Optimal Bounds for Floating-Point Addition in Constant Time

Mak Andrlon*, Peter Schachte*, Harald Søndergaard* and Peter J. Stuckey†
*School of Computing and Information Systems
The University of Melbourne, Victoria 3010, Australia
†Faculty of Information Technology
Monash University, Caulfield East, Victoria 3145, Australia

Abstract—**Reasoning about floating-point numbers is notoriously difficult, owing to the lack of convenient algebraic properties such as associativity. This poses a substantial challenge for program analysis and verification tools which rely on precise floating-point constraint solving. Currently, interval methods in this domain often exhibit slow convergence even on simple examples. We present a new theorem supporting efficient computation of exact bounds of the intersection of a rectangle with the preimage of an interval under floating-point addition, in any radix or rounding mode. We thus give an efficient method of deducing optimal bounds on the components of an addition, solving the convergence problem.**

1. Introduction

Floating-point arithmetic has been the primary way of performing calculations with fractional values on modern computers for many years. Though its numerous corner cases and counter-intuitive behaviors have given it a reputation for fickleness [1], it is nowadays used even in safety-critical applications such as avionics control software [2]. Formal software verification systems are often used in such areas to verify critical safety requirements. Such systems typically rely on automatic theorem provers and constraint solvers to answer logical queries about the program being verified. However, our understanding of floating-point arithmetic is somewhat limited, so our ability to verify floating-point software is also limited.

The main difficulty in working with floating-point numbers is that they do not always obey the laws of real number arithmetic. Although they superficially appear to satisfy such properties, they are finite and thus *cannot* represent every real number. When the exact result of a numerical operation is not representable as a floating-point number, it must be rounded. As even basic operations can produce unrepresentable results, cases of unexpected behavior abound. For instance, it is always true over the reals that $x + 1 > x$. However, if interpreted as a formula of floating-point arithmetic, the inequality may not hold if x is so large that the gap between it and the next larger floating-point number is greater than 1. There are many other examples where real and floating-point arithmetic differ. Hence, directly applying decision procedures for real arithmetic to floating-point problems may lead to unsound conclusions.

Interval methods are frequently used in decision procedures for numerical problems. Floating-point arithmetic is no exception, and many different procedures based on interval reasoning have been proposed [3]–[6]. They over-approximate (as intervals) the sets of feasible values for the problem variables and use satisfiability-preserving operations to narrow the intervals, thus pruning the search space. The simplest such operation is directly computing and propagating the application of a function. For example, consider the ternary constraint $f(x, y) = z$ and bounding intervals X, Y and Z for x, y, z, resp. To reduce Z, we could compute $Z' = Z \cap f[X, Y]$ and then set $Z := \Box Z'$, where $f[X, Y]$ is the image of $X \times Y$ under f and $\Box Z'$ denotes the smallest interval containing Z' as a subset. However, computing $f[X, Y]$ exactly may be slow, and hence may need to over-approximate by finding $\Box f[X, Y]$ instead. Note that this is particularly simple if X and Y are closed and f is non-decreasing in both of its arguments, as it follows that

$$\begin{aligned}\Box f[X, Y] &= [\min f[X, Y], \max f[X, Y]] \\ &= [f(\min X, \min Y), f(\max X, \max Y)] \,.\end{aligned}$$

Despite being an over-approximation, this can still yield optimal results in some common cases. If $f[X, Y] \subseteq Z$, then $Z' = Z \cap f[X, Y] = f[X, Y]$ and thus $\Box Z' = \Box f[X, Y]$. Generally, $\Box Z' = \Box f[X, Y]$ if and only if both the minimum and maximum of $f[X, Y]$ are in Z. Otherwise, at least one endpoint is missing from Z and so there is at least one pair in $X \times Y$ which does not correspond to any element of Z. This may happen whenever there are multiple constraints on z, as it is possible that some other constraint has already reduced Z. We can eliminate the spurious values of X and Y by considering the preimage $f^{-1}[Z]$. We could find $W = (X \times Y) \cap f^{-1}[Z]$ and set $X := X'$ and $Y := Y'$ where $X' \times Y' = \Box W$ is the smallest rectangle containing W as a subset. However, unless f is invertible with a monotone inverse, finding the exact preimage may be difficult.

In this paper, we first study the computation of exact extremal bounds on the addends of a floating-point sum. These bounds are known for binary floating-point arithmetic [7], but we generalize the known results to arbitrary radices. However, what we are particularly interested in are exact bounds on subsets of the *preimage*. In the case of rounded invertible unary real functions, Michel [8] gives an optimal solution.

In the most common case of binary floating-point addition rounded to nearest with ties to even, the recent results of Gallois-Wong, Boldo and Cuoq [9] give an optimal solution. Otherwise, we can use the unary result to construct and over-approximation. But in many instances, this approximation can be very far from the optimal result, as we will see now.

Example 1. Consider IEEE 754 double-precision binary floating-point numbers. Let $\overline{\mathbb{F}}$ denote the set of floating-point numbers (including infinities), and let $\oplus$ denote floating-point addition. In this example, we assume that rounding is upward. Consider the floating-point addition constraint $x \oplus y = z$. Suppose the initial bounds on the floating-point variables x and y are $X_0 = Y_0 = [-100, 100] \cap \overline{\mathbb{F}}$, and assume that z is restricted to the singleton set $Z_0 = \{2^-\}$ where a^- is the greatest floating-point number strictly less than a. As $X_0 \oplus Y_0 \supseteq Z_0$, we cannot narrow the bounds on z using the current bounds on x and y. However, there are many values in X_0 and Y_0 which do not sum to 2^-. For instance, all values in $X_0 \oplus -100 = -100 \oplus Y_0$ are strictly less than 2^-, and thus -100 can be safely removed. The preimage bounds of Michel [8] give us a way of accelerating this process. To obtain a tighter lower bound on x, we find the least floating-point w such that $w \oplus \max X_0 \geq \min Z_0$. This is given by $w = \mathrm{RD}(2^- - 100)^+ = -98$ where $\mathrm{RD}(x)$ is the downward rounding of x and a^+ is the least floating-point number greater than a. The corresponding upper bound is given by $\mathrm{RD}(2^- + 100) = 102^-$ and thus we have

$$\begin{aligned} X_1 &= X_0 \cap [-98, 102^-] = [-98, 100] \cap \overline{\mathbb{F}}\,, \\ Y_1 &= Y_0 \cap [-98, 100^-] = [-98, 100^-] \cap \overline{\mathbb{F}}\,. \end{aligned}$$

After this point, however, convergence slows down drastically. If we perform the process again, we obtain

$$\begin{aligned} X_2 &= X_1 \cap [\mathrm{RD}(2^- - 100^-)^+, \mathrm{RD}(2^- + 98)] \\ &= [-(98^-), 100^-] \cap \overline{\mathbb{F}}\,, \\ Y_2 &= Y_1 \cap [\mathrm{RD}(2^- - 100^-)^+, \mathrm{RD}(2^- - 100^-)] \\ &= [-(98^-), 100^{--}] \cap \overline{\mathbb{F}}\,. \end{aligned}$$

From this point, every iteration will narrow the bounds on x and y by exactly one floating-point number on each side. The optimal solution in this case is $X_\infty = Y_\infty = [-(2^-), 4^-] \cap \overline{\mathbb{F}}$ and in fact $-(2^-) + 4^- = 2^-$ exactly. However, as there are quadrillions of floating-point numbers between -98 and -2; finding this solution will take an enormous amount of time.

