Further remarks on Kahan summation with decreasing ordering

Claude-Pierre Jeannerod

FURTHER REMARKS ON KAHAN SUMMATION WITH DECREASING ORDERING

CLAUDE-PIERRE JEANNEROD

Abstract. We consider Kahan's compensated summation of n floating-point numbers ordered as $|x_1| \geqslant \cdots \geqslant |x_n|$, and show that in IEEE arithmetic a large relative error can occur for a dimension as small as $n = 4$. This answers a question raised in particular by Priest [7] and Higham [3, Problem 4.10].

Introduction

Kahan's compensated summation [4] is a common way to produce more accurate floating-point sums without having to resort explicitly to extended precision. Its main feature is a small backward relative error, that holds in a very general setting: if each addition or subtraction has relative error at most u, then the exact sum of n floating-point numbers $x_1, \ldots, x_n$ is approximated by $\widehat{s}_n = x_1(1+\epsilon_1) + \cdots + x_n(1+\epsilon_n)$ with $\max_k |\epsilon_k| = O(u)$ as $u \to 0$; to first order in u this backward error bound is independent of n, which contrasts with the backward error of recursive summation, whose largest value can be up to about $(n-1)u$. (See for example [6, p. 615] and [1] for detailed proofs.)

What about the forward relative error? In general it can be large, and this already with $n = 3$ and IEEE arithmetic. Thus, from a worst-case perspective on the forward relative error and given arbitrary x_k, Kahan summation does not perform better than recursive summation: for all $n \geqslant 3$, damaging cancellation can occur despite compensation. Following Higham's early analysis and experiments [2], Priest studied further the case where the x_k are arranged in decreasing order of magnitude [7, §4.1]: he showed that if $|x_1| \geqslant |x_2| \geqslant |x_3|$, then Kahan's computed sum $\widehat{s}_3$ has relative error $O(u)$, which is now in clear contrast with recursive summation $(x_1 + x_2) + x_3$, that can produce a totally wrong answer even for decreasing ordering. Priest also gave an $n = 6$ example for which $|x_1| \geqslant \cdots \geqslant |x_n|$ and the relative error of $\widehat{s}_n$ is large in IEEE arithmetic, and asked for the smallest such n. This question of *the smallest dimension for which Kahan summation with decreasing ordering can yield a large relative error* also appears in [3, Problem 4.10] (and already in the 1996 edition, as Problem 4.9).

In what follows, we review first the case $n = 3$ and then, assuming decreasing ordering, the case $n = 4$. In both cases we give input examples leading to large relative errors in IEEE arithmetic, thus answering in particular the above question of Higham and Priest. Hopefully, these remarks are just the first step towards a more general study we are aiming at and which is devoted to identifying and explaining rapid transitions from 'always highly accurate' to 'possibly totally wrong' in various compensated and augmented-precision algorithms and for various arithmetics.

December 11, 2021.

2010 *Mathematics Subject Classification.* Primary 65G50.

Large relative error for $n = 3$

When $n = 3$ Kahan summation approximates the exact sum $s_3 = x_1 + x_2 + x_3$ using five floating-point operations as

$$\widehat{s}_2 := \mathrm{fl}(x_1 + x_2), \quad \widehat{c}_2 := \mathrm{fl}\big(\mathrm{fl}(\widehat{s}_2 - x_1) - x_2\big), \quad \widehat{y}_3 := \mathrm{fl}(x_3 - \widehat{c}_2), \quad \widehat{s}_3 := \mathrm{fl}(\widehat{s}_2 + \widehat{y}_3).$$

Here, $\widehat{c}_2$ is an estimate of the error of the first addition $\mathrm{fl}(x_1 + x_2)$, that is used as a correcting term to update x_3 before proceeding to the second addition. With fl denoting rounding to nearest, this implies that if $|x_2| \ll |x_1|, |x_3|$, then one may have $\widehat{s}_2 = x_1$, $\widehat{c}_2 = -x_2$, $\widehat{y}_3 = x_3$, and $\widehat{s}_3 = \mathrm{fl}(x_1 + x_3)$; if in addition x_2 is nonzero and $x_1 + x_3 = 0$, then we arrive at $s_3 = x_2 \neq 0 = \widehat{s}_3$ and the forward relative error $|\widehat{s}_3 - s_3|/|s_3|$ is one.

