Additive symmetries: the non-negative case

Marc Daumas*, Philippe Langlois

Laboratoire de l’Informatique du Parallélisme, UMR 5668 CNRS, INRIA, ENS de Lyon, Ecole Normale Supérieure de Lyon, 46, allée d’Italie, F-69364 Lyon, France

Received 22 January 2001; received in revised form 28 May 2001; accepted 8 June 2001

Abstract

An additive symmetric value \( b \) of \( a \) with respect to \( c \) satisfies \( c = (a + b)/2 \). Existence and uniqueness of such \( b \) are basic properties in exact arithmetic that fail when \( a \) and \( b \) are floating point numbers and the computation of \( c \) performed in IEEE-754-like arithmetic. We exhibit and prove conditions on the existence, the uniqueness and the consistency of an additive symmetric value when \( b \) and \( c \) have the same sign. © 2002 Elsevier Science B.V. All rights reserved.

Keywords: Floating point arithmetic; Additive symmetric value; Correction; IEEE-754 standard

1. Introduction

“Floating point arithmetic is by nature inexact.” This quotation from Knuth [9] summarizes that floating point arithmetic only approximates real arithmetic. The discrepancies between the approximate and the exact arithmetics are numerous and are due to the finite precision of \( F \), the set of the floating point numbers. Failures of fundamental laws of algebra are well known for floating point arithmetic. For example, the associativity of the addition or the multiplication, the cancellation or the distributivity laws are no longer valid in \( F \) [15]. Two other fundamental axioms for real algebra state the existence and the uniqueness of the additive inverse \((-a)\) and the multiplicative reciprocal \((1/b)\) for \( a \in \mathbb{R} \) and \( b \in \mathbb{R}^*\). These axioms also fail in floating point arithmetic as proved in [10,14] and illustrate well how subtle the discrepancies are.

Here, we consider the existence and the uniqueness of an additive symmetric value in the floating point number set \( F \). An additive symmetric value \( a \) of \( b \) with respect
to $c$, satisfies
\[
  c = fl\left(\frac{a + b}{2}\right),
\]
where $b$ and $c$ are two given floating point numbers in $\mathbb{F}$. The classic notation $fl(x) \in \mathbb{F}$ represents the rounded floating point value of $x \in \mathbb{R}$. Of course,
\[
a_e = 2c - b
\]
is the unique additive symmetric value where rounding affects none of relations (2) and (1).

Let $b$ and $c$ in $\mathbb{F}$ when $\mathbb{F}$ is a set of floating point numbers à la IEEE-754 arithmetic—a symmetric set of binary floating point numbers with denormalized (subnormal) numbers—and for the “round to the nearest (even)” rounding mode.

Existence: Does an additive symmetric value $a$ exist within $\mathbb{F}$?
Uniqueness: Is the additive symmetric value unique?
Consistency: Does $a = fl(a_e)$?

In this paper, we answer to these three questions when $b$ and $c$ have the same sign. Choosing for instance $a$ and $b$ non-negative, we consider the non-negative case of the additive symmetry.

To derive the answers to these questions, we interpret additive symmetry as a correcting operator. Let $v$ be a floating point number in $\mathbb{F}$. Correcting $v$ with the correcting term $z \in \mathbb{F}$ means computing
\[
  t(z) = fl(v + z).
\]
Let $y$ be a correcting term that yields the target $t \in \mathbb{F}$ from the value $v$, that is $t(y) = t$. This correcting term $y$ is an additive symmetric value of $v$ with respect to $t/2$ when $t/2 \in \mathbb{F}$. Again, where rounding affects neither the computation of the correcting term, nor the correction in (3)
\[
  y_e = t - v
\]
is the unique exact correcting term. Existence, uniqueness and consistency of the additive symmetric value correspond now to the following three questions Q1-3 we examine in the sequel.

Q1. Does a correcting term $y$ exist within $\mathbb{F}$?
Q2. Is the correcting term unique?
Q3. Does $y = fl(y_e)$?

