Additive Symmetric: the Non-Negative Case

Marc Daumas, Philippe Langlois

INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Additive symmetric: the non-negative case

Marc Daumas, CNRS
Philippe Langlois, INRIA

No 4115

Février 2001

THÈME 2

ISSN 0249-6399 ISRN INRIA/RR--4115--FR+ENG

RHÔNE-ALPES

Additive symmetric: the non-negative case

Marc Daumas, CNRS
Philippe Langlois, INRIA

Thème 2 — Génie logiciel
et calcul symbolique
Projet Arénaire

Rapport de recherche n ° 4115 — Février 2001 — 13 pages

Abstract: An additive symmetric 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 exact correspondence of an additive symmetric when b and c have the same sign.

Key-words: Floating point arithmetic, additive symmetric, correction, IEEE-754 standard.

(Résumé : tsvp)

This text is also available as a research report of the Laboratoire de l'Informatique du Parallélisme `http://www.ens-lyon.fr/LIP`.

Unité de recherche INRIA Rhône-Alpes
655, avenue de l'Europe, 38330 MONTBONNOT ST MARTIN (France)
Téléphone : 04 76 61 52 00 - International: +33 4 76 61 52 00
Télécopie : 04 76 61 52 52 - International: +33 4 76 61 52 52

Symétrique additif: le cas positif

Résumé : Un symétrique additif b de a par rapport à c vérifie $c = (a + b)/2$. L'existence et l'unicité d'un tel b est une propriété de base en arithmétique exacte qui disparaît quand a et b sont des nombres à virgule flottante et quand le calcul de c est effectué dans une arithmétique de type IEEE-754. Nous présentons et nous prouvons des conditions sur l'existence, l'unicité et l'égalité avec le symmétrique exact dans le cas où b et c sont de même signe.

Mots-clé : Arithmétique à virgule flottante, symétrique additif, correction, norme IEEE-754.

1 Introduction

"Floating point arithmetic is by nature inexact". This quotation from Knuth [1] 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 $\mathbb{F}$, 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 $\mathbb{F}$ [2]. 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 [3] and [4] and illustrate well how subtle the discrepancies are.

Here we consider the existence and the uniqueness of an *additive symmetric* in the floating point number set $\mathbb{F}$. An additive symmetric a of b with respect to c, satisfies

$$c = fl\left(\frac{a+b}{2}\right), \tag{1}$$

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,

$$a_e = 2c - b, \tag{2}$$

is the unique additive symmetric 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 a exist within $\mathbb{F}$?

Uniqueness. Is the additive symmetric unique?

Consistency. Does $a = \widehat{a}_e$, where $\widehat{a}_e = fl(a_e)$?

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

To derive the answers to these questions, we interpret additive symmetry as a correcting operator. Let $\widehat{x}$ be a floating point number in $\mathbb{F}$. Correcting $\widehat{x}$ with the correcting term $z \in \mathbb{F}$ means computing

$$\overline{x}(z) = fl(\widehat{x} + z). \tag{3}$$

Let y be a correcting term that yields $x \in \mathbb{F}$ from $\widehat{x}$, that is $\overline{x}(y) = x$. This correcting term y is an additive symmetric of $\widehat{x}$ with respect to $x/2$ when $x/2 \in \mathbb{F}$. Again, where no rounding affects neither the computation of the correcting term, nor the correction in relation (3),

$$y_e = x - \widehat{x}, \tag{4}$$

is the unique *exact* correcting term. Existence, uniqueness and consistency of the additive symmetric 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 $y = \widehat{y}_e$, where $\widehat{y}_e = fl(y_e)$?

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 symmetric in Section 3 and devote following Section 4 to the proofs — the main result of the paper is the summary Figure 1. We conclude describing questions that remain open.

• **Notations.** We use the classic following notations defined for $\mathbb{F}$. We denote $u(x) = \text{ulp}(x)$, one unit in the last place of the floating point $x \in \mathbb{F}$, and $\mathbf{u} = u(1)$, one ulp of one. Let $x = (-1)^s 1.f \times 2^e$, with a p bits fraction-nary part f. We have $u(x) = \max\{2^{e-p}, \lambda_d\}$, and $u(0) = \lambda_d$, where λ_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 $\mathbb{F}$. The IEEE-754 standard defines $p = 23$ for single precision and $p = 52$ for double precision. Since $\mathbb{F}$ 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 $x = 2^k$, $x^- = x - u(x)$ 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 symmetric. 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 and uniqueness of a corresponding additive symmetric. We denote $AS[b, c]$ an additive symmetric of b with respect to c. We use the correspondence between the correcting term y that yields x from $\widehat{x}$, and the additive symmetric of $b = \widehat{x}$ with respect to $c = x/2$ when $x/2 \in \mathbb{F}$. We have $y = AS(\widehat{x}, x/2)$.