With the results of this paper, we can calculate the tightest bounds on the addends using only a fixed small number of floating-point operations, regardless of choice of base or rounding mode. Thus, the main contributions of this paper are the following:

- A simpler proof of a result [7] concerning extremal bounds for binary floating-point addition.
- A generalization of this result to floating-point addition in arbitrary bases.
- A stronger variant of a well-known result of Sterbenz on exact floating-point subtractions [10].
- A set of results that allows for the constant-time computation of optimal bounds on the addends of a floating-point addition, given initial bounds on the addends and the sum.

In the next section we introduce necessary notation and preliminary results. In Section 3 we prove a generalized form of Marre and Michel's result [7] on universal preimage bounds for addition. In Section 4 we use the bounds to develop a stronger version of Sterbenz's lemma. In Section 5 we show how to perform optimal floating-point addition bounds tightening in constant time. We conclude in Section 6.

2. Preliminaries

We assume the reader has some familiarity with standard representations of floating-point numbers. Nevertheless, we now recall some basics, mainly to fix our notation.

A finite floating-point number is generally written $\pm d_1.d_2d_3 \cdots d_p \times \beta^e$, where each digit d_i is an integer in $[0, \beta - 1]$. The $\pm d_1.d_2d_3 \cdots d_p$ part is the *significand*, its length p is the *precision*, β is the *base* (commonly 2 or 10), and e is the *exponent*. For example, with precision $p = 4$ and base $\beta = 2$, we can represent the real number 0.40625 (here written in decimal) as 1.101×2^{-2}, assuming $e = -2$ is an allowed exponent, that is, one that can be represented given the number of bits allocated for this. Under the same parameters, the real number 0.1 (again in decimal) has no exact finite representation.

To avoid multiple representations for the same real number, floating-point numbers are usually *normalized*, that is, the digit d_1 before the radix point is precluded from being 0. Special patterns are used to represent non-standard floats: (positive and negative) zero, (positive and negative) infinity, and NaN ("not a number").

While the IEEE 754 standard assumes normalized floating-point numbers, it makes an important exception for numbers close to 0: when the exponent is the smallest allowed exponent ($e_{\min}$), the significand does not need to be normalized. This exception leads to a number of advantages; most importantly, it guarantees that for any finite x and y, $x \ominus y = 0$ if and only if $x = y$—one of the rare cases where a useful law carries over from the reals.

We now formalize these concepts. Following convention, we denote the integers by $\mathbb{Z}$, the reals by $\mathbb{R}$, and the (affinely) extended reals by $\overline{\mathbb{R}} = \mathbb{R} \cup \{+\infty, -\infty\}$. We denote the image of a set X under the function g by $g[X]$, and the preimage of X under g by $g^{-1}[X]$. To simplify exposition, we will denote the images of $X \times Y$ and $X \times \{y\}$ under a binary operator $\circ$ simply by $X \circ Y$ and $X \circ y$, respectively.

A *floating-point format* is a quadruple of integers $(\beta, p, e_{\min}, e_{\max})$ where $\beta \geq 2$ is the *base*, $p \geq 1$ is the *precision*, and $e_{\min} < e_{\max}$ are the minimum and maximum *exponents*, respectively. The set of finite nonzero floating-point numbers is defined as

$$\mathbb{F}^* = \left\{ M \cdot \beta^{e-p+1} \;\middle|\; \begin{array}{r} M, e \in \mathbb{Z}, 0 < |M| < \beta^p, \\ e_{\min} \leq e \leq e_{\max} \end{array} \right\}$$

We extend the set $\mathbb{F}^*$ as follows:

$$\mathbb{F} = \{0\} \cup \mathbb{F}^*, \qquad \overline{\mathbb{F}} = \mathbb{F} \cup \{-\infty, +\infty\}.$$

So $\overline{\mathbb{F}}$ is the set of all floating-point numbers (for a given floating-point format). For convenience, we also define the set of nonzero floating-point numbers with unbounded exponent:

$$\mathbb{F}^*_\infty = \{M \cdot \beta^q \mid M, q \in \mathbb{Z}, 0 < |M| < \beta^p\}.$$

Note that $\mathbb{F}^*_\infty$ is always a superset of $\mathbb{F}^*$ regardless of the choice of $e_{\min}$ and $e_{\max}$.

A finite nonzero floating-point number $f \in \mathbb{F}^*$ is *normal* iff $|f| \geq \beta^{e_{\min}}$. Otherwise, it is *subnormal*. Zero is neither normal nor subnormal (we have no need to consider signed zero in this paper).

The *exponent* of f is defined as $e = \lfloor \log_\beta |f| \rfloor$ if f is normal, or $e = e_{\min}$ otherwise. Equivalently, $e = \max\{\lfloor \log_\beta |f| \rfloor, e_{\min}\}$. The *quantum exponent* of f is $q = e - p + 1$. The (quantum) exponent of zero is undefined. We denote the minimum and maximum quantum exponents by $q_{\min} = e_{\min} - p + 1$ and $q_{\max} = e_{\max} - p + 1$, respectively.

The *significand* of f is given by $m = f/\beta^e$, and the *integral significand* of f is $M = m\beta^{p-1}$. Note that M is an integer and that $f = m\beta^e = M\beta^q$. The (integral) significand of zero is zero.

The *predecessor* x^- and *successor* x^+ of a floating-point number x are defined as the greatest floating-point number less than x and the least floating-point number greater than x, respectively. Note that the quantum exponent q satisfies $|x|^+ = |x| + \beta^q$ iff $0 < |x| < \max \mathbb{F}$.

We denote by $\mathrm{fl} : \overline{\mathbb{R}} \to \overline{\mathbb{F}}$ an arbitrary nondecreasing, *faithful* rounding function. A rounding function is faithful if and only if it is equal to the identity function for all floating-point numbers and otherwise returns one of the two floating-point values on either side of its input. Note that a faithful rounding function is therefore idempotent and surjective. The five standard IEEE rounding modes are all nondecreasing and faithful. A real number x can be rounded up (to $\mathrm{RU}(x)$), down (to $\mathrm{RD}(x)$), away from zero, to the nearest floating-point with ties to even ($\mathrm{RNE}(x)$), or to the nearest with ties away from zero. That is, $\mathrm{RU}(x)$ is the least floating-point number no less than x, and $\mathrm{RD}(x)$ is the greatest floating-point number no greater than x. Note that for any x, $\mathrm{RD}(x) \leq x$ and $\mathrm{RU}(x) \geq x$, with equality attained if and only if x is a floating-point number.