The paper is organized as follows. We present the motivations with an introductory example and connected results in the next Section. We summarize the properties of the additive symmetry in Section 3 and devote following Section 4 to the proofs—the main result of the paper is the summary Fig. 1. We conclude describing questions that remain open.
Notations. We use the classic following notations defined for $F$. We denote $u(x) = ulp(x)$, one unit in the last place of the floating point $x \in F$, and $u = u(1)$, one $ulp$ of one. Let $x = (-1)^{s} \times f \times 2^{e}$, with a $p$ bits fractionary part $f$. We have $u(x) = \max \{2^{e-p}, \lambda_{d}\}$, and $u(0) = \lambda_{d}$, where $\lambda_{d}$ is the smallest positive denormalized floating point number. We verify that $u(2^{k} \times x) = 2^{k} u(x)$, for $x$ and $(2^{k} \times x)$ in $F$. The IEEE-754 standard defines $p = 23$ for single precision and $p = 52$ for double precision. Since $F$ is a discrete set, we denote $x^{-}$, the predecessor and $x^{+}$, the successor of each floating point number $x$ provided no overflow occurs. For $x > 0$, these floating point numbers verify, $x^{-} = x - u(x)/2$ when $x = 2^{k}$ and $x$ is normalized, $x^{-} = x - u(x)$ elsewhere, and $x^{+} = x + u(x)$.

2. Motivations and connected results

We illustrate the motivations of the questions Q1-3 with an introductory example and then discuss more general applications of the additive symmetry. We end this section presenting connected results about the additive inverse and the multiplicative reciprocal.

2.1. An introductory example

We propose three simple pairs of floating point numbers $(b, c)$ that exhibit the different possible cases of existence, uniqueness and consistency of a corresponding additive symmetric value. We denote $AS[b, c]$ an additive symmetric value of $b$ with respect to $c$. We use the correspondence between the correcting term $y$ that yields $t$ from $v$, and the additive symmetric value of $b = v$ with respect to $c = t/2$ when $t/2 \in F$. We have $y = AS(v, t/2)$.
**Example 1** (The additive symmetric value $\text{AS}[2^+,1/2]$ exists, is unique and consistent). We consider the corresponding correction of $v=2^+$ that returns $t=1$. The exact correcting term $y_e = -(1 + u(2))$ belongs to $\mathbb{F}$, so $\hat{y}_e = y_e$. The correcting term $y_e$ verifies

$$v + y_e = 1$$

that yields the expected value $t=1$. The two neighbors of $y_e$ in $\mathbb{F}$ are $y_e^+ = -1 - u$, and $y_e^- = -1 - 3u$. We have

$$v + y_e^+ = 1 + u = t^+$$

and

$$v + y_e^- = 1 - u = t^{--},$$

where $t^{--}$ is the predecessor of $t^-$. Thus, the corrected values of $v$ corresponding to consecutive correcting terms are

$$t(y_e^-) = t^{--}, \quad t(y_e) = t \quad \text{and} \quad t(y_e^+) = t^+.$$

The monotonicity of the rounding map $f/l$ ensures that $y_e$ is the unique correcting term that returns $t$. So it is for the additive symmetric value of $b$ with respect to $c$ for the considered value $(2^+,1/2)$.

We remark that $t(y_e^-)$ and $t(y_e)$ are not consecutive and $t(y_e^-) < 1^- < t(y_e)$. We verify that no correcting term exists for $t=1^-$ and $v=2^+$, or similarly, no additive symmetric value of $b=2^+$ with respect to $c=1/2 - u/4$. We detail a similar case of non-existence in the following example.

**Example 2** (The additive symmetric value $\text{AS}[5,(1/2)^+]$ does not exist). Relation (3) gives no correction of $v=5$ that returns $t=1^+$. We have $y_e = -4 + u = -4 + u(2)/2$. The tie-breaking strategy of the “round to the nearest (even)” yields $\hat{y}_e = -4$, and

$$v + \hat{y}_e = 1 < t.$$

With the next larger value $\hat{y}_e^+ = -4 + u(2)$, we have now

$$v + \hat{y}_e^+ = 1 + u(2) = 1^{++},$$

where $1^{++}$ is the successor of $t$. The corrected values that correspond to the two consecutive correcting terms $\hat{y}_e$ and $\hat{y}_e^+$ are

$$t(\hat{y}_e) = t^- \quad \text{and} \quad t(\hat{y}_e^+) = t^+.$$

These values enclose the target value $t$ but neither are equal to $t$. Therefore, an additive symmetric value of $b=5$ with respect to $c=(1/2)^+$ does not exist.
Example 3 (The additive symmetric value $AS[1, 1]$ is non-unique). This case is less surprising rephrased as follows: the three correcting terms $\hat{y}_e = y_e$, $y_e^-$ and $y_e^+$ return $t = 2$ from $v = 1$. For these values, the exact and rounded correcting terms verify $y_e = 1$. We have

\[
v + y_e = 2,
\]
\[
v + y_e^+ = 2 + u
\]

and

\[
v + y_e^- = 2 - u/2.
\]

Using the tie-breaking strategy of the round the nearest (even) in the two last equalities, we conclude that

\[
t(y_e) = t(y_e^-) = t(y_e^+) = t.
\]

2.2. Motivations

The original motivation to the theoretical study of the additive symmetry comes from the CENA method introduced by one of the authors in [12].