Example 1. The additive symmetric $AS[2^+, 1/2]$ exists and is unique. We consider the corresponding correction of $\widehat{x} = 2^+$ that returns $x = 1$. The exact correcting term $y_e = -(1+u(2))$ belongs to $\mathbb{F}$, so $\widehat{y}_e = y_e$. The correcting term $\widehat{y}_e$ verifies

$$\widehat{x} + \widehat{y}_e = 1,$$

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

$$\widehat{x} + \widehat{y}_e^+ = 1 + \mathbf{u} = x^+,$$

and

$$\widehat{x} + \widehat{y}_e^- = 1 - \mathbf{u} = x^{--},$$

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

$$\overline{x}(\widehat{y}_e^-) = x^{--}, \quad \overline{x}(\widehat{y}_e) = x, \quad \text{and} \quad \overline{x}(\widehat{y}_e^+) = x^+.$$

The monotonicity of the rounding map fl ensures that $\widehat{y}_e$ is the unique correcting term that returns x. So it is for the additive symmetric of b with respect to c for the considered value $(2^+, 1/2)$.

We remark that $\overline{x}(\widehat{y}_e^-)$ and $\overline{x}(\widehat{y}_e)$ are not consecutive and $\overline{x}(\widehat{y}_e^-) < 1^- < \overline{x}(\widehat{y}_e)$. We verify that no correcting term exists for $x = 1^-$ and $\widehat{x} = 2^+$, or similarly, no symmetric additive of $b = 2^+$ with respect to $c = 1/2 - \mathbf{u}/4$. We detail a similar case of non-existence in the following example.

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

$$\widehat{x} + \widehat{y}_e = 1 < x.$$

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

$$\widehat{x} + \widehat{y}_e^+ = 1 + u(2) = 1^{++},$$

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

$$\overline{x}(\widehat{y}_e) = x^-, \quad \text{and} \quad \overline{x}(\widehat{y}_e^+) = x^+.$$

These values enclose the target value x but neither are equal to x. Therefore it does not exist an additive symmetric of $b = 5$ with respect to $c = (1/2)^+$.

Example 3. The additive symmetric $AS[1,1]$ is non-unique. This case is less surprising rephrased as follows: the three correcting terms $\widehat{y}_e$, $\widehat{y}_e^-$ and $\widehat{y}_e^+$ return $x = 2$ from $\widehat{x} = 1$. For these values, the exact and rounded correcting terms verify $y_e = \widehat{y}_e = 1$. We have

$$\widehat{x} + \widehat{y}_e = 2,$$

$$\widehat{x} + \widehat{y}_e^+ = 2 + \mathbf{u},$$

and

$$\widehat{x} + \widehat{y}_e^- = 2 - \mathbf{u}/2.$$

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

$$\overline{x}(\widehat{y}_e) = \overline{x}(\widehat{y}_e^-) = \overline{x}(\widehat{y}_e^+) = x.$$

2.2 Motivations

The original motivation to the theoretical study of the additive symmetric comes from the CENA method introduced by one of the authors in [5], [6].

• **The CENA method.** The CENA method provides an automatic correction of the first-order effect of floating point rounding errors on the result of numerical algorithms. This correction is applied to the final result of a computation but correcting sensitive intermediate variables is sometimes interesting. 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.

