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

Error estimates for the summation of real numbers with application to floating-point summation

Marko Lange · Siegfried M. Rump

Received: date / Accepted: date

Abstract Standard Wilkinson-type error estimates of floating-point algorithms involve a factor $\gamma_k := k\mathbf{u}/(1-k\mathbf{u})$ for $\mathbf{u}$ denoting the relative rounding error unit of a floating-point number system. Recently, it was shown that, for many standard algorithms such as matrix multiplication, LU- or Cholesky decomposition, γ_k can be replaced by $k\mathbf{u}$, and the restriction on k can be removed. However, the arguments make heavy use of specific properties of both the underlying set of floating-point numbers and the corresponding arithmetic.

In this paper, we derive error estimates for the summation of real numbers where each sum is afflicted with some perturbation. Recent results on floating-point summation follow as a corollary, in particular error estimates for rounding to nearest and for directed rounding.

Our new estimates are sharp and decover the necessary properties of floating-point schemes to allow for a priori estimates of summation with a factor omitting higher order terms.

Keywords Floating-point, summation, Wilkinson-type error estimates, error analysis, real numbers

Mathematics Subject Classification (2000) 65G50, 65F05

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

M. Lange
Waseda University, Faculty of Science and Engineering, 3–4–1 Okubo, Shinjuku-ku, Tokyo 169–8555, Japan E-mail: m.lange@aoni.waseda.jp

S. M. Rump
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 E-mail: rump@tuhh.de

1 Introduction

A floating-point number system in accordance to the IEEE 754 standard [2,3] with basis β, mantissa length p, and exponent range $[e_{\min}, e_{\max}]$ can be defined via

$$\mathbb{F} := \{m \cdot \beta^e : m \in \beta^{-p}\mathbb{Z}, |m| < 1, e_{\min} \le e \le e_{\max}\}. \tag{1.1}$$

The error analysis regarding such a system is typically based on the two standard models for floating-point arithmetic [1, Eq. (2.4) and (2.5)]. For some operation $\tilde{+} : \mathbb{F} \times \mathbb{F} \to \mathbb{F}$ approximating a real sum according to these models, we have

$$a, b \in \mathbb{F}: \quad a \tilde{+} b = (a+b)(1+\varepsilon_1) = \frac{a+b}{1+\varepsilon_2} \quad \text{for some} \quad |\varepsilon_i| \le \mathbf{u}, \tag{1.2}$$

where $\mathbf{u}$ is a constant associated to $\mathbb{F}$. It is typically referred to as the relative rounding error unit.[1] If $a \tilde{+} b$ approximates $a+b$ by the nearest number in $\mathbb{F}$ and $a+b$ lies in the range of normalized numbers $\pm[\beta^{e_{\min}-1}, (1-\beta^{-p})\beta^{e_{\max}}]$, then (1.2) is satisfied for $\mathbf{u} := \frac{1}{2}\beta^{1-p}$.

Let $s \in \mathbb{F}$ denote the result of a summation of $x_1, \ldots, x_n \in \mathbb{F}$, where each individual sum uses $\tilde{+}$. An immediate consequence of (1.2) is

$$\Big|s - \sum_{i=1}^{n} x_i\Big| \le \big((1+\mathbf{u})^{n-1} - 1\big) \sum_{i=1}^{n} |x_i|. \tag{1.3}$$

In order to avoid the awkward factor in (1.3), usually the standard Wilkinson-type estimate is used:

$$\Big|s - \sum_{i=1}^{n} x_i\Big| \le \gamma_{n-1} \sum_{i=1}^{n} |x_i|, \tag{1.4}$$

where $\gamma_k := k\mathbf{u}/(1-k\mathbf{u})$. The estimate (1.4) is true for any order of evaluation provided that $(n-1)\mathbf{u} < 1$.

In a sequence of papers starting with [8], it was shown that for floating-point systems following the IEEE 754 standard the constant γ_{n-1} in (1.4) can be replaced by $(n-1)\mathbf{u}$ without restriction on n and for any order of evaluation [4], i.e.,

$$\Big|s - \sum_{i=1}^{n} x_i\Big| \le (n-1)\mathbf{u} \sum_{i=1}^{n} |x_i|. \tag{1.5}$$

In the sequel, this principle - the replacement of γ_k by $k\mathbf{u}$ without restriction on k - was extended to dot products, LU-decomposition, Cholesky decomposition, forward and backward substitution, and more [9]. However, all proofs make heavy use of the particular properties of an IEEE 754 like floating-point arithmetic.

In the present paper, we consider the summation of real numbers, where each individual sum is somehow perturbed; abandoning the concept of a fixed floating-point grid as in (1.1). We prove an estimate equivalent to (1.5) for an arbitrary subset $\mathbb{F}$ of $\mathbb{R}$ and operation $\tilde{+}$ with the only assumption that

$$a, b \in \mathbb{F}: \quad |(a \tilde{+} b) - (a+b)| \le \min\{|a|, |b|\}. \tag{1.6}$$

[1] Note that the relative error ε_1 with respect to the true result is, in fact, bounded by $\frac{\mathbf{u}}{1+\mathbf{u}}$, see (3.3).

In particular, any rounding to nearest implies (1.6), see (3.2). In fact, our result is more general, not requiring any standard model, no relative rounding error unit nor any assumption on $\mathbb{F}$. Moreover, weaker assumptions than (1.6) are sufficient. The details are given in the next section.

