_{i\in L_r} |x_i|\right) - c_h b_r\\
&< h\varepsilon \sum_{i\in L_r} |x_i| - (\eta-h)\max\left\{b_r - \varepsilon \sum_{i\in L_r} |x_i|, 0\right\}\\
&\qquad + (\eta-h+2)(h-1)\varepsilon^2 \sum_{i\in L_r} |x_i| - c_h\varepsilon \sum_{i\in L_r} |x_i|.
\end{aligned}$$

The quadratic expression $(\eta-h+2)(h-1)$ has its maximum at $h=\frac{3}{2}+\frac{\eta}{2}$. This implies

$$(\eta-h+2)(h-1)\varepsilon^2 \le \left(\frac{1}{2}+\frac{\eta}{2}\right)^2 \varepsilon^2 = \left(\frac{1}{2}+\frac{2\sqrt{c_h\varepsilon^{-1}}-1}{2}\right)^2 \varepsilon^2 = c_h\varepsilon$$

and finishes the proof. □

Proposition 3 now follows as a corollary of the result given above. As in the previous section, for the discussion of the results regarding some m-digit floating-point number system in base β, we again make use of the *unit in the first place* notation, i.e., $\mathrm{ufp}(s_k)$ denotes the largest power of β less than or equal to $|s_k|$, with the convention $\mathrm{ufp}(0) := 0$.

Proof of Proposition 3. Let T denote the considered summation tree, where to each inner node j of T we associate the respective intermediate summation result s_j including the perturbations δ_i. Using the notation as in Theorem 7 it follows $s_j = \sum_{i\in L_j} x_i + \sum_{i\in N_j} \delta_i$. Furthermore, let $b = \varepsilon = \eta$, where $\eta = \mathbf{u}$ in case of nearest-addition and $\eta = 2\mathbf{u}$ in case of faithful-addition. Define $b_j := \eta\,\mathrm{ufp}(s_j)$ for all inner nodes j. This definition of b_j complies with the assumption (21), i.e., $b_j \in \{0\} \cup \{\beta^m \eta \mid m \in \mathbb{Z}\}$. Moreover,

$$|\delta_j| \leq b_j = \eta\,\mathrm{ufp}(s_j) \leq \eta |s_j - \delta_j| \leq \eta \left(\sum_{i\in L_j} |x_i| + \sum_{i\in N_j\setminus\{j\}} |\delta_i| \right)$$

validates the assumption (22). Finally, for $\alpha = 2$,

$$h \leq \begin{cases} \eta^{-\frac{1}{2}} - 1 = 2\sqrt{(\beta^{-1} - \beta^{-2})\eta^{-1}} - 1 & \text{if } \ \beta = \alpha \\ \sqrt{4 - 8\beta^{-1}}\eta^{-\frac{1}{2}} - 1 = 2\sqrt{(1 - \alpha\beta^{-1})\eta^{-1}} - 1 & \text{otherwise} \end{cases}$$

shows the equivalence of (23) and (9). Thus (10) follows. □

Proof of Corollary 4. Let T denote the tree to the evaluation of the inner product $a^T b$, i.e., a tree with leaves $a_1, b_1, a_2, b_2, \ldots, a_n, b_n$, where all inner nodes are added via $\tilde{+}$ and the leaf pairs a_i, b_i are multiplied via the corresponding floating-point multiplication. Since underflow is barred, the tree T can be transformed into a pure summation tree by replacing the leaf pairs a_i, b_i with x_i, y_i satisfying $a_i b_i = x_i + y_i$. The remainder of the argument follows from the proof of Proposition 3. □

Remark 8. In case of nearest-addition it can be shown that inequality (24) in Theorem 7 still holds valid if we substitute $\frac{u}{1+u}$ for ε and replace condition (22) with

