Improved error bounds for inner products in floating-point arithmetic

Claude-Pierre Jeannerod, Siegfried M. Rump

IMPROVED ERROR BOUNDS FOR INNER PRODUCTS IN FLOATING-POINT ARITHMETIC

CLAUDE-PIERRE JEANNEROD* AND SIEGFRIED M. RUMP†

Abstract. Given two floating-point vectors x, y of dimension n and assuming rounding to nearest, we show that if no underflow or overflow occurs, any evaluation order for inner product returns a floating-point number $\widehat{r}$ such that $|\widehat{r} - x^T y| \leqslant nu|x|^T|y|$ with u the unit roundoff. This result, which holds for any radix and with no restriction on n, can be seen as a generalization of a similar bound given in [7] for recursive summation in radix 2, namely $|\widehat{r} - x^T e| \leqslant (n-1)u|x|^T e$ with $e = [1, 1, \ldots, 1]^T$. As a direct consequence, the error bound for the floating-point approximation $\widehat{C}$ of classical matrix multiplication with inner dimension n simplifies to $|\widehat{C} - AB| \leqslant nu|A||B|$.

Key words. floating-point inner product, rounding error analysis, unit in the first place

AMS subject classifications. 65G50

1. Introduction. Consider IEEE standard floating-point arithmetic, with a set $\mathbb{F}$ of finite floating-point numbers in radix β and precision p, and with a round-to-nearest function $\mathrm{fl} : \mathbb{R} \to \mathbb{F} \cup \{\pm\infty\}$. Then by the standard model, for any real t and in the absence of underflow and overflow,

$$\mathrm{fl}(t) = t(1+\delta), \qquad |\delta| \leqslant u, \tag{1.1}$$

with $u = \frac{1}{2}\beta^{1-p}$ denoting the unit roundoff.

For ease of readability, in this introduction we loosely denote by float(expression) the computed approximation obtained by evaluating the expression in floating-point arithmetic, *no matter what the order of evaluation.* In particular, when the expression is a sum or a sum of products, such a notation covers recursive summation and pairwise summation, but also any other scheme. For example, $x^T y$ with $x, y \in \mathbb{F}^6$ may be evaluated as $(((x_1y_1 + x_4y_4) + x_5y_5) + x_3y_3) + (x_6y_6 + x_2y_2)$, or alike.

For inner products the standard model leads to the most classical a priori bound on the absolute error. Given input vectors $x, y \in \mathbb{F}^n$ and barring underflow and overflow, the repeated application of (1.1) gives

$$|\mathrm{float}(x^T y) - x^T y| \leqslant B_n|x|^T|y| \qquad \text{with } B_n = (1+u)^n - 1 \text{ for all } n; \tag{1.2}$$

see for example [4, p. 62]. Often, B_n is bounded by the commonly used quantity

$$\gamma_n = \frac{nu}{1-nu} \qquad \text{if } nu < 1. \tag{1.3}$$

As noted by Higham in [4, p. 77] a better bound on B_n, attributed to Kiełbasiński and Schwetlick [5, 6], is

$$\gamma'_n = \frac{nu}{1-nu/2} \qquad \text{if } nu < 2. \tag{1.4}$$

Notice that the condition on n is weaker.

*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).

†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@tu-harburg.de).

In all cases[1] the bounds above are of the form $|\mathrm{float}(x^T y) - x^T y| \leqslant (nu + \mathcal{O}(u^2))\,|x|^T|y|$, involving a term in $\mathcal{O}(u^2)$ and possibly with a restriction on n. In the special case of summation, however, this can be improved: assuming that overflow does not occur, it has been shown in [7] that for radix 2 and with $\mathrm{float}(\sum)$ denoting recursive summation, the absolute error satisfies

$$|\mathrm{float}(\textstyle\sum_{i=1}^n x_i) - \sum_{i=1}^n x_i| \leqslant (n-1)u \sum_{i=1}^n |x_i| \qquad \text{for all } n. \tag{1.5}$$

In other words, this gives an absolute error bound that is both unconditional with respect to n and free of any $\mathcal{O}(u^2)$ term.