Besides the new estimates, one purpose of this paper is to identify the properties particularly necessary to allow for (1.5). It will turn out that almost none of the previous assumptions on the floating-point system are actually necessary.

We also prove a similar result for an even more general approximation concept with no consideration of (1.6). The corresponding estimate for other rounding modes, in particular for directed rounding, in an IEEE 754 compliant arithmetic follows as a trivial corollary.

The paper is organized as follows. In Section 2 the new estimates for the summation of real numbers are stated. For didactic purposes, first, the surprisingly simple proof for recursive summation is presented. This is followed by the same result for arbitrary order of summation, which implies the estimate (1.5) for rounding to nearest as a corollary. The third new result, the proof of which is again surprisingly simple, covers a similar result for floating-point arithmetic with faithful rounding as a corollary. In particular, this includes a floating-point arithmetic with directed rounding.

The corollaries for floating-point schemes are presented in Section 3, including a result similar to (1.5) for dot products.

2 Error estimates for perturbed sums of real numbers

The first result on summation is formulated so that the corresponding result for floating-point summation in rounding to nearest follows as a corollary. As mentioned, we begin by presenting the proof for recursive summation, followed by the general proof for arbitrary summation order.

Lemma 2.1 *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 $2 \le k \le n$. Furthermore, to every index $k = 1, \dots, n$, define

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

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

$$|\delta_k| \le \big(1 + \sum_{i=1}^{k} \xi_i\big)|x_k| \qquad \text{for} \quad 2 \le k \le n. \tag{2.1}$$

Then $\Delta_n := s_n - \sum_{i=1}^{n} x_i$ satisfies

$$|\Delta_n| \le \sum_{i=1}^{n} |\delta_i| \le \sum_{i=1}^{n} \xi_i \sum_{i=1}^{n} |x_i| \le \sum_{i=1}^{n} |\varepsilon_i| \sum_{i=1}^{n} |x_i|. \tag{2.2}$$

The estimate is sharp in the sense that for arbitrary nonnegative $x_1, \varepsilon_{1\dots n}$ there exist $x_{2\dots n}$ such that (2.1) is satisfied and there are equalities in (2.2).

Remark 2.1 Inequality (2.1) is the only assumption in Lemma 2.1 to be satisfied. Not even an upper bound for the relative perturbations, such as $\varepsilon_i \le 1$, is assumed. Evidently, the validity of (2.1) is implied by property (1.6).

Remark 2.2 Moreover, there is no assumption whatsoever on the size of ε_1 or δ_1. One choice is $\varepsilon_1 = \delta_1 = 0$. In that case the constant in (2.2) is the sum of the $n-1$ relative errors $\varepsilon_{2\ldots n}$, corresponding to the familiar estimates (1.4) or (1.5).

Remark 2.3 Since the input data x_i are real numbers, $\varepsilon_1 = \delta_1 = 0$ might be thought of as the generic choice. However, when evaluating a scalar product using FMA, every product can be considered as a real number, and the very first product would be perturbed as well. In any case, allowing for general ε_1 and δ_1 seems to ease notation.

Remark 2.4 The estimate (2.2) not only regards the local relative errors ε_i of s_i with respect to $x_i + s_{i-1}$ but gives a tighter bound in correspondence to the relative errors ξ_i, which are 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 i-th step.

Proof of Lemma 2.1 The proof is by induction, whereby the result for $n = 1$ is evident. By definition, we have $\Delta_n = \sum_{i=1}^{n} \delta_i$ and therefore

$$|\Delta_n| = \Big|\sum_{i=1}^{n} \delta_i\Big| \le \sum_{i=1}^{n} |\delta_i| = |\delta_n| + \sum_{i=1}^{n-1} |\delta_i|. \tag{2.3}$$

We distinguish two cases. First, assume $|x_n| < \xi_n \sum_{i=1}^{n-1} |x_i|$. The assumption (2.1) and the induction hypothesis for (2.2) imply

$$\begin{aligned}\sum_{i=1}^{n} |\delta_i| &\le \big(1 + \sum_{i=1}^{n} \xi_i\big)|x_n| + \sum_{i=1}^{n-1} \xi_i \sum_{i=1}^{n-1} |x_i| \\ &< \xi_n \sum_{i=1}^{n-1} |x_i| + \sum_{i=1}^{n} \xi_i\, |x_n| + \sum_{i=1}^{n-1} \xi_i \sum_{i=1}^{n-1} |x_i| \\ &= \sum_{i=1}^{n} \xi_i \sum_{i=1}^{n} |x_i|.\end{aligned}$$

Secondly, suppose $\xi_n \sum_{i=1}^{n-1} |x_i| \le |x_n|$. Then the definition of ξ_n and the induction hypothesis give

$$\begin{aligned}\sum_{i=1}^{n} |\delta_i| &= \xi_n \sum_{i=1}^{n} |x_i| + \xi_n \sum_{i=1}^{n-1} |\delta_i| + \sum_{i=1}^{n-1} |\delta_i| \\ &\le \xi_n \sum_{i=1}^{n} |x_i| + \xi_n \sum_{i=1}^{n-1} \xi_i \sum_{i=1}^{n-1} |x_i| + \sum_{i=1}^{n-1} \xi_i \sum_{i=1}^{n-1} |x_i| \\ &\le \xi_n \sum_{i=1}^{n} |x_i| + \sum_{i=1}^{n-1} \xi_i\, |x_n| + \sum_{i=1}^{n-1} \xi_i \sum_{i=1}^{n-1} |x_i| \\ &= \sum_{i=1}^{n} \xi_i \sum_{i=1}^{n} |x_i|.\end{aligned}$$

Finally, using $s_{k-1} = \sum_{i=1}^{k-1}(x_i + \delta_i)$,

$$|\varepsilon_k| = \frac{|\delta_k|}{|x_k + s_{k-1}|} \geq \frac{|\delta_k|}{|x_k| + \sum_{i=1}^{k-1}(|x_i| + |\delta_i|)} = \xi_k \tag{2.4}$$

proves (2.2).