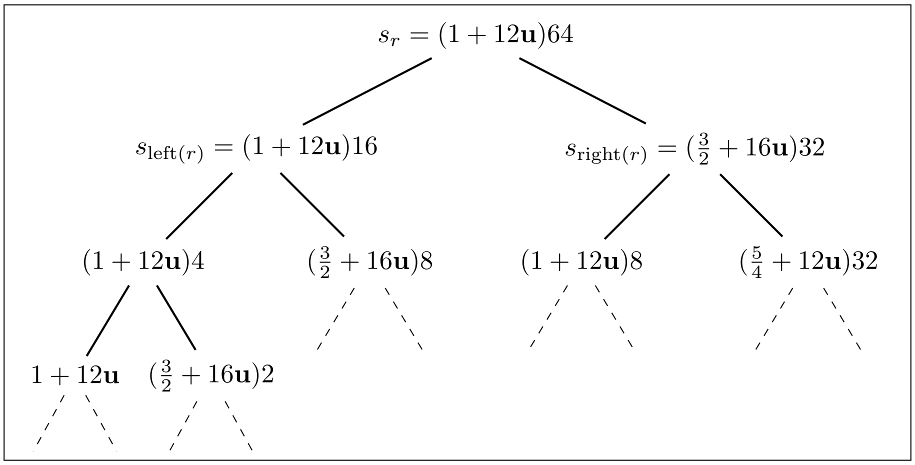
$$|\delta_j| \leq b_j \leq \begin{cases} \dfrac{u}{1+2u}|s_j| & \text{if } \ |s_j - \delta_j| > |s_j| \\ u\,|s_j| & \text{otherwise,} \end{cases} \tag{29}$$

where $s_j := \sum_{i\in L_j} x_i + \sum_{i\in N_j} \delta_i$. For nearest rounding this modification allows us to replace the factor $h\mathbf{u}$ in the estimate (10) as well as (11) with the smaller factor $\frac{h\mathbf{u}}{1+\mathbf{u}}$.

Since the replacement of the factor $h\mathbf{u}$ with $\frac{h\mathbf{u}}{1+\mathbf{u}}$ seems of little practical relevance and requires a considerably more complicated argument, we skip the proof[2] here.

Another remarkable fact is the sharpness of the estimate given in Theorem 7, which has already been shown for the subcase of recursive summation in [4]. For the discussion of the bound on the height h, we will exploit the following auxiliary result.

[2] a proof can be found at `www.ti3.tuhh.de/rump/paper/BinTreeAppendix.pdf`



$s_r = (1 + 12\mathbf{u})64$
$s_{\mathrm{left}(r)} = (1 + 12\mathbf{u})16$
$s_{\mathrm{right}(r)} = (\frac{3}{2} + 16\mathbf{u})32$
$(1 + 12\mathbf{u})4$
$(\frac{3}{2} + 16\mathbf{u})8$
$(1 + 12\mathbf{u})8$
$(\frac{5}{4} + 12\mathbf{u})32$
$1 + 12\mathbf{u}$
$(\frac{3}{2} + 16\mathbf{u})2$

Figure 1: Binary summation tree with exponential increasing error

Lemma 9. *Consider a binary summation in* IEEE 754 *double floating-point format with* $\mathbf{u} = 2^{-53}$ *in rounding to nearest mode. Let* $c \in \mathbb{F}$ *be an even positive floating-point number, where the property* even *denotes that the last bit of the mantissa is zero. Moreover, let* $\frac{c}{8}$ *lie in the normalized range of* $\mathbb{F}$. *Then, there exist even positive* $a, b \in \mathbb{F}$ *that satisfy*

$$c = a \tilde{+} b, \quad c - (a + b) = \mathbf{u}\,\underline{\mathrm{ufp}}(c), \quad \mathbf{u}\,\underline{\mathrm{ufp}}(a) + \mathbf{u}\,\underline{\mathrm{ufp}}(b) \geq (\tfrac{3}{5}\mathbf{u} - 4\mathbf{u}^2)\, c,$$

where $\underline{\mathrm{ufp}}(c)$ *denotes the largest power of* 2 *being strictly smaller than* $|c|$.

Proof. For the specific choice $a = (\frac{1}{4} + 3\mathbf{u})\,\mathrm{ufp}(c)$ and $b = c - a - \mathbf{u}\,\underline{\mathrm{ufp}}(c)$, it is straightforward to verify the statements. □

As a consequence of this lemma, it is possible to generate a binary summation tree for IEEE 754 double floating-point format with an error growing exponentially in h. In each step the growing factor is greater than or equal to $1 + \frac{3}{5}\mathbf{u} - 4\mathbf{u}^2$, such that

$$s_r - \sum_{i \in L_r} x_i \geq \Big(\big(1 + \tfrac{3}{5}\mathbf{u} - 4\mathbf{u}^2\big)^h - 1\Big) \sum_{i \in L_r} |x_i|.$$

This shows that a restriction on the height is mandatory although the upper bound on h may not be sharp. Figure 1 demonstrates the reverse generation of such a summation tree with root $s_r = (1 + 12\mathbf{u})64$.