In this paper we first note that the summation error bound in (1.5) in fact holds for any radix β and no matter what the order of evaluation. Furthermore, in the more general case of inner products, we present the following unconditional and $\mathcal{O}(u^2)$-free bound: for any radix β and no matter what the order of evaluation,

$$|\mathrm{float}(x^T y) - x^T y| \leqslant nu\,|x|^T|y| \qquad \text{for all } n. \tag{1.6}$$

To achieve this result, which holds assuming underflow and overflow does not occur, we use the notion of ufp (unit in the first place) together with tighter error bounds than by the standard model (1.1).

The factors $(n-1)u$ and nu in (1.5) and (1.6), respectively, are both sharp up to $\mathcal{O}(u^2)$. However, from a practical point of view, both do, in general, grossly overestimate the true error. Nevertheless it may be worth noting that the quantities γ_{n-1} and γ_n can simply be replaced by $(n-1)u$ and nu, respectively.

2. Assumptions, notation, and preliminary properties. Throughout this paper we make the customary assumptions that $\beta \geqslant 2$ and $p \geqslant 2$. This implies that the unit roundoff $u = \frac{1}{2}\beta^{1-p}$ satisfies $u \leqslant 1$. Note also that $\mathbb{F}$ being a set of finite IEEE standard floating-point numbers, it is in particular symmetric:

$$s \in \mathbb{F} \qquad \Rightarrow \qquad -s \in \mathbb{F}. \tag{2.1}$$

For the round-to-nearest function fl, we take any map from $\mathbb{R}$ to $\mathbb{F} \cup \{\pm\infty\}$ such that for all real t and in the absence of overflow,

$$|\mathrm{fl}(t) - t| = \min_{s \in \mathbb{F}} |s - t|. \tag{2.2}$$

In particular, in this paper we only require (2.2), which makes no assumption on the way of breaking ties.

A consequence of (2.1) and (2.2) is the following lemma, which will be used to prove the error bound for inner products.

LEMMA 2.1. *For $t \in \mathbb{R}$ and in the absence of overflow,*

$$\big|\mathrm{fl}(t) - t\big| = \big|\mathrm{fl}(|t|) - |t|\big|.$$

Proof. Assume $t < 0$, for otherwise the result is trivial. If t is not exactly halfway between two consecutive floating-point numbers, then (2.2) defines $\mathrm{fl}(t)$ uniquely and using (2.1) gives $\mathrm{fl}(|t|) = -\mathrm{fl}(t)$, from which the conclusion follows. If t is a midpoint

[1] Sometimes the unpleasant denominators in (1.3) and (1.4) are avoided by using $B_n \leqslant 1.01nu$ provided $nu \leqslant 0.01$. More generally, if $nu \leqslant \epsilon$ with $\epsilon > 0$ small enough (say, $\epsilon < 1.79$), then $B_n \leqslant (1+\epsilon)nu$.

then, using again (2.1) and writing s for the largest floating-point number less than t, we have $|\mathrm{fl}(t) - t| = t - s = |s| - |t| = \big|\mathrm{fl}(|t|) - |t|\big|$. □

To establish the error bound for inner products another useful tool is Rump's *unit in the first place* [8], defined by

$$\mathrm{ufp}(t) = \begin{cases} 0 & \text{if } t = 0, \\ \beta^{\lfloor \log_\beta |t| \rfloor} & \text{if } t \in \mathbb{R}\backslash\{0\}; \end{cases}$$

in other words, $\mathrm{ufp}(t)$ is either zero or the largest integer power of β not larger than $|t|$. Note also that $\mathrm{ufp}(t) \in \mathbb{F}$ provided $\mathrm{fl}(|t|)$ is between the smallest and largest positive floating-point numbers. Combining this definition with (2.2), we see that for $t \in \mathbb{R}$ and in the absence of underflow and overflow,

$$|\mathrm{fl}(t) - t| \leqslant u \cdot \mathrm{ufp}(t) \leqslant u|t|. \tag{2.3}$$

Here the second inequality is equivalent to (1.1) and leads to the classical relative error bound u, while the first one gives $u \cdot \mathrm{ufp}(t)/|t|$. This improves the classical bound by a factor of up to almost β and reflects the well-known effect of wobbling precision [4, p. 39].