In order to show that this estimate is sharp, let arbitrary nonnegative $x_1, \varepsilon_{1...n}$ be given, and define $x_k := \varepsilon_k \sum_{i=1}^{k-1} x_i$ for $k = 2, \ldots, n$. The nonnegativity of $x_{1,...,n}$, $\delta_1 = \varepsilon_1 x_1 \geq 0$, and $\delta_k = \varepsilon_k(x_k + s_{k-1}) \geq 0$ implies $s_k = \sum_{i=1}^{k}(x_i + \delta_i) = \sum_{i=1}^{k}|x_i| + \sum_{i=1}^{k}|\delta_i|$ and therefore $\varepsilon_k = \xi_k$ for all $k = 1, \ldots, n$. We proceed by induction to prove that the assumption (2.1) of Lemma 2.1 is satisfied and that there are equalities in (2.2). For $n = 1$, we have $s_1 - x_1 = \delta_1 = \varepsilon_1 x_1$. Suppose that, up to $k \leq n-1$, (2.2) is satisfied with equalities, i.e.

$$\sum_{i=1}^{k} \delta_i = \sum_{i=1}^{k} \xi_i \sum_{i=1}^{k} x_i = \sum_{i=1}^{k} \varepsilon_i \sum_{i=1}^{k} x_i.$$

Then, by the definition $x_k = \varepsilon_k \sum_{i=1}^{k-1} x_i$, the nonnegativity of all quantities, and the induction hypothesis, we have

$$\begin{aligned}
\delta_k &= \varepsilon_k \big(x_k + \sum_{i=1}^{k-1} x_i + \sum_{i=1}^{k-1} \delta_i\big) \\
&= \varepsilon_k x_k + \varepsilon_k \sum_{i=1}^{k-1} x_i + \varepsilon_k \sum_{i=1}^{k-1} \varepsilon_i \sum_{i=1}^{k-1} x_i \\
&= \varepsilon_k x_k + x_k + \sum_{i=1}^{k-1} \varepsilon_i x_k \\
&= \big(1 + \sum_{i=1}^{k} \varepsilon_i\big) x_k = \big(1 + \sum_{i=1}^{k} \xi_i\big) x_k
\end{aligned}$$

for $k = 2, \ldots, n$. Hence, (2.1) is satisfied with equality. Finally, using $\delta_n = \varepsilon_n(x_n + s_{n-1}) = \varepsilon_n(\sum_{i=1}^{n} x_i + \sum_{i=1}^{n-1} \delta_i)$,

$$\begin{aligned}
s_n - \sum_{i=1}^{n} x_i &= \delta_n + \sum_{i=1}^{n-1} \delta_i \\
&= \varepsilon_n \sum_{i=1}^{n} x_i + \varepsilon_n \sum_{i=1}^{n-1} \delta_i + \sum_{i=1}^{n-1} \delta_i \\
&= \varepsilon_n \sum_{i=1}^{n} x_i + \varepsilon_n \sum_{i=1}^{n-1} \varepsilon_i \sum_{i=1}^{n-1} x_i + \sum_{i=1}^{n-1} \varepsilon_i \sum_{i=1}^{n-1} x_i \\
&= \varepsilon_n \sum_{i=1}^{n} x_i + \sum_{i=1}^{n-1} \varepsilon_i x_n + \sum_{i=1}^{n-1} \varepsilon_i \sum_{i=1}^{n-1} x_i \\
&= \sum_{i=1}^{n} \varepsilon_i \sum_{i=1}^{n} x_i,
\end{aligned}$$

so that also (2.2) holds true with equalities. □

The following result generalizes Lemma 2.1 to summation in arbitrary order. The proof requires slightly more effort.

Theorem 2.1 *Let a binary tree T with root r be given. For a node j of T, denote the set of inner nodes of the subtree with root j by N_j, and the set of its leaves by L_j. To each leaf $i \in L_r$ associate a real number x_i, and let to each inner node $j \in N_r$ a real number ε_j be associated. Define*

$$s_j := \begin{cases} x_j & \text{if } j \in L_r \\ (s_{\mathrm{left}(j)} + s_{\mathrm{right}(j)})(1+\varepsilon_j) & \text{if } j \in N_r, \end{cases}$$

where $\mathrm{left}(j)$ *and* $\mathrm{right}(j)$ *denote the left and right child of an inner node j, respectively. Furthermore, define for all inner nodes j*

$$\delta_j := s_j - s_{\mathrm{left}(j)} - s_{\mathrm{right}(j)}$$

as well as, with the convention $\frac{0}{0} := 0$,

$$\xi_j := \frac{|\delta_j|}{\sum_{i\in L_j} |s_i| + \sum_{i\in N_j\setminus\{j\}} |\delta_i|}.$$