The CENA method. The CENA method provides an automatic correction of the first-order effect of floating point rounding errors to the result of numerical algorithms. This correction is applied to the final result or to sensitive intermediate variables of a computation. Depending on some algorithm properties, the final correction improves the accuracy of the computed result and the intermediate correction enhances the numerical stability of the algorithm.

Given a computed value $v$, the CENA method yields a corrected $t$ defined as

\[
t = fl(v + A_L), \tag{5}
\]

where the correcting term $A_L$ is the computed linearization of the error in $v$ with respect to the rounding errors introduced in the intermediate computations (see [11] for a complete description).

The two main limitations for the efficiency of correction (5) come from the significance of the correcting term $A_L$ and the accuracy of the correcting addition. The CENA method computes a significant correcting term when computed $v$ suffers mainly from the first-order effect of the elementary rounding errors. This paper highlights the intrinsic limitation of the correcting addition: we prove conditions that ensure the existence (or the non-existence) of $A_L$ in the floating point set $\mathbb{F}$ that yields the expected corrected result $t$.

Another motivation for the additive symmetry is described in the next paragraph.

New value = Old value + Correction. Numerical methods often consist of such an update strategy, for example, see the remarks on p. 30 in Higham [8]. Computing a more accurate approximate adding a correcting term to a previous approximate is the core of
iterative methods. Newton’s method and the classic iterative refinement that improves the accuracy of the solution to a linear system $Ax=b$ are examples that implement this strategy. It is also the case for integration schemes for ordinary differential equations. To ensure the convergence of iterative methods, the correcting term is designed to tend to zero. For example, the correcting term of the iterative refinement is the residual $r = b - Ax$. This small correction is a particular case of the additive symmetry. A full precision accuracy of the new value is necessary in specific iterations, as for example in the computation of the elementary functions. Again, the answers to questions Q1–3 highlight the limitations of this general strategy when it is applied in finite precision.

2.3. Connected results

In the Introduction, we have noted that the existence and the uniqueness of the additive inverse $(-a)$ and the multiplicative reciprocal $(1/b)$, $b \neq 0$, fail in floating point arithmetic. We summarize some recent results concerning these connected problems.

The additive inverse. An additive inverse $(-a) \in F$ satisfies $fl(a + (-a)) = 0$, for $a \in F$. In [10], Kulisch discusses the uniqueness of the additive inverse in a discrete symmetric subset $S \subset \mathbb{R}$, and for a general rounding map $fl$ defined from $\mathbb{R}$ to $S$. He proves that $fl^{-1}(0) \subseteq \varepsilon, \varepsilon [0, \min_{x, y \in F, x \neq y} |x - y|]$, guarantees the existence of a unique additive inverse. This result applies to the four rounding modes of the IEEE-754 floating point arithmetic—the round to the nearest (even) default rounding mode, and the directed rounding modes “round towards zero” and “round towards infinities”. It is not surprising that unique additive inverses exist for the four rounding modes when denormalized (subnormals) floating point numbers and gradual underflow are available—as it is the case for the IEEE-754 arithmetic. Kulish also points out that the “rounding away from zero” mode satisfies $fl(x) = 0 \iff x = 0$ for $x \in S$, with or without denormalized numbers.

The multiplicative reciprocal. The multiplicative reciprocal $(1/b)$ satisfies $fl(b \times (1/b)) = 1$ for $b \in F^*$. The existence and the uniqueness of this reciprocal depends on the rounding mode and the value of $b$. Muller proves the following results in an IEEE arithmetic-like context when underflow and overflow are neglected [14]. A unique multiplicative reciprocal exists for the ‘round towards $+\infty$’ rounding mode. The other modes yield at most 2 reciprocals for $b \in F^*$: the floating point numbers that enclose (the real value) $1/b$. The existence of the multiplicative reciprocal depends on the mantissa length $p$ of the floating point number. Muller conjectures that the number $\gamma(p)$ of floating point numbers with no multiplicative reciprocal is such that the ratio $\gamma(p)/2^p$ tends to the constant $(1 - 3 \log(4/3))/2$ when $p$ goes to infinity. A recent proof of this conjecture is proposed in [5].