The inequalities in (2.3) apply in particular to the case where t is the sum or the product of two floating-point numbers, at least in the absence of underflow and overflow. For floating-point addition, however, we have the following sharper bound, which holds even if underflow occurs.

LEMMA 2.2. *Let $a, b \in \mathbb{F}$. If $\mathrm{fl}(a+b)$ does not overflow, its absolute error satisfies*

$$|\mathrm{fl}(a + b) - (a + b)| \leqslant \min\{|a|, |b|, u \cdot \mathrm{ufp}(a + b)\}.$$

This result is classical at least in radix 2; see for example Shewchuk [9, Lemma 1] and Rump [7, Lemma 3.1]. For completeness, we give a proof in radix β.

Proof. From (2.2) it follows that $|\mathrm{fl}(a + b) - (a + b)| \leqslant |f - (a + b)|$ for all $f \in \mathbb{F}$. This inequality thus holds in particular for $f = a$ and for $f = b$, leading to the upper bounds $|b|$ and $|a|$, respectively. On the other hand, in the absence of underflow then (2.3) holds, while if $a + b$ is in the subnormal range of $\mathbb{F}$ then it equals $\mathrm{fl}(a + b)$; see for example [3, Theorem 3.4.1] or [4, solution to Problem 2.19]. □

3. Extending Rump's forward error bound for summation. In radix 2 and for recursive summation Rump [7, Theorem 3.3] shows that the leading coefficient γ_{n-1} that typically appears in the forward error bound can be improved to $(n - 1)u$, provided overflow does not occur. We show below that such a bound holds in radix β and also no matter what the order of evaluation.

PROPOSITION 3.1. *For $n \in \mathbb{N}_{>0}$ and given $x_1, \ldots, x_n \in \mathbb{F}$, any order of evaluation of the sum $\sum_{i=1}^n x_i$ produces an approximation $\widehat{r}$ such that, in the absence of overflow,*

$$|\widehat{r} - \textstyle\sum_{i=1}^n x_i| \leqslant (n - 1)u \sum_{i=1}^n |x_i|.$$

Before proving this bound, note that it is valid for any n and in particular does not require $(n - 1)u < 1$.

Proof. The proof is by induction on n, the case $n = 1$ being trivial. For $n \geqslant 2$, we assume the result is true up to $n - 1$, and we fix one evaluation order in dimension n. The approximation $\widehat{r}$ obtained with this order has the form $\widehat{r} = \mathrm{fl}(\widehat{r}_1 + \widehat{r}_2)$, where $\widehat{r}_j$ is the result of a floating-point evaluation of $r_j = \sum_{i \in I_j} x_i$, for $j = 1, 2$ and with $\{I_1, I_2\}$

a partition of the set $I = \{1, 2, \dots, n\}$. For $j = 1, 2$ let also $e_j = \widehat{r}_j - r_j$ and let n_j be the cardinality of I_j. Finally, let $r = \sum_{i=1}^n x_i$, $s = \widehat{r}_1 + \widehat{r}_2$, and $\delta = \mathrm{fl}(s) - s = \widehat{r} - s$.

Since $r = r_1 + r_2$, we have $\widehat{r} - r = \delta + e_1 + e_2$. Now, $1 \leqslant n_1, n_2 \leqslant n-1$ and the inductive assumption leads to

$$|\widehat{r} - r| \leqslant |\delta| + u\big((n_1 - 1)\widetilde{r}_1 + (n_2 - 1)\widetilde{r}_2\big),$$

where $\widetilde{r}_j = \sum_{i \in I_j} |x_i|$ for $j = 1, 2$. Since $n = n_1 + n_2$ and $\sum_{i=1}^n |x_i| = \widetilde{r}_1 + \widetilde{r}_2$, it remains to check that $|\delta| \leqslant u\bar{s}$ for $\bar{s} = n_2\widetilde{r}_1 + n_1\widetilde{r}_2$.