Suppose

$$|\delta_j| \le \min_{k\in\{\mathrm{left}(j),\mathrm{right}(j)\}} \Big\{ |s_k| + \sum_{i\in N_j\setminus N_k} \xi_i \sum_{i\in L_k} |s_i| \Big\} \tag{2.5}$$

is true for all inner nodes j. Then $\Delta_r := s_r - \sum_{i\in L_r} s_i$ satisfies

$$|\Delta_r| \le \sum_{i\in N_r} |\delta_i| \le \sum_{i\in N_r} \xi_i \sum_{i\in L_r} |s_i| \le \sum_{i\in N_r} |\varepsilon_i| \sum_{i\in L_r} |s_i|. \tag{2.6}$$

The estimate is sharp in the sense that for arbitrary $\varepsilon_j \in [0,\,1]$ there exists a tree T such that (2.5) *is satisfied and there are equalities in* (2.6).

Remark 2.5 Inequality (2.5) is the only assumption in Theorem 2.1 to be satisfied.

Remark 2.6 By (2.6), Theorem 2.1 gives an error estimate for the summation of n real numbers x_i in arbitrary order. Since $\sum_{i\in L_r} s_i = \sum_{i\in L_r} x_i$ this estimate is equivalent to (2.2) for recursive summation in Lemma 2.1 with $\varepsilon_1 = 0$.

Remark 2.7 In correspondance to Lemma 2.1, the estimate (2.6) not only regards the local relative errors ε_i but gives a tighter bound in using the relative errors ξ_i with respect to the maximally possible sum of absolute values of the leaves and absolute values of the perturbations δ_i in the respective subtree.

Proof of Theorem 2.1 We proceed by induction. For $n = 1$ leaf, there is no inner node and therefore $s_r - \sum_{i\in L_r} s_i = 0$. Let a tree with n nodes and root r be given, and suppose (2.6) is true for trees with up to $n-1$ leaves. Denote the children of the root r by p and q. Furthermore, denote

$$e_t := \sum_{i\in N_t} \xi_i \qquad \text{for} \quad t \in \{p,q,r\},$$

where N_t is empty if t is a leave. Apparently, $e_r = e_p + e_q + \xi_r$. After possible renaming, we henceforth assume without loss of generality that $\sum_{i \in L_q} |s_i| \le \sum_{i \in L_p} |s_i|$. The left inequality in (2.6) is evident by

$$s_r - \sum_{i \in L_r} s_i = \sum_{i \in N_r} \delta_i, \tag{2.7}$$

and the right inequality follows by the same argument (2.4) as in the proof of Lemma 2.1. In the following, we will prove the second inequality in (2.6), namely $\sum_{i \in N_r} |\delta_i| \le e_r \sum_{i \in L_r} |s_i|$.

We distinguish two cases. First, suppose

$$\sum_{i \in L_q} |s_i| \le \xi_r \sum_{i \in L_p} |s_i|. \tag{2.8}$$

By (2.5), the induction hypothesis, (2.7), (2.8), and $\sum_{i \in L_q} |s_i| \le \sum_{i \in L_p} |s_i|$, we derive

$$\begin{aligned}
\sum_{i \in N_r} |\delta_i| &= |\delta_r| + \sum_{i \in N_p} |\delta_i| + \sum_{i \in N_q} |\delta_i| \\
&\le |s_q| + (e_r - e_q) \sum_{i \in L_q} |s_i| + \sum_{i \in N_p} |\delta_i| + \sum_{i \in N_q} |\delta_i| \\
&\le |s_q| + e_r \sum_{i \in L_q} |s_i| + \sum_{i \in N_p} |\delta_i| \\
&\le \sum_{i \in L_q} |s_i| + \sum_{i \in N_q} |\delta_i| + e_r \sum_{i \in L_q} |s_i| + \sum_{i \in N_p} |\delta_i| \\
&\le \sum_{i \in L_q} |s_i| + e_q \sum_{i \in L_q} |s_i| + e_r \sum_{i \in L_q} |s_i| + e_p \sum_{i \in L_p} |s_i| \\
&\le \xi_r \sum_{i \in L_p} |s_i| + e_q \sum_{i \in L_p} |s_i| + e_r \sum_{i \in L_q} |s_i| + e_p \sum_{i \in L_p} |s_i| \\
&= e_r \sum_{i \in L_r} |s_i|.
\end{aligned}$$

Secondly, assume the opposite of (2.8), namely

$$\xi_r \sum_{i \in L_p} |s_i| < \sum_{i \in L_q} |s_i|. \tag{2.9}$$

Then the definition of ξ_r, the induction hypothesis, (2.9), and $\xi_r \sum_{i \in L_q} |s_i| \le \xi_r \sum_{i \in L_p} |s_i| < \sum_{i \in L_q} |s_i| \le \sum_{i \in L_p} |s_i|$ imply

$$\begin{aligned}
\sum_{i \in N_r} |\delta_i| &= \xi_r \sum_{i \in L_r} |s_i| + \xi_r \sum_{i \in N_p} |\delta_i| + \xi_r \sum_{i \in N_q} |\delta_i| + \sum_{i \in N_p} |\delta_i| + \sum_{i \in N_q} |\delta_i| \\
&\le \xi_r \sum_{i \in L_r} |s_i| + \xi_r e_p \sum_{i \in L_p} |s_i| + \xi_r e_q \sum_{i \in L_q} |s_i| + e_p \sum_{i \in L_p} |s_i| + e_q \sum_{i \in L_q} |s_i| \\
&\le \xi_r \sum_{i \in L_r} |s_i| + e_p \sum_{i \in L_q} |s_i| + e_q \sum_{i \in L_p} |s_i| + e_p \sum_{i \in L_p} |s_i| + e_q \sum_{i \in L_q} |s_i| \\
&= e_r \sum_{i \in L_r} |s_i|,
\end{aligned}$$

which proves (2.6).