Edelman shows a less general property for the IEEE-754 double precision floating point numbers and the round to the nearest (even) rounding mode in [4]. After having proved that $x \times fl(1/x) \in \{1 - u/2, 1\}$, he exhibits the smallest double precision number $1 < x < 2$ such that $fl(x \times fl(1/x)) \neq 1$. This result illustrates the well-known fact that $fl(x \times fl(1/y))$ is not necessary equal to $fl(x/y)$. 
3. Properties of the non-negative case of the additive symmetry

Relations between $b$ and $c$ govern the existence and the uniqueness of the additive symmetric value of $b$ with respect to $c$. We answer to questions Q1-3 proving relations on the corresponding correction problem (3) with $t = 2c$ and $v = b$. We chose to parameterize the discussion with respect to the initial value $v$ of the correction problem

$$t = fl(v + y).$$  \hfill (6)

We prove conditions on the target value $t$, depending on a given value $v$, such that a correcting term $y$ that verifies relation (6) exists, is unique and consistent.

Let $t$ be of the same sign than $v$, for example positive. We distinguish six regions $R_1, \ldots, R_6$, that depend on $v$ in $[0, A]$, the positive part of $\mathbb{F}$ where $A$ denotes the largest positive floating point number. The existence, the uniqueness and the consistency of a correcting term $y$ verifying relation (6) vary with respect to the region where $t$ belongs. We summarize these properties with Fig. 1 and devote the next section to the proofs.

These regions depend on the following functions $U$ and $A$. The function $U(x)$ measures the distance between $x$ and its closest floating point neighbor

$$U(x) = \min\{x^+ - x, x - x^-\}.$$ We verify that $U(x) = u(x)$, except for $x = 2^k$ and $x$ normalized where $U(x) = u(x)/2$. In all cases, $U(x) \geq \lambda_d$. The function $A(x)$ defines the smallest number $a_0$ such that $fl(a + |x|) = a$ for all the floating point numbers $a \geq a_0$. We have $A(x) = 2^{e+p+2}$ for $x = (-1)^s f \times 2^e$ ( $f$ has $p$ bits), and $A(0) = 0$. For example, if $x = u/4$, $A(x) = 1$ since for all $b \geq 1$, $fl(b + x) = b$ but $fl(1^- + x) = 1 \neq 1^-$.\hfill (7)

4. Proofs of the properties

We adapt the important result from Sterbenz [15] (also presented in [8, p. 50]) to give a sufficient condition for an “exact subtraction” in $\mathbb{F}$. Hypotheses on $\mathbb{F}$ arithmetic defined in the Introduction insure the use of a guard bit and gradual underflow; these conditions have been formally validated to be sufficient to the following lemma in [3].

**Lemma 1** (Sterbenz [15]). Let $a$ and $b$ be in $\mathbb{F}$ with $a \leq b \leq 2a$. The floating point subtraction $fl(b - a)$ introduces no rounding error, i.e.

$$fl(b - a) = b - a.$$
into consideration of many possible correcting terms. These proofs are presented in Section 4.2.

4.1. Results obtained using properties of \( \mathbb{F} \)

We recall that the floating point numbers \( t \) and \( v \) share the same sign and we suppose that they are both positive. It is straightforward to check that except in the trivial case \( t = v \), any acceptable \( y \) must have the same sign as \( t - v \).

**R_1** \( t = 0 \). This is the additive inverse case which is analyzed with a more general point of view in [10] as previously mentioned. For completeness, we propose the following proof in the current context.

Relation (6) is satisfied for \( y = y_e = -v \). The exact correction \( y_e \) is the unique solution. Let \( y = -v + z \) be any correcting term in \( \mathbb{F} \) with \( z \in \mathbb{R} \). Either \( z = 0 \) or \( |z| \geq U(v) \) by definition of \( U(v) \). Then \( f\ell(v + y) = f\ell(z) \) but \( |f\ell(z)| \geq f\ell(U(v)) > 0 \).

**R_2** \( \lambda_d \leq t < U(v) \). This condition implies that \( v \neq 0 \) and \( y_e = t - v < 0 \). We only look for \( y < 0 \), since \( y \geq 0 \) gives \( f\ell(v + y) \geq v \geq U(v) > t \). We distinguish three cases to prove that \( f\ell(v + y) \neq t \).

(1) When \( -y > 2v \), \( v + y < -v \) and \( f\ell(v + y) < -v < 0 < t \).

(2) When \( -y < v/2 \), we have \( v \geq 2\lambda_d \) for \( y \neq 0 \). It follows that \( v/2 \geq U(v) \) and as \( v + y > v/2 \), we prove that \( f\ell(v + y) \geq f\ell(v/2) \geq U(v) > t \).