In practice, for $\mathbb{F}$ a floating-point set in base β even and precision $p \geqslant 2$, and for any round-to-nearest map $\mathrm{fl} : \mathbb{R} \to \mathbb{F}$, one can take

$$(x_1, x_2, x_3) = (1, \epsilon, -1), \qquad \epsilon = (\beta/2 - 2u)\beta^{-p-1},$$

where $u = \beta^{1-p}/2$. Since $\epsilon \in \mathbb{F} \cap (0, u/\beta)$, we have $\mathrm{fl}(1+\epsilon) = 1$ and $\mathrm{fl}(-1+\epsilon) = -1$, which implies $\widehat{s}_2 = 1$, $\widehat{c}_2 = -\epsilon$, $\widehat{y}_3 = -1$, and $\widehat{s}_3 = 0 \neq \epsilon = s_3$.

On this example recursive summation behaves just as badly as Kahan summation (same undeserved zero result), but in general its behavior can be worse in the sense that a large relative error can occur even for inputs ordered decreasingly: for $\beta = 2 \leqslant p$, if $x_1 = 1$ and $x_2 = x_3 = -(1 - u)/2$, then $|x_1| \geqslant |x_2| \geqslant |x_3|$, $x_1 + x_2 + x_3 = u$, and $\mathrm{fl}(\mathrm{fl}(x_1 + x_2) + x_3)$ is either $u/2$ or $3u/2$.

Large relative error for $n = 4$ and decreasing ordering

When $n = 4$, Kahan summation first evaluates $x_1 + x_2 + x_3$ as shown in the previous paragraph and then incorporates x_4 by computing further

$$\widehat{c}_3 := \mathrm{fl}\big(\mathrm{fl}(\widehat{s}_3 - \widehat{s}_2) - \widehat{y}_3\big), \qquad \widehat{y}_4 := \mathrm{fl}(x_4 - \widehat{c}_3), \qquad \widehat{s}_4 := \mathrm{fl}(\widehat{s}_3 + \widehat{y}_4).$$

For $\mathbb{F}$ a floating-point set in base two and precision $p \geqslant 4$, let $u = 2^{-p}$ and

$$(x_1, x_2, x_3, x_4) = (1 + 4u, 1 + 2u, -1 + u, -1 + u).$$

Clearly, these x_k are in $\mathbb{F}$ and satisfy $|x_1| \geqslant |x_2| \geqslant |x_3| \geqslant |x_4|$. Furthermore, with fl denoting round to nearest even, one can check that $\widehat{s}_2 = 2 + 8u$, $\widehat{c}_2 = 2u$, $\widehat{y}_3 = -1$, $\widehat{s}_3 = 1 + 8u$, $\widehat{c}_3 = 0$, $\widehat{y}_4 = -1 + u$, and $\widehat{s}_4 = 9u$. Since the exact sum is $s_4 = 8u$, the resulting relative error $|\widehat{s}_4 - s_4|/|s_4|$ equals $1/8$.

For round to nearest with ties to away, the same x_k lead to $\widehat{y}_3 = -1 - 2u$, $\widehat{s}_3 = 1 + 6u$, and $\widehat{s}_4 = 7u$ (the other intermediate quantities being the same as for round to nearest even). Hence the relative error is $1/8$ in this case as well. It is also possible to set x_1 and x_2 to $1 + 2u$ and 1, respectively, in order to obtain a relative error of $1/4$.

Several remarks can be done about these examples. First, although the inputs we have chosen lead to large relative errors, they imply mostly exact computations: out of the nine floating-point operations performed by Kahan summation in dimension $n = 4$, only two of them are inexact, namely, the first and third ones, $\mathrm{fl}(x_1 + x_2)$ and $\mathrm{fl}(x_3 - \widehat{c}_2)$.