The sharpness of this estimate can be shown by applying the same induction argument that was already used in the proof of Lemma 2.1. Consider a tree T whose right children are all leaves. To the only leaf that is a left child, associate some nonnegative real number and let all other leaves satisfy $s_{\mathrm{right}(j)} = \varepsilon_j \sum_{i \in L_{\mathrm{left}(j)}} s_i$. For any subtree which consist of a single leaf the equalities in (2.6) are evident. The structure of T implies that $N_j \setminus N_{\mathrm{left}(j)} = \{j\}$ and $N_{\mathrm{right}(j)} = \emptyset$ for all inner nodes j. Assume equalities in (2.6) for all proper subtrees of the tree with root j, then, using $\varepsilon_j \leq 1$,

$$\begin{aligned} s_{\mathrm{right}(j)} + \sum_{i \in N_j} \xi_i s_{\mathrm{right}(j)} &= \varepsilon_j \sum_{i \in L_{\mathrm{left}(j)}} s_i + \varepsilon_j \sum_{i \in N_j} \xi_i \sum_{i \in L_{\mathrm{left}(j)}} s_i \\ &\leq \sum_{i \in L_{\mathrm{left}(j)}} s_i + \sum_{i \in N_j} \xi_i \sum_{i \in L_{\mathrm{left}(j)}} s_i \\ &= \sum_{i \in L_{\mathrm{left}(j)}} s_i + \sum_{i \in N_{l\mathring{e}ft(j)}} \delta_i + \xi_j \sum_{i \in L_{\mathrm{left}(j)}} s_i \\ &= s_{\mathrm{left}(j)} + \xi_j \sum_{i \in L_{\mathrm{left}(j)}} s_i . \end{aligned}$$

Thus, the minimum in (2.5) is always attained for $k = \mathrm{right}(j)$, so that (2.5) is equivalent to

$$\delta_j \leq s_{\mathrm{right}(j)} + \sum_{i \in N_j} \xi_i s_{\mathrm{right}(j)} .$$

Hence, for the considered tree, the assumption (2.5) coincides with (2.1). The remainder argument follows straightforwardly from the last two equalities in the proof of Lemma 2.1. □

The final result is designed to imply an error estimate for floating-point summation with faithful rounding as a corollary. That covers in particular directed rounding.

It is known for this case that the relative error may grow exponentially, so that some restriction on the number of summands is mandatory. However, in the following theorem that restriction is not explicit but hidden in the assumption (2.13). Remarkably, that assumption depends only on the size of the real sum of the absolute values of the input data.

Theorem 2.2 *Let a binary tree T with n leaves be given. To each leaf associate a real number x_i and to each inner node associate a real number s_k forming vectors $x \in \mathbb{R}^n$ and $s \in \mathbb{R}^{n-1}$. Denote by σ_k the sum of the values associated to the children of an inner node k, and define*

$$\delta_k := s_k - \sigma_k \qquad \mathit{for} \quad 1 \leq k \leq n-1. \tag{2.10}$$

Let nonnegative real numbers λ, μ be given such that

$$\lambda \leq \sum_{i=1}^{n} |x_i| < \mu . \tag{2.11}$$

Define for $1 \le k \le n-1$

$$\varepsilon_k := \begin{cases} \frac{|\delta_k|}{\lambda} & \textit{if } |\sigma_k| < \mu, \\ \frac{|\delta_k|}{\mu} & \textit{otherwise,} \end{cases} \tag{2.12}$$

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

$$\sum_{i=1}^{n-1} \varepsilon_i \le \frac{\mu - \lambda}{\lambda} \tag{2.13}$$

implies for $1 \le k \le n-1$ *that*

$$|\sigma_k| \le \sum_{i=1}^{n} |x_i| + \mu \sum_{i=1}^{n-1} \varepsilon_i < \frac{\mu^2}{\lambda}. \tag{2.14}$$

Furthermore, for r denoting the root of T,

$$\Big|s_r - \sum_{i=1}^{n} x_i\Big| \le \sum_{i=1}^{n-1} |\delta_i| \le \sum_{i=1}^{n-1} \varepsilon_i \sum_{i=1}^{n} |x_i|. \tag{2.15}$$

Remark 2.8 Inequality (2.13) is the only assumption in Theorem 2.2 to be satisfied.

Remark 2.9 Again, it is possible to replace σ_k with the sum of absolute values of the leaves x_i and absolute values of perturbations δ_i in the respective subtree. However, for reasons of comprehensibility and because ε_k remains to be defined relative to either λ or μ, here we refrain from applying this modification.