(3) When \( v/2 \leq -y < 2v \), the “exact subtraction” result from Sterbenz applies. This means here \( f\ell(v + y) = v + y \). The number \( v - t \notin \mathbb{F} \), since it is closer to \( v \) than \( v \pm U(v) \), its closest neighbor in \( \mathbb{F} \). Therefore no floating point number \( y \) satisfies relation (6).

No correcting term \( y \) satisfies (6) for \( t \) in this region \( R_2 \).

**R_3** \( U(v) \leq t < v/2 \). We prove the existence of \( y \) in this region for \( t \) such that \( t/uv \) is an integer. We define the positive integers \( m' = v/uv \) and \( m = t/uv \). For \( t \) in this region, \( u(t) < u(v) \) and \( v - t = (m' - m)uv \) verifies \( |m' - m| \leq m' < 2u^{-1} \). Therefore, \( t - v \in \mathbb{F} \) and \( y_e = f\ell(t - v) = t - v \) verifies relation (6).

This proof extends Sterbenz equality for the subtraction \( v - t \) to \( t < v/2 \) provided that \( t/uv \) is an integer. Nothing direct seems to be derivable in region \( R_3 \) about the uniqueness, nor when \( t/uv \) is not an integer. We consider these remaining aspects in Section 4.3.

**R_4** \( v/2 \leq t \leq 2v \). From Sterbenz’s equality, we have \( y_e = f\ell(t - v) = t - v = y_e \) and \( f\ell(v + y_e) = f\ell(t) = t \). Existence and consistency are proved in region \( R_4 \).

Again, discussing the uniqueness and the consistency uses actual values and is considered at the end of Section 4.3.

**R_5** \( 2v < t \leq A(v) \). A similar discussion as in region \( R_3 \) yields analogous partial results on existence and consistency. We chose to present the complete discussion of this region in next Section 4.2.
(R₆) A(v) < t ≤ A. We recall t > A(v) means f(l(t + v)) = t. In these regions, relation (6) is satisfied for y = t = ˆyₑ. We prove uniqueness while discussing region R₅ in following Section 4.2.

4.2. Results obtained using the actual values of the floating point numbers

Using the actual values of the floating point numbers v and t, we prove conditions on existence, uniqueness and consistency for y satisfying relation (6) in the regions R₅ and R₆, that is for

\[ 2v < t \leq A. \]  

Principles of the proof. We compute

\[ fl(v + y) \text{ for } y \in \{ x, \beta, \gamma, \delta \}. \]  

The four numbers \( x < \beta < \gamma < \delta \) are such that

\[ [x, \delta] \cap \mathbb{F} \subseteq \{x, \beta, \gamma, \delta\}, \quad \beta \in \mathbb{F}, \quad \gamma \in \mathbb{F} \text{ and } yₑ \in [\beta, \gamma]. \]  

We note that \( \beta \) and \( \gamma \) are consecutive floating point numbers and \( \beta = ˆyₑ \) or \( \gamma = ˆyₑ \). In most cases, \( x \) and \( \delta \) are such that \( x = \beta^- \) and \( \delta = \gamma^+ \).

Notations for the proof. Let \( v \) and \( t \) be positive integers that verify previous condition (7). We define the integer

\[ m = t/u(t) \]  

and a unique 4-uple \((k, g, r, s)\) with \( k \in \mathbb{N}, \ g \in \{0, 1\}, \ r \in \{0, 1\} \) and \( s \in [0, 1) \) such that

\[ v = \left( k + \frac{g}{2} + \frac{r + s}{4} \right) u(t). \]  

The notation \((g, r, s)\) stands for the guard, round and sticky bit of IEEE-754 arithmetic as in [1].

Normalized representation for \( t \) provides \( u^{-1} \leq m < 2u^{-1} \). The condition \( 2v < t \) yields \( 0 \leq k < m/2 \), so \( m - k \geq m/2 \) and \( m - k \geq m/2 + 1 \geq u^{-1}/2 + 1 \). Equality holds only for \( m = u^{-1} \).