Following Rump's proof of [7, Theorem 3.3], we assume first that $\widetilde{r}_2 \leqslant u\widetilde{r}_1$. Since $u \leqslant 1$ this implies $\widetilde{r}_2 \leqslant \widetilde{r}_1$. On the other hand, Lemma 2.2 gives $|\delta| \leqslant |\widehat{r}_2|$. Hence

$$|\delta| \leqslant |\widehat{r}_2| = |e_2 + r_2| \leqslant |e_2| + \widetilde{r}_2 \leqslant u\big((n_2 - 1)\widetilde{r}_2 + \widetilde{r}_1\big) \leqslant un_2\widetilde{r}_1 \leqslant u\bar{s}.$$

When $\widetilde{r}_1 \leqslant u\widetilde{r}_2$ the same conclusion can be obtained by swapping the indices 1 and 2 in the above analysis.

Assume now that $u\widetilde{r}_1 < \widetilde{r}_2$ and $u\widetilde{r}_2 < \widetilde{r}_1$. By Lemma 2.2 we have $|\delta| \leqslant u|\widehat{r}_1 + \widehat{r}_2|$. Furthermore, $|\widehat{r}_1 + \widehat{r}_2| \leqslant |e_1| + \widetilde{r}_1 + |e_2| + \widetilde{r}_2$ with $|e_j| \leqslant (n_j - 1)u\widetilde{r}_j \leqslant (n_j - 1)\widetilde{r}_k$ for $(j, k) \in \{(1, 2), (2, 1)\}$. Hence $|\widehat{r}_1 + \widehat{r}_2| \leqslant \bar{s}$ and the conclusion follows. □

4. Forward error bounds for inner products. An improvement over the bounds γ_n and γ'_n in (1.3) and (1.4), respectively, is obtained by a direct, naive application of Rump's bound for summation: barring underflow and overflow and writing $z_i = x_i y_i$ and $\widehat{z}_i = \mathrm{fl}(x_i y_i)$ for $i = 1, \dots, n$, we see by (1.1) that any evaluation order yields $\widehat{r} \in \mathbb{F}$ such that $|\widehat{r} - x^T y| \leqslant |\widehat{r} - \sum_{i=1}^n \widehat{z}_i| + \sum_{i=1}^n |\widehat{z}_i - z_i| \leqslant \gamma''_n |x|^T |y|$ with

$$\gamma''_n = nu + (n-1)u^2 \qquad \text{for all } n.$$

Note that this bound is valid for any n and is better than B_n in (1.2); in fact, it is easy to see that

$$nu \leqslant \gamma''_n \leqslant B_n \leqslant \gamma'_n \leqslant \gamma_n.$$

Also, this bound γ''_n is not restricted to inner products and holds more generally when summing the rounded values $\mathrm{fl}(z_i)$ of the entries of a real vector $z = [z_i] \in \mathbb{R}^n$. However, in any case γ''_n still has a term in $\mathcal{O}(u^2)$. The result below shows that this quadratic term can always be removed, thus implying the inner product bound (1.6) as a particular case.

PROPOSITION 4.1. *For $n \in \mathbb{N}_{>0}$, given $z_1, \dots, z_n \in \mathbb{R}$ and in the absence of underflow and overflow, any order of evaluation of the sum $\sum_{i=1}^n \mathrm{fl}(z_i)$ produces an approximation $\widehat{r}$ such that*

$$|\widehat{r} - \textstyle\sum_{i=1}^n z_i| \leqslant nu \sum_{i=1}^n |z_i|.$$

Before proving this result, two remarks are in order. First, it holds for any n and in particular does not require $nu < 1$. Second, unlike for Proposition 3.1 we now assume underflow does not occur and thus, in particular, that all the $\mathrm{fl}(z_i)$ are normalized floating-point numbers. This assumption is necessary even in the special case $z_i = x_i y_i$ of an inner product. For example, writing $e_{\min}$ for the smallest exponent of a given floating-point format, assuming $e_{\min} < -p$, and defining $x_i = y_i = \beta^{e_{\min}}$ for $i = 1, \dots, n$, then all the $\mathrm{fl}(x_i y_i)$ are equal to zero. This means $\widehat{r}$ is equal to zero and the error, which should be bounded by $nu|x|^T|y|$, satisfies $|\widehat{r} - x^T y| = |x|^T|y|$.