The rounded addition operator $\oplus$ is defined whenever x and y are not infinities of opposite sign as $x \oplus y = \mathrm{fl}(x + y)$. Similarly, the rounded subtraction operator $\ominus$ is defined whenever x and y are not infinities of like sign as $x \ominus y = \mathrm{fl}(x - y)$.

For the sake of completeness, we will now list some elementary properties of floating-point arithmetic that will (silently) be used throughout the paper.

Lemma 1. Let $x, y \in \mathbb{F}^*$, and let e_x and e_y be the exponents of x and y, respectively. If $|x| \geq |y|$, then $e_x \geq e_y$.

Lemma 2. Let $x \in \mathbb{R}$. If $\mathrm{fl}(x)$ is finite, then $|\mathrm{fl}(x) - x| < \beta^q$ where q is the quantum exponent of $\mathrm{fl}(x)$ or $q_{\min}$ if $\mathrm{fl}(x) = 0$.

Lemma 3. Let $x, y \in \mathbb{F}^*$ with quantum exponents q_x and q_y, respectively. Let M_y be the integral significand of y. If $|x| > |y|$ and $|M_y| = \beta^p - 1$, then $q_x > q_y$.

Lemma 4. Let $x \in \mathbb{F}^*_\infty$. If $\beta^{e_{\min}} \leq |x| \leq \max \mathbb{F}$, then $x \in \mathbb{F}$.

Lemma 5. Let $x, y \in \mathbb{F}$. If x and y are subnormal, then $x \pm y \in \mathbb{F}$.

3. Universal Bounds on Addends

We now establish the floating-point addition property of Marre and Michel [7], and generalize it to arbitrary radices. Given a fixed nonzero floating-point number z, the property gives the minimum and maximum values of x and y that satisfy the equation $x \oplus y = z$ in binary floating-point arithmetic. As we will see, these bounds are in fact guaranteed to sum exactly to z, immediately giving optimal solutions in cases where the initial bounds are too wide, as in our earlier example. The proof of our generalization is elementary and requires only basic number theory.

To begin, we will need some preliminary results. The following lemma gives a necessary condition for the sum of two floating-point numbers to be exact.

Lemma 6. Let $q_x, q_y, q_z \in \mathbb{Z}$ and let $M_z \in \mathbb{Z}$. Then there exist integers x and y such that $\beta^{q_x} x + \beta^{q_y} y = M_z \beta^{q_z}$ if and only if $\min\{q_x, q_y\} \leq q_z + k$ where k is the greatest integer such that β^k divides M_z.

Proof: Let $n = \min\{q_x, q_y, q_z\}$ and let $a = \beta^{q_x - n}$, $b = \beta^{q_y - n}$ and $c = M_z \beta^{q_z - n}$. As the exponents are nonnegative, a, b and c are integers. Thus by Bézout's lemma, $ax + by = c$ has a solution in $x, y \in \mathbb{Z}$ if and only if c is a multiple of $\gcd(a, b)$. Hence satisfying x and y exist if and only if $M_z \beta^{q_z - n} = M_z \beta^{-k} \cdot \beta^{q_z + k - n}$ is divisible by $\gcd(a, b) = \beta^{\min\{q_x, q_y\} - n}$. Since $M_z \beta^{-k}$ is not divisible by β, this is equivalent to $\min\{q_x, q_y\} \leq q_z + k$. □

From Lemma 6 it follows that, for every nonzero floating-point z, there is an upper bound (independent of $e_{\min}$ and $e_{\max}$) on the magnitude of floating-point values that can sum to exactly to z. This leads us to the following definition:

Definition 1. The universal upper and lower bounds on exact floating-point addition $U, L : \mathbb{F}^*_\infty \to \mathbb{F}^*_\infty$ are defined by

$$U(z) = \max\{x \in \mathbb{F}^*_\infty \mid \exists y \in \mathbb{F}^*_\infty (x + y = z)\},$$
$$L(z) = z - U(z).$$

Note that these bounds extend naturally to subtraction, as we have $x \ominus y = x \oplus -y$, and $x + y = -z$ if and only if $(-x) + (-y) = z$. Hence they are duals: $L(-z) = -U(z)$ and $U(-z) = -L(z)$. We now show that the universal bounds on exact addition are also the universal bounds on rounded addition.

Lemma 7. Let $z \in \mathbb{F}^*$ and let $u, v \in \mathbb{F}$. If $u \oplus v = z$, then $u, v \in [L(z), U(z)]$.

Proof: Let $x = L(z)$ and $y = U(z)$. As $-z \in \mathbb{F}^*$ and $z + (-z) = 0 \in \mathbb{F}$, we have $x \leq -|z| < 0$ and hence $-x = |x| \geq |z|$ and $y = |y| \geq |z| > 0$. Suppose to the contrary that $u \oplus v = z$ but u and v are not both in $[x, y]$. Then the absolute rounding error $\varepsilon = |u+v-z| = |u-x+v-y|$ must be strictly less than β^{q_z} where q_z is the quantum exponent of z. Without loss of generality, we will assume that $u \leq v$. Suppose to the contrary that $u < x$. Then $|u| > |x|$ and hence there is some integer $N_u < 0$ such that $u - x = N_u \beta^{q_x}$ where q_x is the quantum exponent of x. As $|x| \geq |z|$, the exponent of x is no less than that of z, and so $u - x \leq -\beta^{q_x} \leq -\beta^{q_z} < -\varepsilon$. Hence v must be greater than y, and so there is some integer $N_v > 0$ such that $v - y = N_v \beta^{q_y}$ where q_y is the quantum exponent of y. Since $|y| \geq |z|$ and therefore $q_y \geq q_z$, we have $\varepsilon = |K|\beta^{q_z}$ where $K = N_u \beta^{q_x - q_z} + N_v \beta^{q_y - q_z}$ is an integer. As $\varepsilon < \beta^{q_z}$, it must be that $K = 0$ and thus $\varepsilon = 0$. However, x is the least value for which the addition can be exact and u is strictly less than x, so this is a contradiction. Suppose instead that $u \geq x$. Since $u \leq v$ and at least one of u or v must be outside of $[x, y]$, we must have $v > y$. Hence there is some integer $N_v > 0$ such that $v - y = N_v \beta^{q_y}$. However, $u - x \geq 0$ and thus $\varepsilon = u - x + v - y \geq v - y \geq \beta^{q_z}$, which is a contradiction. Thus u and v must be in $[x, y]$. □

With these results in hand, we are very nearly done. However, though have shown that $L(z)$ and $U(z)$ give universal bounds on rounded addition, we still do not have an efficient construction. We now use Lemma 6 to obtain a more concise proof of the property of Marre and Michel [7], which gives an easy way of calculating $L(z)$ and $U(z)$ when $\beta = 2$. We will then generalize their result to higher bases. In the proof below, we will retain the use of the symbol β in order to highlight its most essential aspects.

Theorem 1 (Marre and Michel [7]). Let $z \in \mathbb{F}^*$ with integral significand $M_z > 0$ and quantum exponent q_z, and let k be the greatest integer such that β^k divides M_z. Let $x = -(\beta^p - 1) \cdot \beta^{q_x}$ where $q_x = q_z + k$ and let $y = z - x$. If $\beta = 2$, $L(z) = x$ and $U(z) = y$.

