FastTwoSum revisited

Claude-Pierre Jeannerod and Paul Zimmermann

ARITH, May 7, 2025

The FastTwoSum algorithm

$$x \leftarrow \circ(a+b)$$
$$z \leftarrow \circ(x-a)$$
$$y \leftarrow \circ(b-z)$$

- A classical way to evaluate the rounding error of finite-precision additions.
- The last two ops aim at producing a suitable estimate of the exact error:

$$y \approx e, \qquad e := a+b-x.$$

- Used since a long time:
  - [1950s] within a fixed-point library for the EDSAC [Gill'51]
  - [1960s] in floating-point arithmetic [Kahan'65, Møller'65]
  - [1970s] analysed for correctly-rounded floating-point arithmetic [Dekker'71]

  and now at the heart of many higher-level algorithms: compensated algorithms, double-word arithmetic, ...

Motivation

Together with Dekker's product, FastTwoSum is used as a building block to double the precision:

- two binary32 numbers provide about 48 bits of accuracy;
- two binary64 numbers provide about 106 bits of accuracy.

Double-double arithmetic is used in many software libraries (CRlibm, routine `exact_add` in LLVM libc `double_double.h`, routine `fasttwosum` in CORE-MATH).

Dekker's analysis and beyond

Theorem (Dekker'71)

If $\beta \leq 3$, $\circ \in \mathrm{RN}$, $e_a \geq e_b$, and no underflow/overflow, then $y = a + b - x$.

- Such conditions suffice for FastTwoSum to implement an error-free transform (EFT) for addition: $x + y = a + b$ with $x = \circ(a + b)$.
- In practice, mostly used with $\beta = 2$, $\circ \in \mathrm{RN}$, and $|a| \geq |b|$.
- Since 1971, many extensions to larger bases, other roundings, weaker restrictions on a and b, and underflows/overflows. [Linnainmaa'74, Virkkunen'80, Priest'91, DaumasRideauThéry'01, RumpOgitaOishi'08, DemmelNguyen'13, GraillatJézéquelPicot'15, BoldoGraillatMuller'17, LangeOishi'20, CorbineauZimmermann'24, ...]

Our paper reviews this literature and extends parts of it, starting from two questions:

- To what extent can we relax Dekker's conditions and still ensure $x + y = a + b$?
- When $x + y \neq a + b$, how large $|x + y - (a + b)|$ can be?

Our contributions

- New sufficient conditions for e, $x - a$, and $b - z$ to be FP numbers.
- Five EFTs depending on the rounding mode in $x \leftarrow \circ(a + b)$.
- Examples showing when exactness is lost.
- Tighter bounds on $|x + y - (a + b)|$.
- For directed roundings, conditions ensuring that $x + y$ is either $a + b$, or the correctly-rounded value of $a + b$ in doubled precision and the same direction.

Notation, definitions, and assumptions

Floating-point number set $\mathbb{F} := \{M \cdot 2^E \ : \ M, E \in \mathbb{Z}, |M| < 2^p\}$.

(base 2, precision p, and no underflow/overflow)

Roundings $\circ, \circ', \circ'' : \mathbb{R} \to \mathbb{F}$ may differ for each of the 3 ops of FastTwoSum.

- IEEE 754 roundings: RNE, RNA, RD, RU, RZ.
- Faithfully rounded results: $\circ(r) \in \{\mathrm{RD}(r), \mathrm{RU}(r)\}$.
- If $\circ = \circ' = \circ''$, we write $[x, y] := \mathrm{FastTwoSum}(a, b, \circ)$.
- For conciseness, we write $\circ \in \mathrm{FR}$ and $\circ \in \mathrm{RN}$ (any tie-breaking rule).

Classical tools and properties for the analysis

- Unit roundoff $u := 2^{-p}$.
- Exponent $e_r := \lfloor \log_2 |r| \rfloor$, $\mathrm{ufp}(r) := 2^{e_r}$, and $\mathrm{ulp}(r) := 2u \cdot \mathrm{ufp}(r)$.
- For $r \neq 0$, $|\circ(r) - r|/\mathrm{ulp}(r)$ is $\leq 1/2$ if $\circ \in \mathrm{RN}$, and < 1 if $\circ \in \mathrm{FR}$.

Relaxing Dekker's conditions $\circ \in \mathrm{RN}$ and $e_a \geq e_b$

We relax these two conditions simultaneously:

- For rounding, we consider directed roundings and even faithful rounding.
- For the input, we allow for $a \in \mathrm{ulp}(b)\mathbb{Z}$ and even $|a| < |b|$.