Proof. The proof is by induction on n. If $n = 1$ then $\widehat{r}$ is equal to $\mathrm{fl}(z_1)$, so that the identity in (1.1) gives the result. For $n \geqslant 2$, we assume the result is true up to $n-1$ and choose a fixed evaluation order in dimension n. Then, for this specific evaluation order we have $\widehat{r} = \mathrm{fl}(\widehat{r}_1 + \widehat{r}_2)$, each $\widehat{r}_j$ being the result of a floating-point evaluation of a sum $r_j = \sum_{i \in I_j} z_i$ with I_j defined as in the proof of Proposition 3.1. Similarly to the proof of Proposition 3.1, let us define further $n_j = |I_j|$, $e_j = \widehat{r}_j - r_j$, $s = \widehat{r}_1 + \widehat{r}_2$, and $\delta = \mathrm{fl}(s) - s$. For $r = \sum_{i=1}^{n} z_i = r_1 + r_2$, the identity $\widehat{r} - r = \delta + e_1 + e_2$ still holds and, since $1 \leqslant n_j \leqslant n-1$, the induction hypothesis now gives

$$|\widehat{r} - r| \leqslant |\delta| + u(n_1 \widetilde{r}_1 + n_2 \widetilde{r}_2)$$

with $\widetilde{r}_j = \sum_{i \in I_j} |z_i|$, $j = 1, 2$. Thus, we are left with checking that $|\delta| \leqslant u\bar{s}$ for $\bar{s} = n_2 \widetilde{r}_1 + n_1 \widetilde{r}_2$. To do this, we will consider separately three cases depending on how $\widetilde{r}_i$ compares to $u\widetilde{r}_j$. Although this scheme is similar to the one employed in Proposition 3.1 for the summation of n floating-point numbers, the third case (see (4.1) below) is now more involved. Indeed, the constraint $\bar{s}$ is the same as before but the bounds we have on the $|e_j|$ are now larger by $u\widetilde{r}_j$; we will handle this harder case by combining ufp's and Lemma 2.1.

Assume first that $\widetilde{r}_2 \leqslant u\widetilde{r}_1$. Since $|\delta| \leqslant |\widehat{r}_2|$ by Lemma 2.2 and since $\widetilde{r}_2 \leqslant \widetilde{r}_1$,

$$|\delta| \leqslant |e_2| + \widetilde{r}_2 \leqslant u(n_2 \widetilde{r}_2 + \widetilde{r}_1) \leqslant u(n_2 \widetilde{r}_1 + \widetilde{r}_2) \leqslant u\bar{s}.$$

The case where $\widetilde{r}_1 \leqslant u\widetilde{r}_2$ is handled similarly by exchanging the roles of $\widetilde{r}_1$ and $\widetilde{r}_2$.

Assume now that

$$u\widetilde{r}_1 < \widetilde{r}_2 \qquad \text{and} \qquad u\widetilde{r}_2 < \widetilde{r}_1. \tag{4.1}$$

If $\mathrm{ufp}(s) \leqslant \bar{s}$, then by (2.3) we have $|\delta| \leqslant u \cdot \mathrm{ufp}(s) \leqslant u\bar{s}$, as wanted. Thus, it remains to consider the case where $\bar{s} < \mathrm{ufp}(s)$. In this case applying Lemma 2.1 gives

$$\begin{aligned} |\delta| &= \big|\mathrm{fl}(|s|) - |s|\big| \\ &\leqslant |s| - \mathrm{ufp}(s), \qquad \text{since } \mathrm{ufp}(s) \in \mathbb{F} \text{ and using (2.2)}, \\ &< |s| - \bar{s}. \end{aligned}$$

Furthermore, by the definition of s and $\bar{s}$,

$$\begin{aligned} |s| - \bar{s} &= |\widehat{r}_1 + \widehat{r}_2| - n_2 \widetilde{r}_1 - n_1 \widetilde{r}_2 \\ &\leqslant |e_1| + |e_2| - (n_2 - 1)\widetilde{r}_1 - (n_1 - 1)\widetilde{r}_2, \quad \text{using } |\widehat{r}_j| = |e_j + r_j| \leqslant |e_j| + \widetilde{r}_j, \\ &\leqslant |e_1| + |e_2| - u\big((n_1 - 1)\widetilde{r}_1 + (n_2 - 1)\widetilde{r}_2\big), \quad \text{using (4.1) and } n_j \geqslant 1. \end{aligned}$$