Proof: Note that $z - x = M_z \beta^{q_x - k} + (\beta^p - 1) \cdot \beta^{q_x} = (M_z \beta^{-k} + \beta^p - 1)\beta^{q_x}$. Suppose $\beta = 2$. Since $M_z \beta^{-k}$ and $\beta^p - 1$ are both positive, odd, and less than β^p, their sum is a positive multiple of β less than $2\beta^p$. Thus $M_y = \beta^{-1}(M_z \beta^{-k} + \beta^p - 1)$ is an integral significand and so $y = z - x \in \mathbb{F}^*_\infty$. Hence $L(z) \leq x$. Suppose to the contrary that $L(z) < x$. Then there is some $x' \in \mathbb{F}^*_\infty$ less than x such that $y' = z - x' \in \mathbb{F}^*_\infty$. Since $|M_x|$ is maximal, the quantum exponent $q_{x'}$ of x' must be greater than q_x. However, by Lemma 6, if $x' + y' = z$, then $\min\{q_{x'}, q_{y'}\} = q_{x'} \leq q_z + k = q_x$ where $q_{y'}$ is the quantum exponent of y', which is a contradiction. Therefore $L(z) = x$ and hence $U(z) = z - x$ by definition. □

The key insight from the previous proof is that we need an integral significand M_x such that $M_x + M_z \beta^{-k}$ is a multiple of β. Equivalently, the last digit of $M_x + M_z \beta^{-k}$ in base β is 0. For $\beta = 2$, it suffices that M_x is odd, as $M_z \beta^{-k}$ must itself be odd. We now reuse Lemma 6 to generalize Theorem 1 to arbitrary bases.

Theorem 2. Let $z \in \mathbb{F}^*$ with integral significand $M_z > 0$ and quantum exponent q_z, and let k be the greatest integer such that β^k divides M_z. Let r be the remainder of the division of $M_z \beta^{-k}$ by β. Let $x = -(\beta^p - r) \cdot \beta^{q_x}$ where $q_x = q_z + k$ and let $y = z - x$. Then $L(z) = x$ and $U(z) = y$.

Proof: We will first show that the subtraction $z - x$ is exactly representable in $\mathbb{F}^*_\infty$:

$$\begin{aligned} z - x &= M_z \beta^{q_z} + (\beta^p - r) \cdot \beta^{q_x} \\ &= M_z \beta^{-k} \beta^{q_z + k} + (\beta^p - r) \cdot \beta^{q_z + k} \\ &= (M_z \beta^{-k} - r + \beta^p) \cdot \beta^{q_z + k} \\ &= ((M_z \beta^{-k} - r)\beta^{-1} + \beta^{p-1}) \cdot \beta^{q_z + k + 1} . \end{aligned}$$

Since $M_z \beta^{-k} - r$ is nonnegative, divisible by β, and strictly less than β^p, the first factor is a positive integer strictly less than β^p and therefore an integral significand. As $q_z + k + 1$ is an integer, $z - x$ is hence exactly representable in $\mathbb{F}^*_\infty$, and therefore $L(z) \leq x$.

Suppose to the contrary that $L(z) < x$. Then there is some $w \in \mathbb{F}^*_\infty$ less than x such that $y' = z - w \in \mathbb{F}^*_\infty$. Let M_w and q_w be the integral significand and quantum exponent of w, respectively. As $w < x$, we have $q_w \geq q_x$, and either $M_w < -(\beta^p - r)$ or $q_w > q_x$. By Lemma 6, $\min\{q_w, q_{y'}\} = q_w \leq q_z + k = q_x$ where $q_{y'}$ is the quantum exponent of y'. Hence $q_w = q_x$. Therefore there must be some integer d such that $M_w = -(\beta^p - d)$ where $1 \leq d < r$. As $z - w = (M_z \beta^{-k} - M_w) \cdot \beta^{q_z + k}$ and $M_z \beta^{-k} - M_w$ is greater than β^p, the result is exactly representable only if $M_z \beta^{-k} - M_w = M_z \beta^{-k} - d + \beta^p$ is divisible by β. Since $M_z \beta^{-k} - r$ is a multiple of β and $1 \leq d < r < \beta$, that is impossible. Thus $L(z) = x$ and hence $U(z) = z - x = y$. □

4. Solutions of Addition Inside the Bounds

Although Lemma 7 gives a useful necessary condition for $x \oplus y = z$ to hold, it says nothing about what solutions exist between $L(z)$ and $U(z)$. It is well-known that if two floating-point values are within a factor of two of each other, their difference is exactly representable:

Lemma 8 (Sterbenz [10]). Let $x, y \in \mathbb{F}$. If x and y have the same sign and

$$\frac{|y|}{2} \leq |x| \leq 2|y| ,$$

then $x - y \in \mathbb{F}$.

However, there are many other cases where subtractions are exact. For example, consider $U(x) - \delta$ and $L(x) + \delta$ with $\delta > 0$. If x is positive, the exponent of $U(x)$ is no less than that of $L(x)$. Intuitively, if we reduce the upper bound by a single "step" to the next highest value—its predecessor—we can perfectly mirror that movement on the lower bound by *increasing* it by β^n steps, where n is the difference between their exponents. This changes the bounds by the same amount in opposite directions, so they still sum exactly to x. As we can repeat this process until we completely reduce the lower bound to zero, the upper bound attains every floating-point value between x and $U(x)$. We formalize this intuition in the following lemma.

Lemma 9. Let $x, y \in \mathbb{F}^*$. If x and y have the same sign and $|y| < |x| \leq U(|y|)$, then $x - y \in \mathbb{F}^*$.

Proof: Suppose x and y have the same sign. Let M_x, M_y be the integral significands and q_x, q_y be the quantum exponents of x and y respectively. Then,

$$\begin{aligned} x - y &= M_x\beta^{q_x} - M_y\beta^{q_y} \\ &= \beta^{q_x}(M_x - M_y\beta^{q_y - q_x}) \,. \end{aligned}$$

Let k be the greatest integer such that β^k divides M_y and suppose $q_x \leq q_y + k$. Then $M_y\beta^{q_y-q_x}$ is an integer. As x and y have the same sign, so do M_x and M_y. Therefore, as $q_y \leq q_x$ and $|M_y| \leq \beta^p - 1$, it follows that $M_x - M_y\beta^{q_y-q_x}$ is an integral significand. Since $q_{\min} \leq q_x \leq q_{\max}$, we have $x - y \in \mathbb{F}^*$. Suppose $q_x > q_y + k$ instead. Then by Theorem 2, $|x| \leq U(|y|)$ implies that $q_x = q_y + k + 1$. Since x and y have the same sign,

$$\begin{aligned} |x - y| &= |x| - |y| \\ &\leq U(|y|) - |y| = -L(|y|) \\ &\leq (\beta^p - 1)\beta^{q_y+k} \,. \end{aligned}$$

Dividing by β^{q_y+k}, we obtain $|M_x\beta - M_y\beta^{-k}| \leq \beta^p - 1$ and hence $M_x\beta - M_y\beta^{-k}$ is an integral significand. As $q_{\min} \leq q_y \leq q_y + k < q_x \leq q_{\max}$, $x - y \in \mathbb{F}^*$. □

Combining the previous result and Sterbenz's lemma, we immediately obtain the following theorem.