5 An estimate for recursive summation uniformly bounded by 1

Before discussing some applications for the refined error bounds given in the previous sections, we will show another perturbation estimate for the specific case of recursive summation. Like the other bounds introduced in the sections above, the following error estimate also improves upon previous worst case analyses of summation errors. Unlike our other theorems, however, here no restriction on n is necessary. Again, the given estimate not only regards the local relative errors with respect to the corresponding intermediate sum but gives a tighter bound in correspondence to the relative errors that are defined with respect to the maximally possible sum of absolute values of the summands and absolute values of the perturbations in the i-th step.

Since rounding to nearest implies by (7) that no error is greater than any of the corresponding addends, apparently, the sum of absolute values of the errors $\sum_{i=1}^{n} |\delta_i|$ is bounded by the sum of absolute values of the addends $\sum_{i=1}^{n} |x_i|$. This inequality is, however, not reflected by the factor $n\mathbf{u}$ proved in earlier papers. The following theorem gives a sharp bound for the error of recursive summation under the only assumption that the error introduced in the i-th step is not greater than the i-th addend. As expected, the sharp estimate in (30) is itself bounded by $\sum_{i=1}^{n} |x_i|$. Indeed the introduced factor slowly converges to 1 and is never greater than $\frac{n\mathbf{u}}{1+\mathbf{u}}$ (respectively the sum of the corresponding relative errors). Nevertheless, we mention that Theorem 10 is rather of theoretical interest than of practical relevance, in particular since the evaluation of $\prod_{i=1}^{n} \frac{1-\mathbf{u}}{1+\mathbf{u}}$ would be difficult in practice.

Theorem 10. *Let $x, \varepsilon \in \mathbb{R}^n$ be given. Define vectors $\delta, s \in \mathbb{R}^n$ such that $s_1 = x_1 + \delta_1 = x_1(1+\varepsilon_1)$ and*

$$s_k = x_k + s_{k-1} + \delta_k = (x_k + s_{k-1})(1+\varepsilon_k)$$

for $2 \le k \le n$. For every index $k = 1, \ldots, n$ suppose that $|\delta_k| \le |x_k|$ and define

$$\xi_k := \frac{|\delta_k|}{\sum_{i=1}^{k} |x_i| + \sum_{i=1}^{k-1} |\delta_i|} \qquad \text{and} \qquad q_k := \prod_{i=1}^{k} \frac{1-\xi_i}{1+\xi_i}$$

with the convention $\frac{0}{0} := 0$. Then

$$\left| s_n - \sum_{i=1}^{n} x_i \right| \le \sum_{i=1}^{n} |\delta_i| \le \frac{1-q_n}{1+q_n} \sum_{i=1}^{n} |x_i|. \tag{30}$$

Both inequalities in (30) are sharp in the sense that for arbitrary positive x_1 and $\varepsilon \in [0, 1)^n$, there exist $x_2, \ldots, x_n$ such that the inequalities become equalities. Moreover,

$$\frac{1-q_n}{1+q_n} \le \sum_{i=1}^{n} \xi_i \le \sum_{i=1}^{n} |\varepsilon_i|. \tag{31}$$

Remark 11. It will be clear from (36) in the proof that the first inequality in (31) is strict if $\delta_{i-1}\delta_i \neq 0$ for some $i \in \{2, \ldots, n\}$. That shows that the factor $\frac{1-q_n}{1+q_n}$ in (30) is better than the sum of relative errors ξ_i with respect to the maximally possible sum of absolute values of the summands and absolute values of the perturbations, and a fortiori better than the sum of relative errors $|\varepsilon_i|$ of the individual sums.

Proof. We first prove (30). The only assumption in Theorem 10, namely $|\delta_k| \leq |x_k|$, implies $\xi_k \leq 1$ for all $k \in \{1, \ldots, n\}$. If $\xi_j = 1$ for some index j, then $q_j = 0 = q_n$ and therefore $\frac{1-q_n}{1+q_n} = 1$. In this case the validity of (30) is an immediate consequence of $|\delta_k| \leq |x_k|$. Thus, we may henceforth assume without loss of generality that $q_n > 0$, which means that $q_1, q_2, \ldots, q_n$ is a monotonically decreasing sequence of positive numbers.