Second, these examples can be used to show that a large relative error is also possible for Kahan's modified version [5] of his initial 1965 summation scheme,

which for $n = 4$ returns $\mathrm{fl}(\widehat{s}_4 - \widehat{c}_4)$ instead of $\widehat{s}_4$, where

$$\widehat{c}_4 := \mathrm{fl}\big(\mathrm{fl}(\widehat{s}_4 - \widehat{s}_3) - \widehat{y}_4\big)$$

is the correcting term associated with the last addition $\widehat{s}_4 := \mathrm{fl}(\widehat{s}_3 + \widehat{y}_4)$. Since on our input examples this last addition is exact, we have $\widehat{c}_4 = 0$ ans thus

$$\mathrm{fl}(\widehat{s}_4 - \widehat{c}_4) = \widehat{s}_4.$$

We note in passing that without the decreasing ordering assumption, Kahan's modified summation can be highly inaccurate already in dimension $n = 3$. To see this, it is enough to consider the $(1, -\epsilon, 1)$ example of the previous section, for which the last operation $\widehat{s}_3 := \mathrm{fl}(\widehat{s}_2 + \widehat{y}_3)$ is exact and thus gives $\widehat{c}_3 = 0$.

Third, by Priest's result we know that the relative error of $\widehat{s}_3$ must be $O(u)$. But on our examples we have more than this and for round to nearest even, for example, $\widehat{s}_3$ turns out to be the correctly-rounded value of the exact sum $s_3 = x_1 + x_2 + x_3$, that is, $\widehat{s}_3 = \mathrm{fl}(s_3)$. Recalling that the operations producing $\widehat{s}_3$, $\widehat{c}_3$, $\widehat{y}_4$, and $\widehat{s}_4$ are exact, we deduce that $\widehat{c}_3 = 0$, $\widehat{y}_4 = x_4$, and $\widehat{s}_4 = \widehat{s}_3 + x_4$. Therefore, the relative error of $\widehat{s}_4$ satisfies

$$\frac{|\widehat{s}_4 - s_4|}{|s_4|} = \frac{|\mathrm{fl}(s_3) - s_3|}{|s_3 + x_4|},$$

and the absolute error $|\mathrm{fl}(s_3) - s_3|$, which is at most u for $1 \leqslant s_3 < 2$, is magnified by the factor $1/|s_3 + x_4|$, which is of order $1/u$. Now, one can check that s_3 is a midpoint for $\mathbb{F} \cap [1, 2)$, which implies that $|\mathrm{fl}(s_3) - s_3|$ is exactly u. Hence a sudden change of numerical quality, from best possible for $\widehat{s}_3$ to almost all wrong for $\widehat{s}_4$. (A similar analysis applies to ties to away, with $\widehat{s}_3$ now being a faithful result.)

To summarize, we have shown that in IEEE arithmetic *the smallest dimension for which Kahan summation methods with decreasing ordering can yield a large relative error is* 4. For doing this it was enough to focus on default floating-point characteristics (base two and round to nearest even/away), but we are currently evaluating the impact of other features as well, such as larger bases, alternative tie-breaking rules and alternative roundings, and the so-called augmented addition operation specified in IEEE 754-2019.

References

[1] Eric Hallman and Ilse C. F. Ipsen. Deterministic and probabilistic error bounds for floating point summation algorithms, 2021. arXiv preprint arXiv:2107.01604.

[2] Nicholas J. Higham. The accuracy of floating point summation. *SIAM J. Sci. Comput.*, 14(4): 783–799, 1993.

[3] Nicholas J. Higham. *Accuracy and Stability of Numerical Algorithms*. Second edition, Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2002. xxx+680 pp. ISBN 0-89871-521-0.

[4] W. Kahan. Further remarks on reducing truncation errors. *Comm. ACM*, 8(1):40, 1965.

[5] W. Kahan. A survey of error analysis. In *Proc. IFIP Congress, Ljubljana*, Information Processing 71, North-Holland, Amsterdam, The Netherlands, 1972, pages 1214–1239.

[6] Donald E. Knuth. *The Art of Computer Programming, Volume 2, Seminumerical Algorithms*. Third edition, Addison-Wesley, Reading, MA, USA, 1998. xiii+762 pp. ISBN 0-201-89684-2.

[7] Douglas M. Priest. *On Properties of Floating Point Arithmetics: Numerical Stability and the Cost of Accurate Computations*. PhD thesis, Mathematics Department, University of California, Berkeley, CA, USA, November 1992. 126 pp.

Inria, Laboratoire LIP UMR 5668, Univ. Lyon, CNRS, ENS de Lyon, Inria, UCBL, F-69007 Lyon, France

E-mail address: claude-pierre.jeannerod@inria.fr