Theorem 3. Let $x, y \in \mathbb{F}$. If x and y have the same sign and

$$\frac{|y|}{2} \leq |x| \leq U(|y|) \,,$$

then $x - y \in \mathbb{F}$.

The above result is, in essence, a substantially stronger version of Sterbenz's lemma, as it provides a precise upper bound on x for subtraction to be exact.

Note that $x + y = z$ requires that one of x or y is no less than $z/2$ and the other no greater than $z/2$, and therefore at least one satisfies Theorem 3. We therefore obtain a complete characterization of the exact solutions of $x + y = z$ in x and y over the floating-point numbers: if $z = 0$, then $x + y = z$ if and only if $x = -y$; if $z > 0$, then $x + y = z$ if and only if $z/2 \leq \max\{x, y\} \leq U(z)$ and $\min\{x, y\} = z \ominus \max\{x, y\}$; if $z < 0$, then $x + y = z$ if and only if $L(z) \leq \min\{x, y\} \leq z/2$ and $\max\{x, y\} = z \ominus \min\{x, y\}$.

4.1. Inexact Solutions

Although we now have a precise characterization of the set of solutions to the exact addition, what we are ultimately interested in are the solutions of $x \oplus y = z$. Let $R = \mathrm{fl}^{-1}[\{z\}] - z$. Then $x \oplus y = z$ if and only if $x + y = z + r$ for some $r \in R$. As R is necessarily an interval and $z + R = \mathrm{fl}^{-1}[\{z\}] \subseteq (z^-, z^+)$, we see that every inexact solution pair corresponds to a floating-point interval containing at least one exact solution. However, the solution space is not guaranteed to be connected unless the gap between exact solutions is strictly narrower than R. Otherwise, there may be holes in the solution space if the spacing between points is too wide or when R is an open interval (e.g. if the rounding mode is RNE and the integral significand of z is odd).

5. The Interval Case

Having developed the previous results, we now proceed to solve our initial problem: finding optimal interval bounds on the components of a floating-point addition using a constant number of operations. We will do this by showing that applying the bounds of Lemma 7 and Theorem 2 guarantees that bounds from the unary preimage will converge to the optimal bounds in no more than two iterations.

In the following, we consider a nonempty floating-point interval $Z \subseteq \mathbb{F}$, and arbitrary $l, u \in \mathbb{F}$ between $\min L[Z]$ and $\max U[Z]$ inclusive. Note that, for simplicity's sake, we only consider finite floating-point numbers. Since $x \oplus \pm\infty \in Z$ if and only if $\pm\infty \in Z$, infinities among the addends themselves are easily handled.

We define the criteria for optimality as follows:

Definition 2. We say that $x \in \mathbb{F}$ is *feasible* if and only if there is some $y \in \mathbb{F}$ such that $x \oplus y \in Z$. Otherwise, we say that x is *infeasible*. The pair $(x, y) \in \mathbb{F}^2$ is *satisfying* if and only if $x \oplus y \in Z$ where $x \geq l$ and $y \leq u$. The *optimal lower bound*, denoted $\Lambda_Z(l, u)$, is the least $x \in \mathbb{F}$ such that (x, y) is satisfying for some $y \in \mathbb{F}$, or $+\infty$ if none exists. Similarly, the *optimal upper bound*, denoted $\Upsilon_Z(l, u)$, is the greatest $y \in \mathbb{F}$ such that (x, y) is satisfying for some $x \in \mathbb{F}$, or $-\infty$ if none exists.

Although these functions are parametric in Z, we will henceforth simply refer to them as Λ and Υ as we do not vary the parameter. It is easy to show that, although in our original problem we are concerned with a problem involving two pairs of bounds (one for each variable), we need only consider two bounds at a time:

Lemma 10. Let X and Y be intervals over $\mathbb{F}$. Let

$$\begin{aligned} X' &= \{x \in X \mid \exists y \in Y\,(x \oplus y \in Z)\} \,, \\ Y' &= \{y \in Y \mid \exists x \in X\,(x \oplus y \in Z)\} \,. \end{aligned}$$

Then,

$$\begin{aligned} \square X' &= [\Lambda(\min X, \max Y), \Upsilon(\min Y, \max X)] \,, \\ \square Y' &= [\Lambda(\min Y, \max X), \Upsilon(\min X, \max Y)] \,. \end{aligned}$$

In the remainder of this section, we will show how to calculate Λ and Υ in constant time.

The following functions give lower and upper bounds on the set of feasible floating-point numbers. They are defined in nearly the same manner as the unary inverses of Michel [8] and will form the base of our approach.

Definition 3. Define $\Phi_Z : \mathbb{F} \to \mathbb{F} \cup \{+\infty\}$ and $\Psi_Z : \mathbb{F} \to \mathbb{F} \cup \{-\infty\}$ by

$$\begin{aligned} \Phi_Z(x) &= \min \mathrm{RU}[\mathrm{fl}^{-1}[Z] - x] \,, \\ \Psi_Z(x) &= \max \mathrm{RD}[\mathrm{fl}^{-1}[Z] - x] \,. \end{aligned}$$

As with Λ and Υ, we do not vary the parameter and so we will henceforth refer to these functions simply as Φ and Ψ. Note that, since $-\infty < \mathrm{RU}(x)$ and $\mathrm{RD}(x) < +\infty$ for finite x, we have $-\infty < \Phi(x)$ and $\Psi(x) < +\infty$.

Example 2. Suppose the rounding mode is RD. Then $\mathrm{fl} = \mathrm{RD}$, and hence $\mathrm{fl}^{-1}[\{z\}] = [z, z^+)$. Thus $\mathrm{fl}^{-1}[Z] = [\min Z, (\max Z)^+)$. As $\mathrm{RD}^{-1}[Z]$ is closed on the left,

$$\begin{aligned}\Phi(x) &= \min \mathrm{RU}[\mathrm{RD}^{-1}[Z] - x] \\ &= \mathrm{RU}(\min [\min Z, (\max Z)^+) - x) \\ &= \mathrm{RU}(\min Z - x)\,.\end{aligned}$$

As $\mathrm{RD}^{-1}[Z]$ is open on the right, we have to take some additional care with Ψ, as we cannot take the maximum of a right-open interval:

$$\begin{aligned}\Psi(x) &= \max \mathrm{RD}[\mathrm{RD}^{-1}[Z] - x] \\ &= \max \mathrm{RD}[[\min Z, (\max Z)^+) - x] \\ &= \mathrm{RU}((\max Z)^+ - x)^-\,.\end{aligned}$$

We now show that Φ and Ψ satisfy some expected relations. First, that Φ and Ψ are indeed lower and upper bounds, respectively.

Lemma 11. Let $x, y \in \mathbb{F}$. If $x \oplus y \in Z$, then $x \geq \Phi(y)$ and $y \leq \Psi(x)$.