$a \in \mathrm{ulp}(b)\mathbb{Z}$: x.xxxx × xxxx.xxxx

$|a| < |b|$: x.xxxxxxx × xxxx.xxxx

The relaxed condition $a \in \mathrm{ulp}(b)\mathbb{Z}$ is always true when $e_a \geq e_b$, but also covers some situations where $e_a < e_b$.

Exactness properties

To analyze FastTwoSum, we consider three exactness properties:

(P)	$a + b - x \in \mathbb{F}$	"the error is representable exactly"
(P')	$x - a \in \mathbb{F}$	"the second operation is exact"
(P")	$b - z \in \mathbb{F}$	"the third operation is exact"

Easily checked facts:

- If **(P')** then **(P)** and **(P")** equivalent.
- If **(P')** and any of **(P)** and **(P")** then FastTwoSum is an EFT.

Hence a simple strategy for the analysis: first give sufficient conditions for each of these properties, and then combine them to deduce sufficient conditions for having an EFT.

Sufficient conditions for exactness

Lemma (Sufficient conditions to ensure (P))

For $a, b \in \mathbb{F}$, let $a \in \mathrm{ulp}(b)\mathbb{Z}$ and $x = \circ(a+b)$. If one of the conditions

(i) $\circ \in \mathrm{RN}$,

(ii) $\circ \in \mathrm{FR}$ *and* $e_a - e_b \leq p$,

(iii) $\circ = \mathrm{RD}$ *and* $b \geq 0$,

(iv) $\circ = \mathrm{RU}$ *and* $b \leq 0$,

(v) $\circ = \mathrm{RZ}$ *and* $ab \geq 0$

is satisfied, then the error $e = a + b - x$ is in $\mathbb{F}$.

- The 4 inequalities in (ii)–(v) are needed.
- (i), (ii), (v) already in the literature, but (iii) and (iv) seem to be new.

Sufficient conditions for exactness

For $a, b \in \mathbb{F}$, let $x = \circ(a + b)$ and $z = \circ'(x - a)$.

Lemma (Sufficient conditions to ensure (P'))

If $\circ \in$ FR *and* $a \in \mathrm{ulp}(b)\mathbb{Z}$ *then* $x - a \in \mathbb{F}$.

Lemma (Sufficient conditions to ensure (P''))

If $\circ, \circ' \in$ FR *and* $e_a - e_b \leq p$ *then* $b - z \in \mathbb{F}$.

- Upper bound p is needed and improves upon $p - 3$ from the literature.
- No lower bound on $e_a - e_b$, so **(P'')** always holds in the reversed case $|a| < |b|$.

Error-free transforms

By combining these lemmas, we deduce immediately 5 EFTs, depending on the rounding mode used for the first operation $x = \circ(a + b)$.

For example :

Theorem (EFT for rounding down)

For $a, b \in \mathbb{F}$, let $x = \mathrm{RD}(a + b)$. If the conditions

(i) $a \in \mathrm{ulp}(b)\mathbb{Z}$,

(ii) $b \geq 0$ *or* $e_a - e_b \leq p$

are satisfied then $x + y = a + b$.

- Faitfhful rounding is enough for $\circ'$ and $\circ''$ (last two ops of FastTwoSum).
- $x + y \neq a + b$ is possible when (i) or (ii) not true.
- Similar theorems for $\circ \in \{\mathrm{RN}, \mathrm{RU}, \mathrm{RZ}, \mathrm{FR}\}$.

Tight bounds on $|x + y - (a + b)|$

When we cannot ensure an EFT, how far can $x + y$ be from $a + b$?

Theorem (Tight error bounds when $a \in \mathrm{ulp}(b)\mathbb{Z}$)

For $a, b \in \mathbb{F}$, if $\circ, \circ', \circ'' \in \mathrm{FR}$ and $a \in \mathrm{ulp}(b)\mathbb{Z}$ then

$$|x + y - (a + b)| \leq 2u^2 \mathrm{ufp}(a + b) \leq 2u^2 \mathrm{ufp}(x)$$

and these bounds are asymptotically optimal.

- Hence $x + y$ very close to $a + b$, with bounds 2^p times smaller than for x alone:

$$|x - (a + b)| \leq 2u\,\mathrm{ufp}(a + b) \leq 2u\,\mathrm{ufp}(x).$$

- Asymptotic optimality: for $(a, b) = (1 + 2u, -u^3)$ and RD, it can be checked that $|x + y - (a + b)| = 2u^2 - u^3$, which is equivalent to $2u^2\mathrm{ufp}(x)$ as $u \to 0$.

