Optimal Bounds for Floating-Point Addition in Constant Time

Mak Andrlon[1] Peter Schachte[1]
Harald Søndergaard[1] Peter J. Stuckey[2]

[1]School of Computing and Information Systems
The University of Melbourne

[2]Faculty of Information Technology
Monash University

26th IEEE Symposium on Computer Arithmetic
Kyoto, Japan, June 2019

Problem

When does $x \oplus y = z$ hold?

Assumptions:

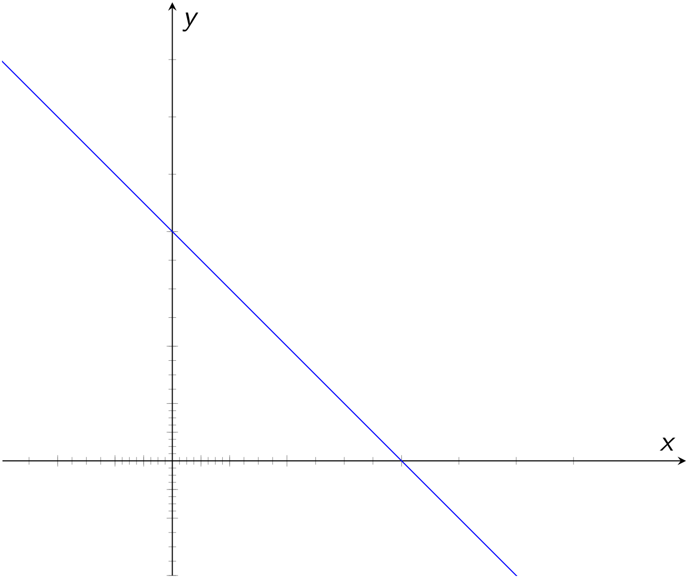
1. x, y and z are drawn from intervals X, Y and Z.
2. IEEE 754 numbers with radix-β, precision p, exponents e_{min} to e_{max}.
3. Rounding function $\mathrm{fl} : \overline{\mathbb{R}} \to \overline{\mathbb{F}}$ is nondecreasing and faithful.

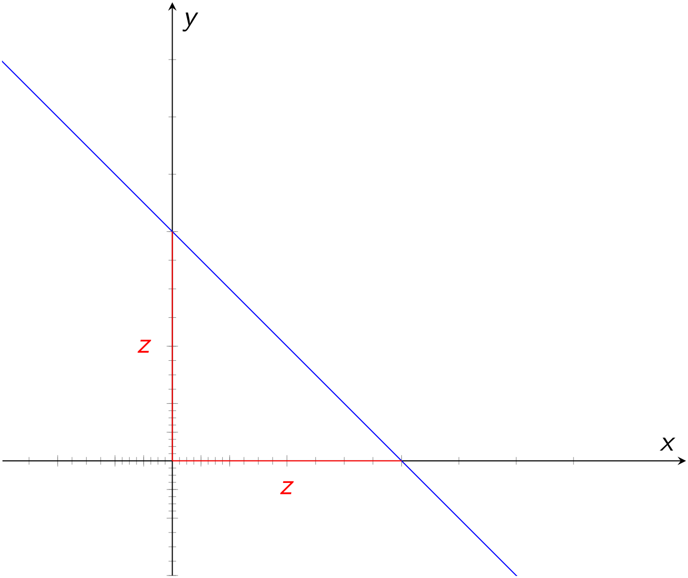
Difficulties

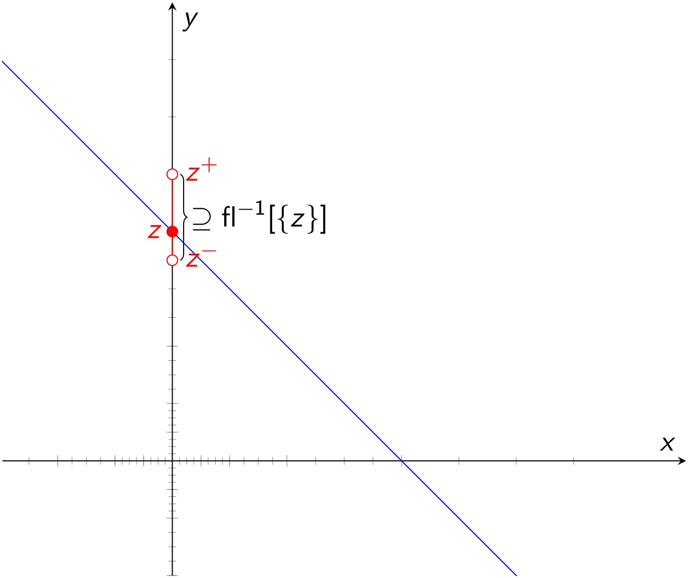
Unary rounded functions are easy, since the preimage of $\mathrm{fl} \circ f$ is just $f^{-1} \circ \mathrm{fl}^{-1}$. However, addition is binary.