The proof of (30) is by induction on n, whereas the validity for $n = 1$ is evident. Henceforth, assume that equation (30) holds valid up to $n - 1$. By definition we have $s_n - \sum_{i=1}^n x_i = \sum_{i=1}^n \delta_i$, so that

$$\left| s_n - \sum_{i=1}^{n} x_i \right| \leq |\delta_n| + \sum_{i=1}^{n-1} |\delta_i|.$$

First, assume

$$\frac{2q_{n-1}}{1+q_{n-1}} \sum_{i=1}^{n-1} |x_i| \geq \frac{2q_n}{1+q_n} \sum_{i=1}^{n} |x_i|. \tag{32}$$

The induction hypothesis, $|\delta_n| \leq |x_n|$, and (32) imply

$$\begin{aligned} \sum_{i=1}^{n} |\delta_i| &\leq |x_n| + \frac{1-q_{n-1}}{1+q_{n-1}} \sum_{i=1}^{n-1} |x_i| \\ &= \sum_{i=1}^{n} |x_i| - \frac{2q_{n-1}}{1+q_{n-1}} \sum_{i=1}^{n-1} |x_i| \\ &\leq \sum_{i=1}^{n} |x_i| - \frac{2q_n}{1+q_n} \sum_{i=1}^{n} |x_i| \\ &= \frac{1-q_n}{1+q_n} \sum_{i=1}^{n} |x_i|. \end{aligned} \tag{33}$$

This proves (30) if (32) is satisfied. Secondly, suppose the opposite case, namely

$$\frac{2q_{n-1}}{1+q_{n-1}} \sum_{i=1}^{n-1} |x_i| < \frac{2q_n}{1+q_n} \sum_{i=1}^{n} |x_i|. \tag{34}$$

For all $n > 1$, the definitions of ξ_n and q_n give

$$|\delta_n| = \xi_n \left(\sum_{i=1}^{n} |x_i| + \sum_{i=1}^{n-1} |\delta_i| \right) = \frac{q_{n-1} - q_n}{q_{n-1} + q_n} \left(\sum_{i=1}^{n} |x_i| + \sum_{i=1}^{n-1} |\delta_i| \right).$$

Together with (34) and the induction hypothesis, we derive

$$\begin{aligned}
\sum_{i=1}^{n} |\delta_i| &= \frac{q_{n-1} - q_n}{q_{n-1} + q_n} \left(\sum_{i=1}^{n} |x_i| + \sum_{i=1}^{n-1} |\delta_i| \right) + \sum_{i=1}^{n-1} |\delta_i| \\
&= \frac{q_{n-1} - q_n}{q_{n-1} + q_n} \sum_{i=1}^{n} |x_i| + \frac{2q_{n-1}}{q_{n-1} + q_n} \sum_{i=1}^{n-1} |\delta_i| \\
&\leq \frac{q_{n-1} - q_n}{q_{n-1} + q_n} \sum_{i=1}^{n} |x_i| + \frac{2q_{n-1}}{q_{n-1} + q_n} \frac{1 - q_{n-1}}{1 + q_{n-1}} \sum_{i=1}^{n-1} |x_i| \\
&\leq \frac{q_{n-1} - q_n}{q_{n-1} + q_n} \sum_{i=1}^{n} |x_i| + \frac{2q_{n}}{q_{n-1} + q_n} \frac{1 - q_{n-1}}{1 + q_{n}} \sum_{i=1}^{n} |x_i| \\
&= \frac{1 - q_n}{1 + q_n} \sum_{i=1}^{n} |x_i| \, .
\end{aligned}$$

This finishes the proof of (30).