Proof: Suppose $x \oplus y \in Z$. Then $x + y = t$ for some $t \in \mathrm{fl}^{-1}[Z]$. Therefore,

$$\mathrm{RU}(x) = \mathrm{RU}(t - y) \geq \min \mathrm{RU}[\mathrm{fl}^{-1}[Z] - y] = \Phi(y)$$

and similarly $\mathrm{RD}(y) = \mathrm{RD}(t - x) \leq \Psi(x)$. As $x, y \in \mathbb{F}$, we have $\mathrm{RU}(x) = x$ and $\mathrm{RD}(y) = y$, and hence the result. □

We now show that Φ and Ψ are nonincreasing functions.

Lemma 12. Let $x, y \in \mathbb{F}$. If $x \leq y$, then $\Phi(x) \geq \Phi(y)$ and $\Psi(x) \geq \Psi(y)$.

Proof: Suppose $x \leq y$ and let $t \in \mathrm{fl}^{-1}[Z]$. Then $t - x \geq t - y$. By the monotonicity of rounding, $\mathrm{RU}(t-x) \geq \mathrm{RU}(t-y)$ and $\mathrm{RD}(t-x) \geq \mathrm{RD}(t-y)$. If we choose t to minimize $\mathrm{RU}(t - x)$, we obtain $\mathrm{RU}(t-x) = \min \mathrm{RU}[\mathrm{fl}^{-1}[Z] - x] = \Phi(x)$ and $\mathrm{RU}(t - y) \geq \min \mathrm{RU}[\mathrm{fl}^{-1}[Z] - y] = \Phi(y)$ and therefore $\Phi(x) \geq \Phi(y)$. Similarly, if we instead choose t to maximize $\mathrm{RD}(t - y)$, we obtain $\Psi(x) \geq \Psi(y)$. □

We now show that if the lower bound from Φ is infinite, then there are no finite solutions, and therefore, no finite optimal lower bound exists. The proof for the case of the upper bound from Ψ is the same, mutatis mutandis.

Lemma 13. Let $x, y \in \mathbb{F}$. If $x \oplus y \in Z$ and $y \leq u$, then $\Phi(u)$ is finite.

Proof: Suppose $x \oplus y \in Z$ and $y \leq u$. Then by Lemmas 11 and 12, $\Phi(u) \leq \Phi(y) \leq x < +\infty$. Since $\Phi(u) \neq -\infty$ by definition, we obtain the result. □

Lemma 14. Let $x, y \in \mathbb{F}$. If $x \oplus y \in Z$ and $x \geq l$, then $\Psi(l)$ is finite.

The following lemma gives a pair of necessary conditions for the bounds to be feasible. As before, the proof of the subsequent lemma is almost identical and hence omitted.

Lemma 15. Let $x, y \in \mathbb{F}$. If $x \oplus y \in Z$ where $y \leq u$, then $x \geq \Phi(u)$ and $y \leq \Psi(\Phi(u))$.

Proof: Suppose $x \oplus y \in Z$ and $y \leq u$. Then by Lemmas 11 and 12, $x \geq \Phi(y) \geq \Phi(u)$. Applying Lemmas 11 and 12 again, we obtain $y \leq \Psi(x) \leq \Psi(\Phi(u))$. □

Lemma 16. Let $x, y \in \mathbb{F}$. If $x \oplus y \in Z$ where $x \geq l$, then $x \geq \Phi(\Psi(l))$ and $y \leq \Psi(l)$.

An immediate corollary of the previous two lemmas is that Φ never overestimates the optimal lower bound, and Ψ never underestimates the optimal upper bound.

The properties we have established here will be used frequently throughout the remainder of this section. We will now show how to find solutions to $x \oplus y \in Z$ such that $x \geq l$ and $y \leq u$. This will be divided into two parts: the easy case of either l or u being feasible, and the hard case of them being infeasible.

5.1. Feasible Bounds

If our initial lower and upper bounds l and u are feasible, then finding the optimal bounds happens to be very simple. If $l \oplus u$ lands below $\min Z$, then if u is feasible, $\Phi(u)$ is the least lower bound which sums with u to something in Z. Similarly, if $l \oplus u > \max Z$, then if l is feasible, then $\Psi(l)$ is the greatest value summing with l to any value in Z.

We first prove that Φ never underestimates the optimal lower bound for a fixed addition.

Lemma 17. For all $x \in \mathbb{F}$, $\Phi(x) \oplus x \geq \min Z$.

Proof: Let $x \in \mathbb{F}$, and let $t \in \mathrm{fl}^{-1}[Z]$ such that $\Phi(x) = \mathrm{RU}(t - x)$. Then $\Phi(x) + x = \mathrm{RU}(t - x) + x \geq t - x + x = t$ and therefore $\Phi(x) \oplus x \geq \mathrm{fl}(t) \in Z$. Since $\mathrm{fl}(t) \geq \min Z$, the result follows. □

We now show that $\Phi(x)$ "overshoots" Z if and only if x is infeasible.

Lemma 18. For all $x \in \mathbb{F}$, $\Phi(x) \oplus x \leq \max Z$ if and only if x is feasible.

Proof: Let $x \in \mathbb{F}$. By Lemma 17, $\min Z \leq \Phi(x) \oplus x$. Thus, as Z is an interval, if $\Phi(x) \oplus x \leq \max Z$, then $\Phi(x) \oplus x \in Z$ and so x is feasible.

Suppose x is feasible. Then there is some $y \in \mathbb{F}$ such that $x \oplus y \in Z$. By Lemma 11, $\Phi(x) \leq y$, and therefore $\Phi(x) \oplus x \leq y \oplus x \leq \max Z$ by monotonicity. □

Putting together the two lemmas above, we immediately obtain the following result:

Lemma 19. Let $x \in \mathbb{F}$. If x is feasible, then $\Phi(x) \oplus x \in Z$.

We now see that a feasible upper bound gives an optimal lower bound:

Lemma 20. If $l \oplus u < \min Z$ and u is feasible, then $\Phi(u)$ is the optimal lower bound and u is the optimal upper bound.

Proof: Suppose $l \oplus u < \min Z$ and u is feasible. Then by Lemma 19, $\Phi(u) \oplus u \in Z$ and hence $\Phi(u) > l$ by monotonicity. Therefore u is the optimal upper bound, and by Lemma 15, $\Phi(u)$ is the optimal lower bound. □

The corresponding lemmas for Ψ and the optimal upper bound can be proved in the same way as Lemmas 17 to 20:

Lemma 21. For all $x \in \mathbb{F}$, $x \oplus \Psi(x) \leq \max Z$.

Lemma 22. For all $x \in \mathbb{F}$, $x \oplus \Psi(x) \geq \min Z$ if and only if x is feasible.

Lemma 23. Let $x \in \mathbb{F}$. If x is feasible, then $x \oplus \Psi(x) \in Z$.

Lemma 24. If $l \oplus u > \max Z$ and l is feasible, then l is the optimal lower, and $\Psi(l)$ the optimal upper, bound.