The principles of the CENA method are the following. Given a computed $\widehat{x}$, the method yields a corrected $\overline{x}$ defined as

$$\overline{x} = fl\left(\widehat{x} + \widehat{\Delta}_L\right). \tag{5}$$

The correcting term $\widehat{\Delta}_L$ is the computed linearization of the error in $\widehat{x}$ with respect to the rounding errors introduced in the intermediate computations. For example, let $\widehat{f}$ be the floating point evaluation of a program that computes $\widehat{x} = \widehat{x}_N$ at the datum $X = (x_1, \ldots, x_n)$. The computation of the intermediate variables $\widehat{x}_{n+1}, \ldots, \widehat{x}_N$, generates the vector of the absolute elementary rounding errors $\delta = (\delta_{n+1}, \ldots, \delta_N)$. So, the first order approximate is $\Delta_L = \sum_{k=n+1}^{N} \frac{\partial \widehat{f}}{\partial \delta_k}(X, \delta) \cdot \delta_k$. The computation of Δ_L uses algorithmic differentiation and rounding error estimates to return the correcting term $\widehat{\Delta}_L$. The CENA method also provides a confidence threshold associated to the corrected result $\overline{x}$. These last more technical aspects are not necessary hereafter and are described in [6].

Let $x = fl(f)$ be the most accurate answer that can be returned by the program. Relation (5) returns this optimal x when $\widehat{\Delta}_L$ is an additive symmetric of the computed result $\widehat{x}$ with respect to $x/2$. The CENA method returns a number close to this x provided that the error of the evaluation $\widehat{f}$ contains mostly first-order effect of the elementary rounding errors. In this paper, we exhibit cases of such functions where the optimal correction is not possible since the computed result $\widehat{x}$ does not have a symmetric with respect to $x/2$.

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

• **New value = Old value + Correction.** Numerical methods often consist in such update strategy, as remarks for example HIGHAM in [7]. 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 symmetric. 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 $(-a) \in \mathbb{F}$ satisfies $fl(a + (-a)) = 0$, for $a \in \mathbb{F}$. In [3], 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 $fl^{-1}(0) \subseteq]-\epsilon, \epsilon[$, where $\epsilon = \min_{x,y \in \mathbb{F}, x \neq y} |x-y|$, 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 and even without denormalized numbers.

• **The multiplicative reciprocal.** The multiplicative reciprocal $(1/b)$ satisfies $fl(b \times (1/b)) = 1$, for $b \in \mathbb{F}^*$. 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 [4]. A unique multiplicative reciprocal exists for the "round towards $+\infty$" rounding mode. The other modes yield at most 2 reciprocals for $b \in \mathbb{F}^*$: 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 [8].

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 [9]. After having proved that $x \times fl(1/x) \in \{1 - \mathbf{u}/2, 1\}$, he exhibits the smallest double precision number $1 < x < 2$ such that $fl(x \times fl(1/x)) \neq 1$. Such a result illustrates that the computations of $fl(x \times fl(1/y))$ and $fl(x/y)$ are not equivalent and remind us a famous chip flaw [10], [11].

3 Properties of the non-negative case of the additive symmetric

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

$$x = fl(\widehat{x} + y). \tag{6}$$

We prove conditions on the target value x, depending on a given $\widehat{x}$, such that a correcting term y that verifies relation (6) exists and is unique.

Let x be of the same sign than $\widehat{x}$, for example positive. We distinguish six regions $\mathbf{R}_1, \ldots, \mathbf{R}_6$, that depend on $\widehat{x}$ in $[0, \Lambda]$, the positive part of $\mathbb{F}$ where Λ denotes the largest positive floating point number. The existence and the uniqueness of a correcting term y verifying relation (6) vary with respect to the region where x belongs. We summarize these properties with Figure 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 $x = 2^k$ 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

$$A(x) = \min_{a \in \mathbb{F}}\{fl(a + |x|) = a\}.$$

We verify that $A(x) = 2^{e+p+2}$ for $x = (-1)^s 1.f \times 2^e$ (f has p bits), and $A(0) = 0$.