To show that both estimates in (30) are sharp, let arbitrary positive x_1 and $\varepsilon \in [0,\, 1)^n$ be given. The summands $x_2, \dots, x_n$ to be defined are nonnegative. In that case, together with the non-negativity of the relative errors ε_k, we have $\delta_k \geq 0$, and therefore $\varepsilon_1 = \xi_1$ as well as

$$\varepsilon_k = \frac{\delta_k}{x_k + s_{k-1}} = \frac{\delta_k}{x_k + \sum_{i=1}^{k-1} x_i + \sum_{i=1}^{k-1} \delta_i} = \frac{|\delta_k|}{\sum_{i=1}^{k} |x_i| + \sum_{i=1}^{k-1} |\delta_i|} = \xi_k$$

for all $k = 2, \dots, n$. The equalities $\varepsilon_k = \xi_k$, in turn, imply

$$\forall 1 \leq k \leq n\colon \quad q_k = \prod_{i=1}^{k} \frac{1 - \varepsilon_i}{1 + \varepsilon_i} \qquad \text{and therefore also} \qquad \varepsilon_k = \frac{q_{k-1} - q_k}{q_{k-1} + q_k},$$

where we use the convention $q_0 := 1$. Define

$$x_k := \frac{1}{q_k} \frac{q_{k-1} - q_k}{1 + q_{k-1}} \sum_{i=1}^{k-1} x_i \qquad \text{for} \quad k = 2, \dots, n. \tag{35}$$

Since the q_k are monotonically decreasing and positive, the x_k are well-defined and nonnegative.

We proceed by induction to prove that the assumption $|\delta_k| \leq |x_k|$ is satisfied for all k, and that there are equalities in (30). For $n = 1$, we have $s_1 - x_1 = \delta_1 = \varepsilon_1 x_1 = \frac{1-q_1}{1+q_1} x_1 \leq x_1$. Suppose that, up to $k \leq n - 1$, (30) is satisfied with equalities. Then, by the non-negativity of all quantities, the induction hypothesis, and (35), we have

$$\begin{aligned}\delta_k &= \varepsilon_k \left(x_k + \sum_{i=1}^{k-1} x_i + \sum_{i=1}^{k-1} \delta_i \right) \\ &= \frac{q_{k-1} - q_k}{q_{k-1} + q_k} \left(\frac{1}{q_k} \frac{q_{k-1} - q_k}{1 + q_{k-1}} \sum_{i=1}^{k-1} x_i + \sum_{i=1}^{k-1} x_i + \frac{1 - q_{k-1}}{1 + q_{k-1}} \sum_{i=1}^{k-1} x_i \right) \\ &= \frac{q_{k-1} - q_k}{q_{k-1} + q_k} \frac{1}{q_k} \frac{q_{k-1} + q_k}{1 + q_{k-1}} \sum_{i=1}^{k-1} x_i \\ &= x_k\end{aligned}$$

for $k = 2, \ldots, n$. Furthermore, the definition of x_k given in (35) implies

$$\begin{aligned}\frac{2q_{n-1}}{1 + q_{n-1}} \sum_{i=1}^{n-1} |x_i| &= \frac{2q_n}{1 + q_n} \sum_{i=1}^{n-1} |x_i| + \frac{2q_n}{1 + q_n} \frac{1}{q_n} \frac{q_{n-1} - q_n}{1 + q_{n-1}} \sum_{i=1}^{n-1} |x_i| \\ &= \frac{2q_n}{1 + q_n} \sum_{i=1}^{n-1} |x_i| + \frac{2q_n}{1 + q_n} |x_n| \\ &= \frac{2q_n}{1 + q_n} \sum_{i=1}^{n} |x_i|.\end{aligned}$$

Hence, the assumption (32) is satisfied with equality. As a consequence, both inequalities in (33) can be replaced with equalities, so that (30) holds true with equalities as well.

Finally, for the proof of (31), we exploit $\frac{1-q_1}{1+q_1} = \xi_1$ as well as

$$q_n \le q_{n-1} \le 1 \implies (1-q_{n-1})(1-q_n) \ge 0 \implies 1+q_n q_{n-1} \ge q_{n-1}+q_n. \quad (36)$$