Since by the inductive assumption $|e_j| \leqslant n_j u\widetilde{r}_j$, we deduce that

$$|\delta| < |s| - \bar{s} \leqslant u(\widetilde{r}_1 + \widetilde{r}_2) \leqslant u\bar{s}.$$

This completes the proof. □

By applying Proposition 4.1 to $z_i = x_i y_i$ with $x_i, y_i \in \mathbb{F}$ for $i = 1, \ldots, n$, we arrive at the announced forward error bound (1.6) for inner products:

THEOREM 4.2. *For $n \in \mathbb{N}_{>0}$ and given $x, y \in \mathbb{F}^n$, any order of evaluation of the inner product $x^T y$ produces an approximation $\widehat{r}$ such that, if no underflows or overflows are encountered,*

$$|\widehat{r} - x^T y| \leqslant nu|x|^T|y|.$$

5. Concluding remarks. We mention some direct applications of our forward error bound (1.6). First, if the input vectors $x, y \in \mathbb{F}^n$ satisfy $x_i y_i \geqslant 0$ for $i = 1, \ldots, n$, then a *relative error* bound for their inner product is nu for all n. This generalizes a similar result obtained for $n = 2$ by Brent, Percival, and Zimmermann in the context of complex floating-point multiplication to arbitrary n [1, pp. 1470-1471].

Another consequence of (1.6) is the following *backward error* result for inner products similar to [4, (3.4)].

COROLLARY 5.1. *For $n \in \mathbb{N}_{>0}$ and given $x, y \in \mathbb{F}^n$, any order of evaluation of the inner product $x^T y$ produces an approximation $\widehat{r}$ which, if no underflows or overflows are encountered, has the following form:*

$$\widehat{r} = (x + \Delta x)^T y = x^T (y + \Delta y)$$

for some $\Delta x, \Delta y \in \mathbb{R}^n$ such that $|\Delta x| \leqslant nu|x|$ and $|\Delta y| \leqslant nu|y|$.

Proof. It suffices to show the first identity for a given evaluation order. By Theorem 4.2, the computed inner product has the form $\widehat{r} = x^T y + \theta\,|x|^T |y|$ for some $\theta \in \mathbb{R}$ such that $|\theta| \leqslant nu$. Hence

$$\begin{aligned}
\widehat{r} &= \sum_{i=1}^{n} \big(x_i y_i + \theta |x_i y_i|\big) \\
&= \sum_{i=1}^{n} x_i y_i (1 + \theta_i), \qquad \theta_i = \operatorname{sign}(x_i y_i)\theta, \quad i = 1, \ldots, n, \\
&= (x + \Delta x)^T y \qquad \text{with } \Delta x \in \mathbb{R}^n \text{ having its } i\text{th entry equal to } x_i \theta_i.
\end{aligned}$$

Since $|\theta_i| = |\theta|$ for all i, we have $|\Delta x| \leqslant nu|x|$, from which the result follows. ∎

Of course, for summation a similar backward error result can be deduced in exactly the same way, starting from Proposition 3.1: no matter what the evaluation order and in the absence of overflow, the computed approximation of the sum of the entries of $x \in \mathbb{F}^n$ satisfies $\widehat{r} = \sum_{i=1}^{n} (x + \Delta x)_i$ with $|\Delta x| \leqslant (n-1)u|x|$.

Finally, given Theorem 4.2 and Corollary 5.1, it is straightforward to improve upon the classical backward and forward error bounds associated with, respectively, matrix-vector and matrix-matrix products [4, §3.5]:

- Given $A \in \mathbb{F}^{m \times n}$ and $x \in \mathbb{F}^n$, let $\widehat{y}$ be the approximation to Ax obtained after m inner products in dimension n, each of them being performed in an arbitrary evaluation order. If no underflow or overflow occurs then

$$\widehat{y} = (A + \Delta A)x, \qquad |\Delta A| \leqslant nu|A|.$$

- Given $A \in \mathbb{F}^{m \times n}$ and $B \in \mathbb{F}^{n \times p}$, let $\widehat{C}$ be the approximation to their product AB returned by using mp inner products in dimension n in any order of evaluation.[2] Then, in the absence of underflow and overflow, we have

$$|\widehat{C} - AB| \leqslant nu|A||B|$$

  as well as the corresponding normwise bounds

$$\|\widehat{C} - AB\|_\alpha \leqslant nu\|A\|_\alpha \|B\|_\alpha, \qquad \alpha = 1, \infty, F,$$

  with $\|\cdot\|_F$ denoting the Frobenius norm.