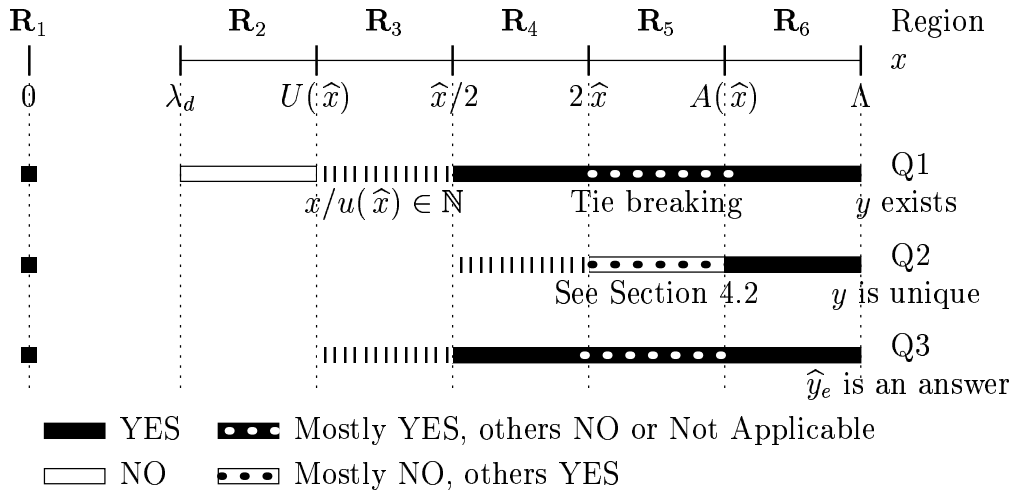

Figure 1: Summary of results on questions Q1-3 of the correction problem. Answers to the original problem are obtained by scaling the figure by a factor $\frac{1}{2}$, $\widehat{x}$ become b and $\frac{x}{2}$ becomes c.

4 Proofs of the properties

We recall the important result from Sterbenz [2] that gives a sufficient condition for an "exact subtraction" in $\mathbb{F}$ (hypothesis on $\mathbb{F}$ arithmetic are defined in the Introduction).

Lemma 1 (Sterbenz) *Let a and b be in $\mathbb{F}$ with $a \leq b \leq 2a$. Provided $b - a$ does not overflow, the floating point subtraction $fl(b-a)$ introduces no rounding error,* i.e.,

$$fl(b-a) = b - a.$$

This property enables us to derive direct proofs of the existence, and in some cases of the uniqueness, of the correcting term y expected in relation (6). We gather this kind of derivation is next Section 4.1. The other proofs need to explicitly discuss a lot of possible correcting terms that depend on the actual values of $\widehat{x}$ and x. We present as simple as possible proofs of these cases in Section 4.2.

4.1 Results obtained using $\mathbb{F}$ properties

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

($\mathbf{R_1}$) $\mathbf{x = 0}$.

This is the additive inverse case which is analyzed in a more general point of view in [3] as previously mentioned. For completeness, we propose the following proof in the current context.

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

($\mathbf{R_2}$) $\lambda_{\mathbf{d}} \leq \mathbf{x} < \mathbf{U}(\widehat{\mathbf{x}})$.

This condition implies that $\widehat{x} \neq 0$ and $y_e = x - \widehat{x} < 0$. We only look for $y < 0$, since $y \geq 0$ gives $fl(\widehat{x} + y) \geq \widehat{x} \geq U(\widehat{x}) > x$. We distinguish three cases to prove that $fl(\widehat{x} + y) \neq x$.

1. When $-y > 2\widehat{x}$, $\widehat{x} + y < -\widehat{x}$ and $fl(\widehat{x} + y) < -\widehat{x} < 0 < x$.

2. When $-y < \widehat{x}/2$, we have $\widehat{x} \geq 2\lambda_d$ for $y \neq 0$. It follows that $\widehat{x}/2 \geq U(\widehat{x})$ and as $\widehat{x} + y > \widehat{x}/2$, we prove that $fl(\widehat{x} + y) \geq fl(\widehat{x}/2) \geq U(\widehat{x}) > x$.
3. When $\widehat{x}/2 \leq -y \leq 2\,\widehat{x}$, the "exact subtraction" result from Sterbenz applies. This means here $fl(\widehat{x} + y) = \widehat{x} + y$. The number $x - \widehat{x} \notin \mathbb{F}$, since it is closer to $\widehat{x}$ than $\widehat{x} \pm U(\widehat{x})$, its closest neighbor in $\mathbb{F}$. Therefore no floating point number y satisfies relation (6).

No correcting term y satisfies relation (6) for x in this region $\mathbf{R}_2$.

($\mathbf{R}_3$) $\mathbf{U(\widehat{x}) \leq x < \widehat{x}/2}$.