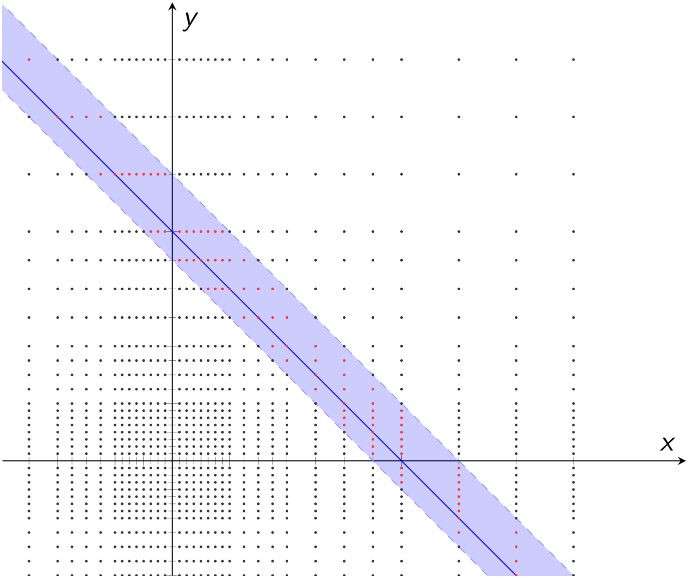
We can partially solve this by fixing one argument and taking the preimage, but that isn't guaranteed to give the optimal answer in one step unless the argument to the preimage is *feasible*.

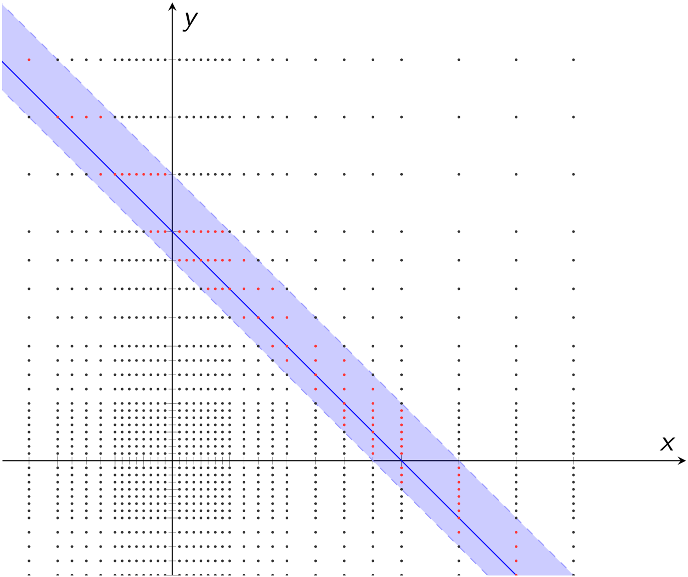
In pathological cases, it can take quadrillions of steps to arrive at the true answer!


y
x


y
z
x
z


y
x
z+
z
z−
⊇ fl−1[{z}]


y
x



Observation: candidate solutions all lie on parallel line *segments!*

Finding the longest line

For $\beta = 2$, B. Marre and C. Michel (2010) give exact extremal bounds on x and y such that $x \oplus y = z$.

These bounds are independent of the exponent range. Further, they are guaranteed to sum *exactly* to z.

Question: does this hold for $\beta > 2$?

Finding the longest line on a *grid*

When is the addition in $x \oplus y = z$ exact? That is, when do we have $x + y = z$?

Observation: the floating-point grid can be decomposed into overlapping (scaled) integer lattices.
Therefore, we are looking for the *lattice points* of $x + y = z$.

Lemma (Bézout's lemma)

$ax + by = c$ has integer solutions iff c is a multiple of $\gcd(a, b)$.

We can apply this by writing x, y and z as scaled integers:

$$M_x \beta^{q_x} + M_y \beta^{q_y} = M_z \beta^{q_z} \, .$$

Finding the longest line on the floating-point grid

Lemma

$M_x\beta^{q_x} + M_y\beta^{q_y} = M_z\beta^{q_z}$ has integer solutions iff $\min\{q_x, q_y\} \le q_z + k$ where k is the largest integer such that β^k divides M_z.

Proof.