Proof of Theorem 2.2 For an inner node k denote the set of leaves by L_k and the set of inner nodes without k itself by N_k. Then (2.10) gives

$$\sigma_k = \sum_{i \in L_k} x_i + \sum_{i \in N_k} \delta_i \qquad \text{for} \quad 1 \le k \le n-1.$$

Thus (2.12), (2.11), and (2.13) imply for all $1 \le k \le n-1$

$$|\sigma_k| \le \sum_{i=1}^{n} |x_i| + \sum_{i=1}^{n-1} |\delta_i| \le \sum_{i=1}^{n} |x_i| + \mu \sum_{i=1}^{n-1} \varepsilon_i < \mu + \frac{\mu - \lambda}{\lambda} \mu = \frac{\mu^2}{\lambda}$$

and prove (2.14).

Note that $|\delta_j| = \varepsilon_j \lambda$ if $|\sigma_j| < \mu$. Thus, if $|\sigma_j| < \mu$ for all $1 \le j \le n-1$, then $\lambda \le \sum_{i=1}^{n} |x_i|$ implies (2.15).

Otherwise, there is an inner node p with $|\sigma_p| \ge \mu$. The assumptions imply that there also exists an inner node q in the subtree with root p whose both children are leaves. Thus $|\sigma_q| = |\sum_{i \in L_q} x_i| \le \sum_{i=1}^{n} |x_i| < \mu$. It follows that there exists a node k such that

$$|\sigma_k| \ge \mu \qquad \text{and} \qquad \forall j \in N_k: \quad |\sigma_j| < \mu.$$

Denote $J := N_k$ and $J' := \{1, \dots, n-1\} \backslash J$, so that $|\delta_j| = \varepsilon_j \lambda$ for all $j \in J$ and

$$\sum_{i=1}^{n} |x_i| + \sum_{i \in J} \varepsilon_i \lambda \geq \sum_{i \in L_k} |x_i| + \sum_{i \in J} \varepsilon_i \lambda \geq |\sigma_k| \geq \mu.$$

Then, using (2.13), it follows

$$\begin{aligned}
\sum_{i=1}^{n-1} |\delta_i| \leq \sum_{i \in J} \varepsilon_i \lambda + \sum_{i \in J'} \varepsilon_i \mu &= \sum_{i=1}^{n-1} \varepsilon_i \mu + \sum_{i \in J} \varepsilon_i (\lambda - \mu) \\
&\leq \sum_{i=1}^{n-1} \varepsilon_i \left(\sum_{i=1}^{n} |x_i| + \sum_{i \in J} \varepsilon_i \lambda \right) + \sum_{i \in J} \varepsilon_i (\lambda - \mu) \\
&= \sum_{i=1}^{n-1} \varepsilon_i \sum_{i=1}^{n} |x_i| + \sum_{i \in J} \varepsilon_i \left(\sum_{i=1}^{n-1} \varepsilon_i \lambda + \lambda - \mu \right) \\
&\leq \sum_{i=1}^{n-1} \varepsilon_i \sum_{i=1}^{n} |x_i| + \sum_{i \in J} \varepsilon_i \left(\frac{\mu - \lambda}{\lambda} \lambda + \lambda - \mu \right) \\
&= \sum_{i=1}^{n-1} \varepsilon_i \sum_{i=1}^{n} |x_i|,
\end{aligned}$$

which proves (2.15). □

3 Application to floating-point systems

Let $\mathbb{F} \subseteq \mathbb{R}$ be an arbitrary set of real numbers. A mapping $\tilde{+}\colon \mathbb{F} \times \mathbb{F} \to \mathbb{F}$ is called "nearest-addition" if for all $a, b \in \mathbb{F}$:

$$|(a \,\tilde{+}\, b) - (a + b)| = \inf\{|f - (a + b)| \colon f \in \mathbb{F}\}. \tag{3.1}$$

That implies that $a \,\tilde{+}\, b = a + b$ if $a + b \in \mathbb{F}$. Note that there is no assumption whatsoever on the set $\mathbb{F}$. In particular, for $a, b, c, d \in \mathbb{F}$ and $a + b = c + d$, not necessarily $a \,\tilde{+}\, b = c \,\tilde{+}\, d$.

In Lemma 2.1 and Theorem 2.1 the only assumptions to be satisfied are (2.1) and (2.5), respectively. Both are a trivial consequence of (1.6). For a "nearest-addition" and $a, b \in \mathbb{F}$, property (1.6), in turn, follows by

$$\begin{aligned}
|(a \,\tilde{+}\, b) - (a + b)| &= \inf\{|f - (a + b)| \colon f \in \mathbb{F}\} \\
&\leq \min\{|f - (a + b)| \colon f \in \{a, b\}\} \\
&= \min\{|a|, |b|\}.
\end{aligned} \tag{3.2}$$

As a consequence, Lemma 2.1 and Theorem 2.1 hold true for any "nearest-addition" (any rounding of ties) over some arbitrary set $\mathbb{F} \subseteq \mathbb{R}$.

For a k-digit floating-point number system in base β following the IEEE 754 standard, the relative rounding error unit is $\mathbf{u} := 0.5\beta^{1-k}$. In the first standard model for error analysis of floating-point operations the error is defined relative to the real result, whereas in the second standard model the definition is relative to the floating-point result. The rounding error unit $\mathbf{u}$ is applicable to both standard models (1.2). However, if we limit our consideration on the first standard model, the sharp estimate

$$|(a\tilde{+}b)-(a+b)| \le \frac{\mathbf{u}}{1+\mathbf{u}}|a+b| \tag{3.3}$$

holds true, see [6, p. 232]. Indeed, (3.3) follows without any reference to a floating-point grid simply as a consequence of (1.2) and rounding to nearest [10]. Therefore, we will use the constant $\frac{\mathbf{u}}{1+\mathbf{u}}$ in the following corollaries where rounding to nearest is assumed. Moreover note that if the result of a floating-point addition is in the underflow range there is no rounding error, i.e. the result is equal to the real addition.