We prove the existence for x in this region such that $x/u(\widehat{x})$ is an integer. We define the positive integers $\widehat{m} = \widehat{x}/u(\widehat{x})$ and $m = x/u(\widehat{x})$. For x in this region, $u(x) < u(\widehat{x})$ and $\widehat{x} - x = (\widehat{m} - m)u(x)$ verifies $|\widehat{m} - m| \leq \widehat{m} < \mathbf{u}^{-1}$. Therefore $x - \widehat{x} \in \mathbb{F}$ and $\widehat{y}_e = fl(x - \widehat{x}) = x - \widehat{x}$ verifies relation (6).

This proof extends Sterbenz equality for the subtraction $\widehat{x} - x$ to $x < \widehat{x}/2$ provided that $x/u(\widehat{x})$ is an integer. Nothing direct seems to be derivable in region $\mathbf{R}_3$ about the uniqueness, nor when $x/u(\widehat{x})$ is not an integer. We consider these remaining aspects in Section 4.3.

($\mathbf{R}_4$) $\mathbf{\widehat{x}/2 \leq x \leq 2\,\widehat{x}}$.

From Sterbenz equality, we have $\widehat{y}_e = fl(x - \widehat{x}) = x - \widehat{x}$ and $fl(\widehat{x} + \widehat{y}_e) = fl(x) = x$. Existence and $\widehat{y}_e = y_e$ is proved in region $\mathbf{R}_4$.

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

($\mathbf{R}_5$) $\mathbf{2\,\widehat{x} < x \leq A(\widehat{x})}$.

A similar discussion as in region $\mathbf{R}_3$ yields analogous partial results on existence and exact correction. We chose to present the complete discussion of this region in next Section 4.2.

($\mathbf{R}_6$) $\mathbf{A(\widehat{x}) < x \leq \Lambda}$.

We recall $x > A(\widehat{x})$ means $fl(x + \widehat{x}) = x$. In these regions, relation (6) is satisfied for $y = x = \widehat{y}_e$. We prove uniqueness while discussing region $\mathbf{R}_5$ 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 $\widehat{x}$ and x, we prove conditions on existence and uniqueness for y satisfying relation (6) in the regions $\mathbf{R}_5$ and $\mathbf{R}_6$, that is for

$$2\,\widehat{x} < x \leq \Lambda. \tag{7}$$

• **Principles of the proof.** We compute

$$fl(\widehat{x} + y) \text{ for } y \in \{\alpha, \beta, \gamma, \delta\}. \tag{8}$$

The four numbers $\alpha < \beta < \gamma < \delta$ are such that

$$[\alpha, \delta] \cap \mathbb{F} \subset \{\alpha, \beta, \gamma, \delta\}, \quad \beta \in \mathbb{F}, \quad \gamma \in \mathbb{F} \quad \text{and} \quad y_e \in [\beta, \gamma]. \tag{9}$$

We note that β and γ are consecutive floating point numbers and $\beta = \widehat{y}_e$ or $\gamma = \widehat{y}_e$. In most cases, α and δ are such that $\alpha = \beta^-$ and $\delta = \gamma^+$.

• **Notations for the proof.** Let $\widehat{x}$ and x be positive integers that verify previous condition (7). We define the integer

$$m = x/u(x), \tag{10}$$

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

$$\widehat{x} = \left(k + \frac{g}{2} + \frac{r+s}{4}\right) u(x). \tag{11}$$

Normalized notation for x provides $\mathbf{u}^{-1} \leq m < 2\mathbf{u}^{-1}$. The condition $2\,\widehat{x} < x$ yields $0 \leq k < m/2$, so $m - k > m/2 \geq m/2 + 1 \geq \mathbf{u}^{-1}/2 + 1$. Equality holds only for $m = \mathbf{u}^{-1}$.

>From relations (10) and (11), $y_e = x - \widehat{x}$ verifies

$$y_e = \left(m - k - 1 + \frac{1-g}{2} + \frac{(1-r)+(1-s)}{4}\right) u(x). \tag{12}$$

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

$$z = (m - k - 1)u(x), \tag{13}$$

that will vary around y_e. Since $\mathbf{u}^{-1}/2 \leq m - k - 1 < 2\mathbf{u}^{-1}$, z is a floating point number not necessarily normalized. We discuss the computation of relation (8) for varying $u(z)$ with the following tables. From relation (13), we have $u(x)/2 \leq u(z) \leq u(x)$. The different cases for the ulp of z and its neighbors are $u(x)/4$, $u(x)/2$ and $u(x)$; these values guide the following discussion. We consider the cases $u(z) = u(x)/2$ and $u(z) = u(x)$ — values for the neighbors appear implicitly in the discussion.