Tight bounds on $|x + y - (a + b)|$

If $|a| < |b|$ instead of $a \in \mathrm{ulp}(b)\mathbb{Z}$, things can be much worse [CorbineauZimmermann'24]:

- For RN, $|x + y - (a + b)| \leq u|x|$ and this bound is attained: $x + y$ is a priori not a better approximation than x alone.
- For RD, RU, RZ, this is worse: the exact sum $x + y$ can be a poorer approximation than x alone! This is shown by the asymptotically optimal bounds in $3u|x|$ from [CorbineauZimmermann'24], which we have slightly refined for optimality:

Theorem (Optimal error bounds for directed roundings and reversed operands)

For $a, b \in \mathbb{F}$, let $[x, y] := \mathrm{FastTwoSum}(a, b, \circ)$. If $|a| < |b|$ then optimal bounds are

$$|x + y - (a + b)| \leq \begin{cases} \frac{3u}{1+4u}|x| & \text{if } \circ = \mathrm{RZ}, \\ \frac{3u}{1+2u}|x| & \text{if } \circ \in \{\mathrm{RD}, \mathrm{RU}\}. \end{cases}$$

Correctly-rounded results in doubled precision

For $\circ, \circ', \circ'' \in \mathrm{FR}$ and $a \in \mathrm{ulp}(b)\mathbb{Z}$, we have seen that

$$|x + y - (a + b)| \leq 2u^2 \mathrm{ufp}(a + b),$$

which is the bound we would have if $a + b$ were rounded correctly in precision $2p$.

For directed roundings and $a \in \mathrm{ulp}(b)\mathbb{Z}$, it turns out that FastTwoSum always yields

- either $a + b$ exactly,
- or such correctly-rounded values in precision $2p$.

Correctly-rounded results in doubled precision

Theorem (EFT or doubled-precision CR result for rounding down)

For $a, b \in \mathbb{F}$, let $[x, y] = \mathrm{FastTwoSum}(a, b, \mathrm{RD})$. If $a \in \mathrm{ulp}(b)\mathbb{Z}$ then

$$x + y = \begin{cases} a + b & \text{if } b \geq 0 \text{ or } e_a - e_b \leq p, \\ \mathrm{RD}_{2p}(a + b) & \text{otherwise.} \end{cases}$$

- If $a \notin \mathrm{ulp}(b)\mathbb{Z}$ then it can happen that $x + y \notin \{a + b, \mathrm{RD}_{2p}(a + b)\}$.
- Similar theorems for RU and RZ.

Correctly-rounded results in doubled precision

Corollary (Interval obtained when running FastTwoSum with RD and RU)

For $a, b \in \mathbb{F}$, let

$$[\underline{x}, \underline{y}] := \mathrm{FastTwoSum}(a, b, \mathrm{RD}) \quad \textit{and} \quad [\overline{x}, \overline{y}] := \mathrm{FastTwoSum}(a, b, \mathrm{RU}).$$

If $a \in \mathrm{ulp}(b)\mathbb{Z}$ then

$$\left[\underline{x} + \underline{y}, \overline{x} + \overline{y}\right] = \begin{cases} \left[a + b, \mathrm{RU}_{2p}(a + b)\right] & \textit{if } b \geq 0, \\ \left[\mathrm{RD}_{2p}(a + b), a + b\right] & \textit{if } b < 0. \end{cases}$$

- The interval $[\underline{x} + \underline{y}, \overline{x} + \overline{y}]$ can thus take only two very specific forms.
- BEWARE: can fail for some variants of FastTwoSum.

Variants

The FastTwoSum scheme we have analyzed corresponds to the parenthesization

$$x + (b - (x - a)),$$

but variants are sometimes considered:

[V1] $x + (b + (a - x))$,

[V2] $x - ((x - a) - b)$.

With directed roundings (which are not anti-symmetric), these variants can behave very differently even if $a \in \mathrm{ulp}(b)\mathbb{Z}$: for example, for RD,

- with V1, $x + y$ is still either $a + b$ or $\mathrm{RD}_{2p}(a + b)$.
- This is not true anymore for V2.

Conclusion

Results obtained so far

- New sufficient conditions for exactness that make it easy to deduce new EFTs
- Asymptotically optimal error bounds when $a \in \mathrm{ulp}(b)\mathbb{Z}$ and faithful rounding
- Optimal error bounds when $|a| < |b|$ and directed roundings
- Further insight into the interval $[\underline{x} + \underline{y}, \overline{x} + \overline{y}]$ with RD/RU

On-going work

- Analysis of TwoSum beyond rounding to nearest
- Larger bases and underflow/overflow