The easy case solved, we can proceed to the more difficult scenario of our suboptimal bound of interest being infeasible.

5.2. Infeasible Bounds

To find an answer when the bound is infeasible, we will first need some auxiliary technical results. We shall begin by eliminating the possibility of Z containing zero in this case.

Lemma 25. If $0 \in Z$, then all $x \in \mathbb{F}$ are feasible.

Proof: For any $x \in \mathbb{F}$, we have $-x \in \mathbb{F}$. Thus if $0 \in Z$, we then have $x \oplus -x = 0 \in Z$. □

As Z is an interval by assumption, a further consequence of the above lemma is that all members of Z have the same sign. Therefore, we can use what we know about the structure of the solution space from Section 4 and Theorem 3 to eliminate many possibilities.

Lemma 26. Let $x \in \mathbb{F}$. If there is a $z \in Z$ with the same sign as x such that $|z|/2 \leq |x| \leq U(|z|)$, then x is feasible.

Proof: Suppose there is such a z. Then by Theorem 3, $z - x \in \mathbb{F}$. As $x \oplus (z - x) = z \in Z$, the result follows. □

We now show that the union of the intervals where Theorem 3 applies is itself an interval.

Lemma 27. Let I be an interval over $\mathbb{F}$ such that $x > 0$ for all $x \in I$. Then $\bigcup_{x \in I} [x/2, U(x)] = [\min I/2, \max U[I]]$.

Proof: Let $x \in I$. As $x/2 \leq x^- < x < U(x)$, $[x/2, U(x)]$ contains x^- and also x^+ if it is finite, and thus overlaps with the intervals surrounding them, if any. Hence $\bigcup_{x \in I} [x/2, U(x)]$ is a union of overlapping intervals and thus itself an interval. And the least value $x/2$ attains is $\min I/2$, while the greatest value $U(x)$ attains is $\max U[I]$. □

The negative case easily follows from Lemma 27:

Lemma 28. Let I be an interval over $\mathbb{F}$ such that $x < 0$ for all $x \in I$. Then $\bigcup_{x \in I} [L(x), x/2] = [\min L[I], \max I/2]$.

With the above lemmas, we significantly shrink the space of infeasible values:

Lemma 29. Let $x \in \mathbb{F}$ and suppose all $z \in Z$ have the same sign as x. Let $Z' = \{|z| \mid z \in Z\}$. If $\min Z'/2 \leq |x| \leq \max U[Z']$, then x is feasible.

The following result is perhaps unsurprising, but it is vital to the proof of the lemma ahead.

Lemma 30. For any $x \in \mathbb{F}$, $\mathrm{RD}(x/2) + \mathrm{RU}(x/2) = x$.

Proof: Let $x \in \mathbb{F}$, and let M and q be its integral significand and quantum exponent, respectively. Then $\mathrm{RD}(x/2) = \lfloor M' \rfloor \beta^{q'}$ with some quantum exponent q' and integral significand $M' = \frac{1}{2} M \beta^{q-q'}$ where $q' \leq q$. If M' is an integer, then $x/2$ is exactly representable and hence the result follows. Suppose M' is a half-integer instead. Then $\mathrm{RU}(x/2) = \mathrm{RD}(x/2)^+$ and $\lfloor M' \rfloor = M' - \frac{1}{2}$. Without loss of generality, we will assume x is positive. Then $\mathrm{RD}(x/2)^+ = (\lfloor M' \rfloor + 1)\beta^{q'} = (M' + \frac{1}{2})\beta^{q'}$ and hence $\mathrm{RD}(x/2) + \mathrm{RU}(x/2) = 2M'\beta^{q'} = M\beta^q = x$. □

We now proceed to the main lemma of this subsection. Given infeasible u, the next lemma provides an optimal lower bound if and only if it exists.

Lemma 31. If u is infeasible and $\Phi(u)$ is finite, then $\Phi(u)$ is feasible.

Proof: Suppose u is infeasible and $\Phi(u)$ is finite. Then by Lemma 25, $0 \notin Z$, and as Z is an interval, all of its elements have the same sign.

Suppose all $z \in Z$ are positive. Then if u is nonpositive, $u < \min Z/2$. If u is instead positive, then by Lemma 29, either $u < \min Z/2$ or $u > \max U[Z]$. By assumption, the latter is false, and thus $u < \min Z/2$. Since u is infeasible, $\Phi(u) + u > \min Z$ and hence $\Phi(u) > \min Z - u > \min Z/2 > 0$. Let $z \in Z$ such that $L(z) = \min L[Z]$. Then,

$$\begin{aligned}
\Phi(u) &= \min \mathrm{RU}[\mathrm{fl}^{-1}[Z] - u] \\
&\leq \mathrm{RU}(z - u) \\
&\leq \mathrm{RU}(z - \min L[Z]) \\
&= \mathrm{RU}(z - L(z)) \\
&= \mathrm{RU}(U(z))\,.
\end{aligned}$$

If $\mathrm{RU}(U(z)) = +\infty$, then $\max U[Z] \geq U(z) > \max \mathbb{F} \geq \Phi(u)$. Otherwise $\max U[Z] \geq U(z) = \mathrm{RU}(U(z)) \geq \Phi(u)$. Therefore, as $0 < \min Z/2 \leq \Phi(u) \leq \max U[Z]$, by Lemma 29, $\Phi(u)$ is feasible.

Suppose instead that all $z \in Z$ are negative. If u is nonnegative, then $u > \max Z/2$. If u is negative, then by Lemma 29 we have either $u < \min L[Z]$ or $u > \max Z/2$. As $u \geq \min L[Z]$ by assumption, $u > \max Z/2$. Since u is infeasible, $\Phi(u) + u > \max Z$. Let $z' \in Z$ such that $U(z') = \max U[Z]$. Then $\Phi(u) > \max Z - u \geq z' - u \geq z' - \max U[Z] = z' - U(z) = L(z') \geq \min L[Z]$. Thus,

$$\begin{aligned}
\Phi(u) &= \min \mathrm{RU}[\mathrm{fl}^{-1}[Z] - u] \\
&\leq \mathrm{RU}(\max Z - u) \\
&\leq \mathrm{RU}(\max Z/2)\,.
\end{aligned}$$

By Lemma 30, $\mathrm{RD}(\max Z/2) \oplus \mathrm{RU}(\max Z/2) = \max Z$ and hence $\mathrm{RU}(\max Z/2)$ is feasible. Suppose $\Phi(u) < \mathrm{RU}(\max Z/2)$. Then $\Phi(u) \leq \mathrm{RD}(\max Z/2) \leq \max Z/2$. Therefore, as $\min L[Z] \leq \Phi(u) \leq \max Z/2 < 0$, by Lemma 29, $\Phi(u)$ is feasible. □

As we know that feasible bounds always have an optimal counterpart, this leads straight to a solution. The variant of Lemma 31 for Ψ can be proved similarly, mutatis mutandis.

Lemma 32. If l is infeasible and $\Psi(l)$ is finite, then $\Psi(l)$ is feasible.