From (10) and (11), \( yₑ = t - v \) verifies

\[ yₑ = \left( m - k - 1 + \frac{1 - g}{2} + \frac{1 - r}{4} + \frac{1 - s}{4} \right) u(t). \]  

The next three tables detail the computation of relation (8) for each value of \( y \) and the different values of the parameters \( r \) and \( s \) that govern the rounding of relation (12) and relation (8). We define the reference point

\[ z = (m - k - 1)u(t) \]  

that will vary around \( yₑ \). Since \( u^{-1}/2 \leq m - k - 1 < 2u^{-1} \), \( z \) is a floating point number not necessarily normalized. We discuss the computation of relation (8) for varying \( u(z) \).
Table 1

| Rounding conditions | \( r' = 0 \) | \( r' = 1 \)
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( s' = 0 )</td>
<td>( s' \neq 0 )</td>
</tr>
<tr>
<td>( y )</td>
<td>( y / u(t) )</td>
<td>( (m - k) u(t) )</td>
</tr>
<tr>
<td>( \delta )</td>
<td>( m - k + 1 )</td>
<td>( m + \frac{r' + s'}{2} + 1 )</td>
</tr>
<tr>
<td>( \gamma )</td>
<td>( m - k )</td>
<td>( m + \frac{r' + s'}{2} )</td>
</tr>
<tr>
<td>( \beta )</td>
<td>( m - k - 1 )</td>
<td>( m + \frac{r' + s'}{2} - 1 )</td>
</tr>
<tr>
<td>( z )</td>
<td>( m - k - \frac{3}{2} )</td>
<td>( m + \frac{r' + s'}{2} - \frac{3}{2} )</td>
</tr>
</tbody>
</table>

with the following tables. From relation (13), we have \( u(t)/2 \leq u(z) \leq u(t) \). The different cases for the \( ulp \) of \( z \) and its neighbors are \( u(t)/4, u(t)/2 \) and \( u(t) \); these values guide the following discussion. We consider the cases \( u(z) = u(t)/2 \) and \( u(z) = u(t) \)—values for the neighbors appear implicitly in the discussion.

The tables also present \( \hat{y}_e \), the rounded value of relation (12). Comparing \( \hat{y}_e \) and \( y \) when \( f l(v + y) = t \) yields the answer to Q3 (consistency). We indicate some cases where \( \hat{y}_e = \beta \) or \( \gamma \) is sufficient to derive a (positive or negative) answer to Q3.

**Discussion.** (1) \( u(z) = u(t) \). When \( u(t) \neq \lambda_d \), it means that \( z \) is normalized and \( m - k - 1 \geq u^{-1} \). Let \( \beta = z \), its successor is \( z^+ = (m - k) u(t) \). Its predecessor is \( z^- = (m - k - 2) u(t) \), except when \( m - k - 1 = u^{-1} \) where \( z^- = (m - k - 3/2) u(t) \). We note that \( m \neq u^{-1} \). If \( u(t) = \lambda_d \), most relations still hold and \( z^- = (m - k - 2) u(t) \).

We verify that the guard bit \( g \) is not used when computing (8) since no value is shifted. Thus we simplify the notations defining \( r' \in \{0, 1\} \) and \( s' \in [0, 1) \) such that

\[
\frac{g}{2} + \frac{r + s}{4} = \frac{r' + s'}{2}.
\]

We introduce the following last notations in the next three tables. The function \( E(m, m^+) \) returns the even value between the two consecutive integers \( m \) and \( m^+ \). The string MNT identifies a cell that does not have to be computed since it differs certainly from the target value applying the monotonicity of the rounding to one of its neighboring value (upward or downward cell). The strict monotonicity yields a similar conclusion when the neighbor value exhibits (with function \( E \)) the use of the tie-breaking rule.

Now we explore the Table 1 with respect to \( (r', s') \) that governs the rounding in \( f l(v + y) \). This explicit computation gives the following answers to questions Q1–3.
Table 2

<table>
<thead>
<tr>
<th>Rounding conditions</th>
<th>( r = 0 )</th>
<th>( r = 1 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( s = 0 )</td>
<td>( s \neq 0 )</td>
<td>( s \neq 0 )</td>
</tr>
</tbody>
</table>

\[
\hat{y}_e = \frac{(m - k + \frac{1 - s}{2} - \frac{1}{2})u(t)}{u(t)}
\]

\[
\beta \text{ or } \gamma
\]

- \( Q_1 \) is positive except when \((r', s') = (1, 0)\) for odd \( m \),
- \( Q_2 \) is positive except when \((r', s') = (1, 0)\),
- \( Q_3 \) is positive with \( Q_1 \), and
- two correcting terms exist when \((r', s') = (1, 0)\) for even \( m \).