For $n > 1$ we then derive

$$\frac{1 - q_n}{1 + q_n} - \frac{1 - q_{n-1}}{1 + q_{n-1}} = \frac{2(q_{n-1} - q_n)}{1 + q_{n-1} + q_n + q_{n-1} q_n} \le \frac{q_{n-1} - q_n}{q_{n-1} + q_n} = \xi_n \le |\varepsilon_n|,$$

and a telescope sum using $q_0 = 1$ shows (31). □

6 Applications

In the final section we improve some well-known error bounds into new ones without higher order terms in $\mathbf{u}$. We denote by $s = \mathrm{float}(\textit{expression})$ the result of the expression with each operation replaced by the corresponding floating-point operation. The evaluation may be in any order but, if applicable, respecting parentheses. First, consider a sum of products

$$s := \sum_{i=1}^{n} \prod_{j=1}^{m} x_{ij} \qquad \text{for } x_{ij} \in \mathbb{F}. \quad (37)$$

Provided $(n+m-2)\mathbf{u} < 1$, the standard Wilkinson-type error estimate gives

$$\left| \text{float}\,\Big(\sum_{i=1}^{n} \prod_{j=1}^{m} x_{ij} \Big) - s \right| \le \gamma_{n+m-2} \sum_{i=1}^{n} \prod_{j=1}^{m} |x_{ij}|$$

using the classical $\gamma_k := \frac{k\mathbf{u}}{1-k\mathbf{u}}$. Exploiting Proposition 1 and Theorem 1.2 in [6] this can be improved as follows.

Proposition 12. *Let $x_{ij} \in \mathbb{F}$ with $1 \le i \le n$ and $1 \le j \le m$ be given for a set $\mathbb{F}$ of floating-point numbers to base β. Assume floating-point addition and multiplication in rounding to nearest. Furthermore, suppose*

$$m \le \beta^{-\frac{1}{2}}\mathbf{u}^{-\frac{1}{2}} \;, \quad n \le 1 + \frac{\beta-1}{2}\mathbf{u}^{-1}, \quad \textit{and} \quad m \le n. \tag{38}$$

If no multiplication causes underflow, then

$$\left| \text{float}\,\Big(\sum_{i=1}^{n} \prod_{j=1}^{m} x_{ij} \Big) - s \right| \le (n+m-2)\mathbf{u} \sum_{i=1}^{n} \prod_{j=1}^{m} |x_{ij}|. \tag{39}$$

For binary floating-point numbers $m \le \mathbf{u}^{-\frac{1}{2}}$ suffices for (39) *to hold true.*

Remark 13. Note that for $m = 2$ the estimate of Proposition 12 includes the error bound $|\,\text{float}(x^T y) - x^T y| \le n\mathbf{u}|x|^T|y|$ for a floating-point computation of the dot product of two vectors $x, y \in \mathbb{F}^n$. That was first proved in [2], however, without restriction on n.

Proof. For $1 \le i \le n$ denote $p_i := \text{float}(\prod_{j=1}^{m} x_{ij})$. Then (38) and [6, Theorem 1.2] imply

$$\left| p_i - \prod_{j=1}^{m} x_{ij} \right| \le (m-1)\mathbf{u} \prod_{i=1}^{m} |x_{ij}|$$

for $1 \le i \le n$. Denoting the left-hand side in (39) by Δ and again using (38), Proposition 1 gives

$$\begin{aligned} \Delta &= \left| \text{float}\,\Big(\sum_{i=1}^{n} p_i \Big) - \sum_{i=1}^{n} p_i + \sum_{i=1}^{n} \left(p_i - \prod_{j=1}^{m} x_{ij} \right) \right| \\ &\le \frac{(n-1)\mathbf{u}}{1+(n-1)\mathbf{u}} \sum_{i=1}^{n} |p_i| + (m-1)\mathbf{u} \sum_{i=1}^{n} \prod_{j=1}^{m} |x_{ij}| \\ &\le \left(\frac{(n-1)\mathbf{u}}{1+(n-1)\mathbf{u}} (1+(m-1)\mathbf{u}) + (m-1)\mathbf{u} \right) \sum_{i=1}^{n} \prod_{j=1}^{m} |x_{ij}|, \end{aligned}$$

and $m \le n$ finishes the argument. □

A direct application of Proposition 12 is a bound on the error of a Vandermonde matrix times a vector. Let

$$V_{ij} = \alpha_j^i \qquad \text{for} \quad 0 \le i, j \le n$$

for given $\alpha_j \in \mathbb{F}$. Then $(Vx)_i = \sum\limits_{j=0}^{n} \alpha_j^i x_j$, so that for a vector $x \in \mathbb{F}^{n+1}$, starting with index 0, we obtain

$$|\operatorname{float}(Vx) - Vx| \le \operatorname{diag}(n\mathbf{u}, n\mathbf{u} + \mathbf{u}, \ldots, 2n\mathbf{u})\,|V|\,|x| \le 2n\mathbf{u}\,|V|\,|x|.$$