[2] Note that this covers in particular blocking strategies, but not more sophisticated methods as by Strassen [10], Coppersmith and Winograd [2], or Vassilevska Williams [11].

To summarize, in all these error bounds the factor nu replaces the classical factor $\gamma_n = nu/(1-nu)$, thus removing $\mathcal{O}(u^2)$ terms, and it is valid for any n, i.e., without the assumption $nu < 1$. Both factors nu and γ_n hold no matter what the order of evaluation and, when this order corresponds to recursive summation, they are sharp up to $\mathcal{O}(u^2)$. For example, assume that the radix is even (which holds in practice as β is either 2 or 10) and consider the n-dimensional vectors x and y given by

$$x^T = [1-u, 1-2u, \ldots, 1-2u] \qquad \text{and} \qquad y^T = [1+2u, u, \ldots, u].$$

Then $x, y \in \mathbb{F}^n$, $(1-u)(1+2u) \in [1, 1+u)$, and $(1-2u)u \in \mathbb{F}$, so that recursive summation of $\mathrm{fl}(x_1y_1), \mathrm{fl}(x_2y_2), \ldots, \mathrm{fl}(x_ny_n)$ yields $\widehat{r} = 1$ and a relative error of the form $nu - \mathcal{O}(u^2)$. (This example applies to any tie-breaking rule, but note that if tie breaking is 'to even' then $1-2u$ can be replaced by 1 in the last $n-1$ entries of x.)

However, for other evaluation orders, those factors are sometimes far from being best possible. For example, for pairwise summation repeated application of the standard model leads to the factor γ_ℓ with $\ell = \lceil \log_2 n \rceil + 1$; see [4, p. 64]. Thus it remains to understand how our ufp-based error analysis can be adapted to such highly parallel evaluation schemes, and whether γ_ℓ—which requires $\ell u < 1$ and has a $\mathcal{O}(u^2)$ term, can be replaced by an unconditional factor of the form ℓu.

Acknowledgments. We thank the referees for their helpful comments and suggestions.

REFERENCES

[1] R. P. Brent, C. Percival, and P. Zimmermann, *Error bounds on complex floating-point multiplication*, Mathematics of Computation, 76 (2007), pp. 1469–1481.
[2] D. Coppersmith and S. Winograd, *Matrix multiplication via arithmetic progressions*, J. Symbolic Computation, 9 (1990), pp. 251–280.
[3] J. R. Hauser, *Handling floating-point exceptions in numeric programs*, ACM Trans. Program. Lang. Syst., 18 (1996), pp. 139–174.
[4] N. J. Higham, *Accuracy and Stability of Numerical Algorithms*, SIAM, Philadelphia, PA, USA, second ed., 2002.
[5] A. Kiełbasiński and H. Schwetlick, *Numerische Lineare Algebra: Eine Computerorientierte Einführung*, VEB Deutscher, Berlin, 1988.
[6] ———, *Numeryczna Algebra Liniowa: Wprowadzenie do Obliczeń Zautomatyzowanych*, Wydawnictwa Naukowo-Techniczne, Warszawa, 1992.
[7] S. M. Rump, *Error estimation of floating-point summation and dot product*, BIT, 52 (2012), pp. 201–220.
[8] S. M. Rump, T. Ogita, and S. Oishi, *Accurate floating-point summation part I: Faithful rounding*, SIAM Journal on Scientific Computing, 31 (2008), pp. 189–224.
[9] J. R. Shewchuk, *Adaptive precision floating-point arithmetic and fast robust geometric predicates*, Discrete and Computational Geometry, 18 (1997), pp. 305–363.
[10] V. Strassen, *Gaussian elimination is not optimal*, Numer. Math., 13 (1969), pp. 354–356.
[11] V. Vassilevska Williams, *Multiplying matrices faster than Coppersmith-Winograd*, in Proceedings of the 44th symposium on Theory of Computing, STOC'12, New York, NY, USA, 2012, ACM, pp. 887–898.