(2) \( u(z) = u(t)/2 \). In this case, \( u^{-1}/2 \leq m - k - 1 < u^{-1} \). The mantissa of \( z \) is shifted to compute (8) and the guard bit \( g \) introduces the quantity \((1 - g)/2\) we consider to round \( y_e \). We use \( \beta = (m - k - 1 + (1 - g)/2)u(t) \). We note that \( u(\beta) = u(z) = u(t)/2 \). This gives us relations to define \( \delta \) and \( \gamma \) as the respective neighbors of \( \delta \) and \( \beta \) satisfying relation (9). We separate two cases with respect to \( m = u^{-1} \) and we build the Tables 2 and 3 with respect to \((g, r, s)\) since \((r, s)\) governs the rounding in \( f(l(v + y))/u(t) \).

(a) When \( m = u^{-1} \), the predecessor of \( t \) is \((m - 1)u(t)\) and the gap between the four considered values for \( y \) is not smaller than \( u(t)/2 \).

We note that \( \delta \) does not necessarily belong to \( \mathbb{F} \). The conclusions are now the followings.

- \( Q_1 \) and \( Q_3 \) are positive in all the cases,
- \( Q_2 \) is positive when \((r, s) = (0, 0)\) and for odd \( m \),
- two correcting terms exist when \( r = 1 \) or \((r, s) = (0, s)\) with \( s \neq 0 \),
- three correcting terms exist when \((r, s) = (0, 0)\) and for even \( m \).

(b) Table 3 gives the answer when \( m = u^{-1} \). In this case, a \( u(t)/4 \) gap may exist between the floating point values \( \alpha \) and \( \beta \).

We derive the following conclusions.

- \( Q_1 \) and \( Q_3 \) are positive in all the cases,
- \( Q_2 \) is positive when \((r, s) = (0, s)\) with \( s \neq 0 \),
- two correcting terms exist when \( r = 1 \) or eventually when \((r, s) = (0, 0)\).

**Conclusion for the regions \( R_5 - R_6 \).** Now we derive the answers for questions Q1–3 in the regions \( R_5 - R_6 \) from the results of the previous discussion.
Table 3

<table>
<thead>
<tr>
<th>Rounding conditions</th>
<th>( r = 0 )</th>
<th>( r = 1 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( s = 0 )</td>
<td>( s \neq 0 )</td>
<td>( \hat{y}_e )</td>
</tr>
<tr>
<td>( y )</td>
<td>( y/u(t) )</td>
<td>( (v + y)/u(t) )</td>
</tr>
<tr>
<td>( \delta )</td>
<td>( m - k + \frac{1-g}{2} )</td>
<td>( m + \frac{r+s}{4} + \frac{1}{2} )</td>
</tr>
<tr>
<td>( \gamma )</td>
<td>( m - k + \frac{1-g}{2} - \frac{1}{2} )</td>
<td>( m + \frac{r+s}{4} )</td>
</tr>
<tr>
<td>( \beta )</td>
<td>( m - k + \frac{1-g}{2} - 1 )</td>
<td>( m + \frac{r+s}{4} - \frac{1}{2} )</td>
</tr>
<tr>
<td>( \alpha )</td>
<td>( m - k + \frac{1-g}{2} - \frac{5}{4} )</td>
<td>( m + \frac{r+s}{4} - \frac{3}{4} )</td>
</tr>
</tbody>
</table>

The value of the cell is obtained from even rounding since \( m = E(m, m + 1) \). It means that \( \delta \) is the largest quantity that may return \( m \). When \( g = 0 \) and \( m - k = u^{-1} \), \( \delta \) is not a floating point number and \( \gamma = \delta + u(t)/2 \) yields the corrected value \( (m + 1)u(t) \). In this case, the solution \( \hat{y}_e \) is unique and consistent.

We prove that Q1–3 are positive in \( R_6 \). In this region, \( A(v) < t \) gives \( k = g = r' = 0 \). When \( u(z) = u(t) \), the column \( r' = 0 \) in Table 1 yields the result.

When \( u(z) = u(t)/2 \), the condition \( u^{-1}/2 \leq m - k - 1 < u^{-1} \) yields \( m - 1 < u^{-1} \) for \( k = 0 \). As \( u^{-1} \leq m < 2u^{-1} \), the case \( u(z) = u(t)/2 \) is only possible when \( m = u^{-1} \). This case corresponds to Table 3. The discussion of this case when \( g = r = 0 \) gives the expected positive answers. The case \( r = 1 \) would imply that \( t = A(v) \) that is not permitted by the definition of \( R_6 \).

We prove that Q1 and Q3 are positive in \( R_5 \) provided the tie breaking mechanism does not prevent any \( y \) to be the correct answer. The answer to Q2 is known from the three tables but we cannot simplify the conditions on \( v \) and \( t \) to a few high level relations. If the answer to Q2 is no, there are at most three acceptable correcting terms.