1. Let $a = \beta^{q_x}$, $b = \beta^{q_y}$, $c = M_z\beta^{-k}\beta^{q_z+k}$.
2. By Bézout's lemma, $aM_x + bM_y = c$ is solvable iff $\gcd(a, b)$ divides c.
3. $M_z\beta^{-k}$ is not divisible by β, but $\gcd(a, b) = \beta^{\min\{q_x,q_y\}}$.
4. Therefore $\gcd(a, b)$ divides c iff $\min\{q_x, q_y\} \le q_z + k$. □

Since integral significands are bounded, there is a finite upper bound $U(z)$ on exact addition independent of exponent range!

Extremal bounds on inexact addition

We now have the upper bound $U(z)$ and lower bound $L(z) = z - U(z)$ for exact addition. But are there any floating-point numbers $x > U(z)$ or $y < L(z)$ such that $x \oplus y = z$ *inexactly?*

Extremal bounds on inexact addition

We now have the upper bound $U(z)$ and lower bound $L(z) = z - U(z)$ for exact addition. But are there any floating-point numbers $x > U(z)$ or $y < L(z)$ such that $x \oplus y = z$ *inexactly?*

No!

Even when $\beta > 2$, the extremal bounds for exact addition are also extremal for rounded addition. The quantum of $U(z)$ and $L(z)$ is simply too coarse-grained.

Solutions between $L(z)$ and $U(z)$

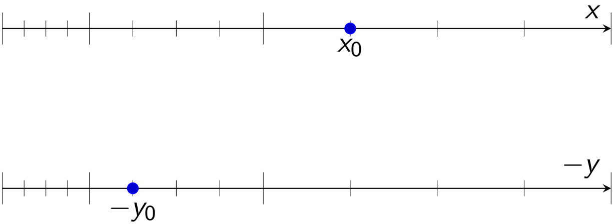
Suppose we have some x and y such that $x \oplus y = z$. Can we use them to find another nearby solution?

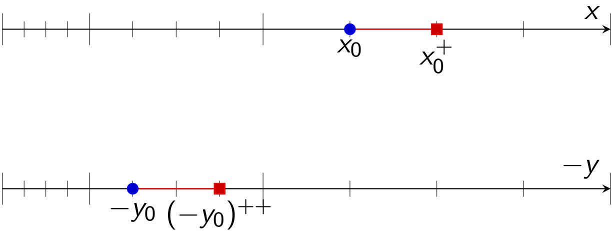
Solutions between $L(z)$ and $U(z)$

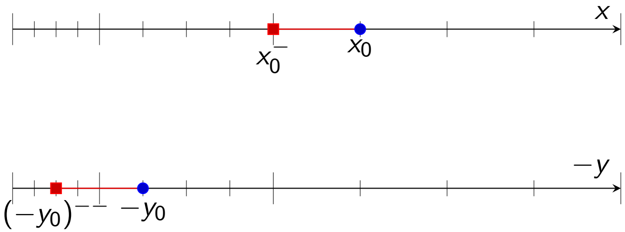
Suppose we have some x and y such that $x \oplus y = z$. Can we use them to find another nearby solution?

Maybe!

We can look at their neighbors to find x' and y' such that $x' + y' = x + y$. (Hint: all points with the same exact sum must be collinear.)


x
x_0
$-y$
$-y_0$


x
x_0
x_0^+
$-y$
$-y_0$
$(-y_0)^{++}$



x
x_0^-
x_0
$-y$
$(-y_0)^{--}$
$-y_0$

Exploiting collinearity

All solutions summing to the same exact value are collinear.
Therefore, all exact solutions are collinear.

$L(z)$ and $U(z)$ are the most extreme solutions, and they are exact.

Exploiting collinearity

All solutions summing to the same exact value are collinear.
Therefore, all exact solutions are collinear.

$L(z)$ and $U(z)$ are the most extreme solutions, and they are exact.

Lemma (Sterbenz)

If x and y are floating-point numbers with the same sign and $|y/2| \leq |x| \leq 2|y|$, then $x - y$ is exactly representable.

Exploiting collinearity

All solutions summing to the same exact value are collinear.
Therefore, all exact solutions are collinear.

$L(z)$ and $U(z)$ are the most extreme solutions, and they are exact.

Lemma (Sterbenz)

If x and y are floating-point numbers with the same sign and $|y/2| \leq |x| \leq 2|y|$, then $x - y$ is exactly representable.

Lemma

If x and y are floating point numbers with the same sign and

$$|y/2| \leq |x| \leq U(|y|)\,,$$

then $x - y$ is exactly representable.

Observation: if $x + y = z$, then we cannot have both $x < z/2$ and $y < z/2$.

Putting things together

With these results in hand, everything becomes relatively straightforward.

Theorem

If the intervals X and Y are within $[\min L[Z], \max U[Z]]$, the algorithm based on unary preimages converges in at most two steps.

Questions?