Corollary 3.1 *Let $\mathbb{F}$ be a k-digit floating-point number system in base β, and denote by s the result of a floating-point summation of $a_1,\ldots,a_n \in \mathbb{F}$ in some nearest-addition in any order. Then, with $\mathbf{u} = 0.5\beta^{1-k}$,*

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

Proof The estimate is an immediate consequence of Theorem 2.1, inequality (3.2), and the relative rounding error bound $\mathbf{u}/(1+\mathbf{u})$ for the first standard model. □

For the application to dot products, we adapt a similar proof as in [5, Theorem 4.2].

Corollary 3.2 *Let $\mathbb{F}$ be a k-digit floating-point number system in base β, and denote by s the result of a floating-point dot product in rounding to nearest of $a,b \in \mathbb{F}^n$. Then, barring underflow,*

$$|s-a^Tb| \le n\mathbf{u}\sum_{j=1}^{n}|a_jb_j|.$$

Proof Denote the floating-point approximation to a_jb_j by $x_j \in \mathbb{F}$, so that s is the floating-point sum of the x_j. Then, abbreviating $\mathbf{v} = \mathbf{u}/(1+\mathbf{u})$, Corollary 3.1 implies $|s-\sum_{j=1}^{n} x_j| \le (n-1)\mathbf{v}\sum_{j=1}^{n}|x_j|$. Moreover, rounding to nearest implies $|x_j - a_jb_j| \le \mathbf{v}|a_jb_j|$, so that $|x_j| \le (1+\mathbf{v})|a_jb_j|$. Hence,

$$|s-a^Tb| \le |s-\sum_{i=1}^{n} x_i| + |\sum_{i=1}^{n} x_i - a_jb_j| \le (\mathbf{v}+(n-1)\mathbf{v}(1+\mathbf{v}))\sum_{j=1}^{n}|a_jb_j|,$$

and it is easily checked that $\mathbf{v}+(n-1)\mathbf{v}(1+\mathbf{v}) \le n\mathbf{v}(1+\mathbf{v}) \le n\frac{\mathbf{v}}{1-\mathbf{v}} = n\mathbf{u}$. □

One application of Theorem 2.2 is the computation of a sum in directed rounding. An improvement of the classical Wilkinson bound has been proved by Bünger in [7, Theorem 1]. However, his arguments are involved making heavy use of specific properties of IEEE 754 floating-point arithmetic. A generalization of his result is as follows.

A mapping $\tilde{+}: \mathbb{F}\times\mathbb{F} \to \mathbb{F}$ is called "faithful-addition" if, for all $a,b \in \mathbb{F}$, $a\tilde{+}b$ is the only floating-point number in the convex hull $(a+b)\underline{\cup}(a\tilde{+}b)$. As before, necessarily $a\tilde{+}b = a+b$ if $a+b \in \mathbb{F}$.

Note that for a faithful-addition upward and downward rounding may be arbitrarily mixed, and, as before, $a,b,c,d \in \mathbb{F}$ and $a+b = c+d$ does not imply $a\tilde{+}b = c\tilde{+}d$.

Corollary 3.3 *Let $\mathbb{F}$ be a k-digit floating-point number system in base β, and denote by s the result of a floating-point summation of $x_1,\dots,x_n \in \mathbb{F}$ using some faithful-addition. If $n \le 1+\frac{\beta-1}{2}\mathbf{u}^{-1} = 1+(\beta-1)\beta^{k-1}$, then*

$$\Big|s-\sum_{j=1}^{n} x_j\Big| \le (n-1)\cdot 2\mathbf{u}\sum_{j=1}^{n}|x_j|.$$

Remark 3.1 For infinite exponent range, recursive summation, and rounding upwards, sufficiently small ε and $x=(1,\varepsilon,\varepsilon,\dots)$ show that the estimate is sharp. The restriction on n is mandatory, and it is sharp in the sense that the upper bound on n cannot be replaced by the next larger integer.

Proof of Corollary 3.3 Let $m\in\mathbb{Z}$ such that $\lambda := \beta^m \le \sum_{i=1}^{n}|x_i| < \beta^{m+1} =: \mu$. Let σ_k be as in Theorem 2.2, and denote by $\mathrm{ufp}(\sigma_k)$ the largest power of β being less than or equal to $|\sigma_k|$. If $|\sigma_k|<\mu$, then $\mathrm{ufp}(\sigma_k)\le\lambda$ and (2.12) implies

$$\varepsilon_k = \frac{|\delta_k|}{\lambda} \le \frac{|\delta_k|}{\mathrm{ufp}(\sigma_k)} = \frac{|\mathrm{fl}(\sigma_k)-\sigma_k|}{\mathrm{ufp}(\sigma_k)} \le 2\mathbf{u},$$

and otherwise $\mu \le |\sigma_k| < \mu^2/\lambda = \beta\mu$ shows $\mathrm{ufp}(\sigma_k)=\mu$ and

$$\varepsilon_k = \frac{|\delta_k|}{\mu} = \frac{|\delta_k|}{\mathrm{ufp}(\sigma_k)} = \frac{|\mathrm{fl}(\sigma_k)-\sigma_k|}{\mathrm{ufp}(\sigma_k)} \le 2\mathbf{u}.$$