The tables also present $\widehat{y}_e$, the rounded value of relation (12). Comparing $\widehat{y}_e$ and y when $fl\,(\widehat{x} + y) = x$ yields the answer to Q3. We indicate some cases where $\widehat{y}_e = \beta$ or γ is sufficient to derive a (positive or negative) answer to Q3.

• **Discussion.**

1. $\mathbf{u(z) = u(x)}$.

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

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

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

   Now we explore the table with respect to (r', s') that governs the rounding in $fl\,(\widehat{x} + y)$.

Rounding conditions			$r' = 0$	$r' = 1$	
				$s' = 0$	$s' \neq 0$
		$\widehat{y}_e$	$m - k$	β or γ	$m - k - 1$
y	$y/u(x)$	$(\widehat{x} + y)/u(x)$	$fl\,(\widehat{x} + y)\,/u(x)$		
δ	$m - k + 1$	$m + \frac{r'+s'}{2} + 1$	$m + 1$	$\geq m + 1$	$\geq m + 1$
γ	$m - k$	$m + \frac{r'+s'}{2}$	m	$even(m, m+1)$	$m + 1$
β	$m - k - 1$	$m + \frac{r'+s'}{2} - 1$	$m - 1$	$even(m-1, m)$	m
α	$m - k - \frac{3}{2}$	$m + \frac{r'+s'}{2} - \frac{3}{2}$	$\leq m - 1$	$\leq m - 1$	$m - 1$

   This explicit computation gives the following answers to questions Q1-3.

   - Q1 is positive except when $(r', s') = (1, 0)$ for odd m,
   - Q2 is positive except when $(r', s') = (1, 0)$,

- Q3 is positive with Q1, and
- two correcting terms exist when $(r', s') = (1, 0)$ for even m.

2. $\mathbf{u(z) = u(x)/2}$.

In this case, $\mathbf{u}^{-1}/2 \leq m-k-1 < \mathbf{u}^{-1}$. The mantissa of z is shifted to compute relation (8) and the guard bit g introduces the quantity $(1-g)/2$ we consider to round y_e. We use $\beta = (m-k-1+\frac{1-g}{2})u(x)$. We note that $u(\beta) = u(z) = u(x)/2$. This gives us relations to define α and γ as the respective neighbors of δ and β satisfying relation (9). We separate two cases with respect to $m = \mathbf{u}^{-1}$ and we build the table with respect to (g, r, s) since (r, s) governs the rounding in $fl\,(\widehat{x} + y)$.

(a) When $m \neq \mathbf{u}^{-1}$.

Rounding conditions			$r=0$		$r=1$
			$s=0$	$s \neq 0$	
		$\widehat{y}_e$	$m-k+\frac{1-g}{2}-\frac{1}{2}$		β or γ
y	$y/u(x)$	$(\widehat{x}+y)/u(x)$	$fl\,(\widehat{x}+y)\,/u(x)$		
δ	$m-k+\frac{1-g}{2}$	$m+\frac{r+s}{4}+\frac{1}{2}$	$even(m, m+1)$	$m+1$	$m+1$
γ	$m-k+\frac{1-g}{2}-\frac{1}{2}$	$m+\frac{r+s}{4}$	m	m	m
β	$m-k+\frac{1-g}{2}-1$	$m+\frac{r+s}{4}-\frac{1}{2}$	$even(m-1, m)$	m	m
α	$m-k+\frac{1-g}{2}-\frac{3}{2}$	$m+\frac{r+s}{4}-1$	$m-1$	$m-1$	$m-1$

We note that δ does not necessary belong to $\mathbb{F}$. The conclusions are now the followings.

- Q1 and Q3 are positive in all the cases,
- Q2 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) This last table gives the answer when $m = \mathbf{u}^{-1}$.