Another application is a new error estimate for blocked summation. Let a vector $x \in \mathbb{F}^{mn}$ be given and consider blocked floating-point summation of all elements of x with fixed block size m, that is,

$$s := \operatorname{float}\left(\sum_{i=1}^{n} \left(\sum_{j=1}^{m} x_{ij} \right) \right). \tag{40}$$

Then the standard Wilkinson-type error estimate $|s - \sum_{ij} x_{ij}| \le \gamma_{n+m-2} \sum_{ij} |x_{ij}|$ can be improved as follows.

Proposition 14. *Let $x_{ij} \in \mathbb{F}$ with $1 \le i \le n$ and $1 \le j \le m$ be given for a set $\mathbb{F}$ of floating-point numbers to base β. Assume floating-point addition in rounding to nearest and suppose*

$$\max(m, n) \le 1 + \frac{\beta - 1}{2}\mathbf{u}^{-1}. \tag{41}$$

Then s as defined in (40) *satisfies*

$$\left| s - \sum_{ij} x_{ij} \right| \le (n + m - 2)\mathbf{u} \sum_{ij} |x_{ij}|. \tag{42}$$

Proof. Denoting the left-hand side in (42) by Δ and proceeding as in the proof of Proposition 12, we obtain

$$\Delta \le \left(\frac{(n-1)\mathbf{u}}{1 + (n-1)\mathbf{u}} \left(1 + \frac{(m-1)\mathbf{u}}{1 + (m-1)\mathbf{u}} \right) + \frac{(m-1)\mathbf{u}}{1 + (m-1)\mathbf{u}} \right) \sum_{i,j} |x_{ij}|.$$

The result then follows by

$$\frac{p}{1+p} \left(1 + \frac{q}{1+q} \right) + \frac{q}{1+q} = p + q + \frac{pq - p^2 - q^2 - (p+q)pq}{(1+p)(1+q)} \le p + q,$$

where p and q are substitutes for $(n-1)\mathbf{u}$ and $(m-1)\mathbf{u}$, respectively. □

Finally, we mention without proof another application of the sharper estimate in Proposition 1. For given $A \in \mathbb{F}^{m \times n}$, $y \in \mathbb{F}^m$, and $x \in \mathbb{F}^n$, the standard Wilkinson-type error bound for the floating-point computation of $y^T A x$ reads $|\operatorname{float}(y^T A x) - y^T A x| \le \gamma_{m+n} |y|^T |A|\,|x|$. If $\max(m, n) \le \mathbf{u}^{-1}$ and $n + m \ge 10$, this bound can be improved to

$$|\operatorname{float}(y^T A x) - y^T A x| \le (m + n)\,\mathbf{u}\,|y|^T |A|\,|x|.$$

References

[1] Nicholas J. Higham. *Accuracy and Stability of Numerical Algorithms*. Other Titles in Applied Mathematics. SIAM Publications, Philadelphia, 2nd edition, 2002.

[2] Claude-Pierre Jeannerod and Siegfried M Rump. Improved error bounds for inner products in floating-point arithmetic. *SIAM J. Matrix Anal. & Appl.*, 34(2):338–344, 2013.

[3] Donald E. Knuth. *The Art of Computer Programming*. Pearson Education (US), 1997.

[4] Marko Lange and Siegfried M. Rump. Error estimates for the summation of real numbers with application to floating-point summation. *BIT Numer. Math.*, 57(3):927–941, 2017.

[5] Walter F. Mascarenhas. Floating point numbers are real numbers. ArXiv:1605.09202, 2016.

[6] Siegfried M. Rump, Florian Bünger, and Claude-Pierre Jeannerod. Improved error bounds for floating-point products and Horner's scheme. *BIT Numer. Math.*, 56(1):293–307, 2016.

[7] Siegfried M. Rump and Claude-Pierre Jeannerod. Improved backward error bounds for LU and Cholesky factorizations. *SIAM J. Matrix Anal. & Appl.*, 35(2):684–698, 2014.