Thus, all ε_k are bounded by $2\mathbf{u}$. Additionally, the limit on n implies

$$\sum_{i=1}^{n-1}\varepsilon_i \le (n-1)2\mathbf{u} \le \beta-1 = \frac{\mu-\lambda}{\lambda},$$

so that the assumption (2.13) in Theorem 2.2 is satisfied. Thus,

$$\Big|s-\sum_{i=1}^{n}x_i\Big| \le \sum_{i=1}^{n-1}|\delta_i| \le (n-1)\cdot 2\mathbf{u}\sum_{i=1}^{n}|x_i|,$$

and the proof is finished. □

As a final example, consider a logarithmic number system

$$\mathbb{F} := \{\pm c^k\colon k\in\mathbb{Z}\}\cup\{0\} \tag{3.4}$$

for some $1<c\in\mathbb{R}$. Let $a,b\in\mathbb{F}$ with $c^m \le |a+b| < c^{m+1}$. For a nearest-addition $\tilde{+}\colon \mathbb{F}\times\mathbb{F}\to\mathbb{F}$, it follows, similar to (3.3),

$$\left|\frac{(a\,\tilde{+}\,b)-(a+b)}{a+b}\right| \le \frac{\frac{1}{2}(c^{m+1}-c^m)}{\frac{1}{2}(c^{m+1}+c^m)} = \frac{c-1}{c+1},$$

so that $\frac{c-1}{c+1}<1$ is the relative rounding error unit. Hence, by Theorem 2.1, the result s of a floating-point summation of $a_1,\dots,a_n\in\mathbb{F}$ in some nearest-addition in any order satisfies

$$\Big|s-\sum_{j=1}^{n}a_j\Big| \le (n-1)\,\frac{c-1}{c+1}\sum_{j=1}^{n}|a_j|.$$

For a faithful-addition $\tilde{+}: \mathbb{F} \times \mathbb{F} \to \mathbb{F}$, it follows

$$\left| \frac{(a \tilde{+} b) - (a+b)}{a+b} \right| < \frac{c^{m+1} - c^m}{c^m} = c - 1, \tag{3.5}$$

so that, as in the proof of Corollary 3.3 and with the notation of Theorem 2.2, we conclude $\varepsilon_k \le c-1$ for all $1 \le k \le n-1$. Note that, although (3.5) is a strict inequality, the right-hand side $c-1$ cannot be replaced by a smaller constant. Thus the assumption (2.13) is surely satisfied if $(n-1)(c-1) \le \frac{\mu-\lambda}{\lambda} = c-1$, limiting the number of summands to $n \le 2$.

Indeed, for $1 = c^0 \in \mathbb{F}$ and $e \in \mathbb{F}$ with $0 < e < c-1$, rounding upwards implies $(1 \tilde{+} e) \tilde{+} e = c \tilde{+} e = c^2$. Thus, for small enough e, the left-hand side in (2.15) comes arbitrarily close to $c^2 - 1$, whereas the right-hand side tends to $2(c-1)$. Hence (2.15) is not necessarily satisfied for $n > 2$.

A reason is as follows. Let N denote the maximal number of summands so that the condition (2.13) in Theorem 2.2 is surely satisfied. We concluded that $N = 1 + (\beta - 1)\beta^{k-1}$ for a k-digit floating-point number system, and $N = 2$ for a logarithmic number system. The reason is that in the first case the interval $[\lambda, \mu]$ in (2.11) corresponds to some $I := [\beta^k, \beta^{k+1}]$. In this interval, the absolute error of a faithful-addition is constantly bounded by $2\mathbf{u} = \beta^{1-k}$, and N is equal to the number $\beta^k - \beta^{k-1} + 1$ of elements of $\mathbb{F}$ in I. Similarily, for the logarithmic number system, $[\lambda, \mu]$ corresponds to $I := [c^k, c^{k+1}]$ consisting of $N = 2$ elements of $\mathbb{F}$.

References

1. N. J. Higham. *Accuracy and stability of numerical algorithms*. SIAM Publications, Philadelphia, 2nd edition, 2002.
2. *ANSI/IEEE 754-1985: IEEE Standard for Binary Floating-Point Arithmetic*. New York, 1985.
3. *ANSI/IEEE 754-2008: IEEE Standard for Floating-Point Arithmetic*. New York, 2008.
4. C.-P. Jeannerod and S.M. Rump. Improved error bounds for inner products in floating-point arithmetic. *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, 2014.
6. D.E. Knuth. *The Art of Computer Programming: Seminumerical Algorithms*, volume 2. Addison Wesley, Reading, Massachusetts, third edition, 1998.
7. K. Ozaki, T. Ogita, F. Bünger, and S. Oishi. Accelerating interval matrix multiplication by mixed precision arithmetic. *Nonlinear Theory and Its Applications, IEICE*, 6(3):364–376, 2015.
8. S.M. Rump. Error estimation of floating-point summation and dot product. *BIT Numerical Mathematics*, 52(1):201–220, 2012.
9. S.M. Rump and C.-P. Jeannerod. Improved backward error bounds for LU and Cholesky factorizations. *SIAM J. Matrix Anal. & Appl. (SIMAX)*, 35(2):684–698, 2014.
10. S.M. Rump and M. Lange. On the Definition of Unit Roundoff. *BIT Numerical Mathematics*, 56(1):309317, 2015.