Lemma 33. If $l \oplus u < \min Z$ and u is infeasible and $\Phi(u)$ is finite, then $\Phi(u)$ is the optimal lower bound and $\Psi(\Phi(u))$ is the optimal upper bound.

Proof: Suppose the conditions hold. Then by Lemma 18, $\Phi(u) \oplus u > \max Z$ and hence $\Phi(u) > l$ by monotonicity. By Lemma 31, $\Phi(u)$ is feasible. Therefore, by Lemma 23, $\Phi(u) \oplus \Psi(\Phi(u)) \in Z$ and hence $\Psi(\Phi(u)) < u$ by monotonicity. Hence Lemmas 15 and 16 imply $\Phi(u)$ is the optimal lower bound and $\Psi(\Phi(u))$ is the optimal upper bound. □

In the corresponding case of an infeasible lower bound, we have the following result:

Lemma 34. If $l \oplus u > \max Z$ and l is infeasible and $\Psi(l)$ is finite, then $\Phi(\Psi(l))$ is the optimal lower bound and $\Psi(l)$ is the optimal upper bound.

With the case of infeasible bounds now also handled, we can finish finding optimal bounds in the general case.

5.3. Synthesis

Putting everything together, we now state our two main theorems. The first is a concise description of the optimal bounds $\Lambda(l, u)$ and $\Upsilon(l, u)$:

Theorem 4. Λ and Υ satisfy the following:

$$\Lambda(l,u) = \begin{cases} l & l \oplus u \in Z \\ \Phi(u) & l \oplus u < \min Z \\ \Phi(\Psi(l)) & l \oplus u > \max Z \text{ and } \Psi(l) > -\infty \\ +\infty & \text{otherwise} \end{cases}$$

and

$$\Upsilon(l,u) = \begin{cases} u & l \oplus u \in Z \\ \Psi(\Phi(u)) & l \oplus u < \min Z \text{ and } \Phi(u) < +\infty \\ \Psi(l) & l \oplus u > \max Z \\ -\infty & \text{otherwise} \end{cases}$$

Finally, from Lemma 10 and Theorem 4, we conclude that the optimal bounds can be found in constant time under mild conditions:

Theorem 5. Let X and Y be intervals over $\mathbb{F}$. Let

$$X' = \{x \in X \mid \exists y \in Y \, (x \oplus y \in Z)\},$$
$$Y' = \{y \in Y \mid \exists x \in X \, (x \oplus y \in Z)\}.$$

If floating-point operations, L, U, Φ and Ψ can be computed in constant time, then $\Box X'$ and $\Box Y'$ can also be computed in constant time.

Although the specifics of computing Φ and Ψ depend on the definition of fl, the work of Michel [8] provides a way of computing Φ and Ψ for any of the standard IEEE rounding modes using a constant number of floating-point operations. Thus, for the overwhelming majority of architectures, Theorem 4 gives us a way of computing an optimal answer in constant time.

6. Conclusion

Correct reasoning about floating-point operations is notoriously difficult, and at the same time of the utmost importance. For applications such as program analysis and program verification, performing *precise* reasoning without at the same time jeopardizing soundness is a considerable challenge.

In this paper we have generalized a result by Marre and Michel and established a number of other results about floating-point addition and subtraction. Based on these results we suggest an optimal interval bounds adjustment algorithm for floating-point constraints of form $x \oplus y = z$. The algorithm would run in constant time, in the sense that it would use a constant number of floating-point operations. This has practical applications for non-bit-blasting floating-point constraint solvers.

References

[1] D. Monniaux, "The pitfalls of verifying floating-point computations," *ACM Transactions on Programming Languages and Systems*, vol. 30, no. 3, pp. 12:1–12:41, May 2008.

[2] J. Souyris, V. Wiels, D. Delmas, and H. Delseny, "Formal verification of avionics software products," in *Proceedings of the 16th International Symposium on Formal Methods*, ser. Lecture Notes in Computer Science, A. Cavalcanti and D. R. Dams, Eds., vol. 5850. Springer Berlin Heidelberg, 2009, pp. 532–546.

[3] L. Haller, A. Griggio, M. Brain, and D. Kroening, "Deciding floating-point logic with systematic abstraction," in *Proceedings of the 12th Conference on Formal Methods in Computer-Aided Design*, ser. FMCAD '12, G. Cabodi and S. Singh, Eds. IEEE, Oct. 2012, pp. 131–140.

[4] M. Brain, V. D'Silva, A. Griggio, L. Haller, and D. Kroening, "Deciding floating-point logic with abstract conflict driven clause learning," *Formal Methods in System Design*, vol. 45, no. 2, pp. 213–245, Oct. 2014.

[5] K. Scheibler, F. Neubauer, A. Mahdi, M. Fränzle, T. Teige, T. Bienmüller, D. Fehrer, and B. Becker, "Accurate ICP-based floating-point reasoning," in *Proceedings of the 16th Conference on Formal Methods in Computer-Aided Design*, ser. FMCAD '16, R. Piskac and M. Talupur, Eds. Austin, TX: FMCAD Inc., Oct. 2016, pp. 177–184.

[6] B. Marre, F. Bobot, and Z. Chihani, "Real behavior of floating point numbers," in *Proceedings of the 15th International Workshop on Satisfiability Modulo Theories*, Jul. 2017. [Online]. Available: http://smt-workshop.cs.uiowa.edu/2017/papers/SMT2017_paper_21.pdf

[7] B. Marre and C. Michel, "Improving the floating point addition and subtraction constraints," in *Proceedings of the 16th International Conference on Principles and Practice of Constraint Programming*, ser. Lecture Notes in Computer Science, D. Cohen, Ed., vol. 6308. Springer Berlin Heidelberg, Sep. 2010, pp. 360–367.

[8] C. Michel, "Exact projection functions for floating point number constraints," in *Seventh International Symposium on Artificial Intelligence and Mathematics*, ser. AI&M 15-2002, Fort Lauderdale, FL, USA, Jan. 2002.

[9] D. Gallois-Wong, S. Boldo, and P. Cuoq, "Optimal inverse projection of floating-point addition," *Numerical Algorithms*, 2019, in press. [Online]. Available: https://hal.inria.fr/hal-01939097v1

[10] P. H. Sterbenz, *Floating-Point Computation*, ser. Prentice-Hall Series in Automatic Computation. Englewood Cliffs, NJ: Prentice-Hall, 1973.

Author/s:
Andrlon, M; Schachte, P; Sondergaard, H; Stuckey, PJ

Title:
Optimal Bounds for Floating-Point Addition in Constant Time

Date:
2019-06

Citation:
Andrlon, M., Schachte, P., Sondergaard, H. & Stuckey, P. J. (2019). Optimal Bounds for Floating-Point Addition in Constant Time. Takagi, N (Ed.) Boldo, S (Ed.) Langhammer, M (Ed.) Proceedings of the 2019 IEEE 26th Symposium on Computer Arithmetic (ARITH), 2019-June, pp.159-166. IEEE. https://doi.org/10.1109/arith.2019.00038.

Persistent Link:
http://hdl.handle.net/11343/241776

File Description:
Accepted version