4.3. Results that could be obtained using the actual values of the floating point numbers

The following cases remain from the direct discussion of Section 4.1. We explicit the remaining cases to prove using similar explicit computations.

- Q2 in region \( R_3 \) when \( t/u(v) \in \mathbb{Z} \), and Q1–3 otherwise,
- Q2 in region \( R_4 \).

Similar explicit computation will yield the answers of these questions. We do not propose another long discussion in this paper but we illustrate the answers presented with Fig. 1 exhibiting examples of the different cases.

Region \( R_3 \). Example 1 in Section 2.1 illustrates the positive answers to questions Q1–3 in the region \( R_3 \). We verify that \( t = 1 \) and \( v = 2^+ \) satisfy \( t/u(v) \in \mathbb{N} \).
**Region R₃.** Example 2 in Section 2.1 illustrates the negative answer to question Q1 in the same region R₃. Here \( t = 1^+ \) and \( v = 5 \) are not such that \( t/u(v) \in \mathbb{N} \).

**Region R₄.** Example 3 in Section 2.1 illustrates the negative answer to question Q2 in the region R₄.

**Region R₄.** With the following Example 4, we exhibit a case where the correcting term is unique and so a positive answer to Q2 in the same region R₄.

**Example 4.** We chose \( v = 1^+ = 1 + u \) and \( t = (1/2)^+ = (1 + u)/2 \). The exact correcting term is

\[
y_e = - \left( \frac{1}{2} + \frac{1}{2} u \right) = \hat{y}_e.
\]

The computation of the correction \( f(l(v + y)) \) for the two neighbors of \( y_e \) yields

\[
fl(v + y_e^-) = t^- \quad \text{and} \quad fl(v + y_e^+) = t^+.
\]

The unique correcting term is \( y_e \).

5. Conclusion

The additive symmetric operator has well-known properties and is used often in exact arithmetic. We considered how its properties changes when floating point arithmetic is used rather than exact arithmetic. Results and proofs have been presented using the corresponding correction operator and restricting our study to a particular arithmetic—IEEE-754 arithmetic and the round to the nearest (even) rounding mode.

Motivations of this study are its connections with the automatic correcting method CENA and the classic update strategy “New value = Old value + Correction.” As regions of non-existence of a correcting term have been exhibited, we validate intrinsic limitations of the CENA method. The main domain of limitation (Region R₂) corresponds to an inaccurate initial (absolute) result that is too large with respect to the exact value to be corrected in finite precision arithmetic. When inaccuracy grows with the computation, such a limitation is circumvented by the correction of intermediate variables. Of course designing a general dynamic choice of such a switch is a difficult task. The existence of a correcting term in region R₄ validates the classic update strategy since the correcting factor is designed to tend to zero. Hence, the difficulty remains of choosing a good initial value for these iterative processes so convergence is assured.

The way the results are proved in this paper are significant of the derivation of properties satisfied by floating point arithmetic. When general properties apply, direct proofs are possible and consist of simple algebraic derivations. Unfortunately, most cases require long and tedious computations to cover all possibilities. As this kind of derivation may suffer from human mistakes, automatic formal provers should be applied to validate such results. Examples of formal validation of floating point properties are
Of course, validated results are the general theorems that promote future direct proofs.

When the additive symmetric value does not exist, the following natural question arises: what is the best approximate additive symmetric value? When the existence is proved, we note that the consistent additive symmetric value—or correcting term—is always (one of) the solution(s) of the problem. Let us consider now the following slight generalization of our correction problem (3). Let the value \( v \) be a floating point number and the target \( t \) be a real number. Is the answer of question Q3 still positive when Q1 is positive? The following example suggested by Muller illustrates it is not the case. When \( t = 8 + u(1) - u(10) - \xi \), with \( \xi > 0 \) small enough, we have

\[
fl(t) = 8^+.
\]

Choosing \( v = 1 - u(1) \), we verify that \( \hat{y} = 7^+ = 7 + u(4) \) and

\[
fl(v + \hat{y}) = 8 = fl(t)^-,
\]

whereas

\[
fl(v + \hat{y}^+) = 8^+ = fl(t).
\]

The non-exact \( \hat{y}^+ \) yields the expected corrected value. "Floating point arithmetic is by nature inexact."

Acknowledgements

We thank the anonymous referees for their detailed and constructive comments that help us to improve the presentation and the correctness of the manuscript.

References