Rounding conditions			$r=0$		$r=1$
			$s=0$	$s \neq 0$	
		$\widehat{y}_e$	$m-k+\frac{1-g}{2}-\frac{1}{2}$		β or γ
y	$y/u(x)$	$(\widehat{x}+y)/u(x)$	$fl\,(\widehat{x}+y)\,/u(x)$		
δ	$m-k+\frac{1-g}{2}$	$m+\frac{r+s}{4}+\frac{1}{2}$	m (1)	$m+1$	$m+1$
γ	$m-k+\frac{1-g}{2}-\frac{1}{2}$	$m+\frac{r+s}{4}$	m	m	m
β	$m-k+\frac{1-g}{2}-1$	$m+\frac{r+s}{4}-\frac{1}{2}$	$m-\frac{1}{2}$	$m-\frac{1}{2}$	m
α	$m-k+\frac{1-g}{2}-\frac{5}{4}$	$m+\frac{r+s}{4}-1$	$\leq m-\frac{1}{2}$	$\leq m-\frac{1}{2}$	$< m$ (2)

We derive the following conclusions.

- Q1 and Q3 are positive in all the cases,
- Q2 is positive when $(r, s) = (0, s)$ with $s \neq 0$,

[1] The value of the cell is obtained from even rounding since $m = even(m, m+1)$. It means that δ is the last quantity that may return m. When $g = 0$ and $m-k = \mathbf{u}^{-1}$, δ is not a floating point number and we may use $\delta + u(x)/2$ to obtain a correction to $(m+1)u(x)$. In this case, the solution $\widehat{y}_e$ is unique.

[2] Here again, the value of the cell is obtained by even rounding $m = even(m, m+1)$. We present the safest bound for α, that only occurs for $m-k+\frac{1-g}{2}-1 = \mathbf{u}^{-1}/2$. This conditions yields that $k - \frac{g}{2} + \frac{1}{2} = \mathbf{u}^{-1}/2$ and finally $k = \mathbf{u}^{-1}/2$, $g = 1$ and $r = 1$. As it is impossible, we conclude that the result is strictly less than m.

- two correcting terms exist when $r = 1$ or eventually when $(r, s) = (0, 0)$.

• **Conclusion for the regions $\mathbf{R}_5$-$\mathbf{R}_6$.** Now we derive the answers for questions Q1-3 in the regions $\mathbf{R}_5$-$\mathbf{R}_6$ from the results of the previous discussion.

We prove that Q1, Q2 and Q3 are positive in $\mathbf{R}_6$. In this region, $A(\widehat{x}) < x$ gives $k = g = r' = 0$. When $u(z) = u(x)$, the column $r' = 0$ in the first table yields the result.

When $u(z) = u(x)/2$, the condition $\mathbf{u}^{-1}/2 \leq m - k - 1 < \mathbf{u}^{-1}$ yields $m - 1 < \mathbf{u}^{-1}$ for $k = 0$. As $\mathbf{u}^{-1} \leq m < 2\mathbf{u}^{-1}$, the case $u(z) = u(x)/2$ is only possible when $m = \mathbf{u}^{-1}$. This case corresponds to the last table. The discussion of this case when $g = r = 0$ gives the expected positive answers. The case $r = 1$ would imply that $x = A(x)$ that is not permitted by the definition of $\mathbf{R}_6$.

We prove that Q1 and Q3 are positive in $\mathbf{R}_5$ 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 $\widehat{x}$ and x 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 $\mathbf{R}_2$ when $x/u(\widehat{x}) \in \mathbb{Z}$, and Q1-3 otherwise,
- Q2 in region $\mathbf{R}_3$.

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 Figure 1 exhibiting examples of the different cases.

Region $\mathbf{R}_3$. Example 1 in Section 2.1 illustrates the positive answers to questions Q1-3 in the region $\mathbf{R}_3$. We verify that $x = 1$ and $\widehat{x} = 2^+$ satisfy $x/u(\widehat{x}) \in \mathbb{F}$.

Region $\mathbf{R}_3$. Example 2 in Section 2.1 illustrates the negative answer to question Q1 in the same region $\mathbf{R}_3$. Here $x = 1^+$ and $\widehat{x} = 5$ are not such that $x/u(\widehat{x}) \in \mathbb{F}$.

Region $\mathbf{R}_4$. Example 3 in Section 2.1 illustrates the negative answer to question Q2 in the region $\mathbf{R}_4$.

Region $\mathbf{R}_4$. 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 $\mathbf{R}_4$.

Example 4. We chose $\widehat{x} = 1^+ = 1 + \mathbf{u}$ and $\widehat{x} = \frac{1}{2}^+ = \frac{1}{2} + \frac{1}{2}\mathbf{u}$. The exact correcting term is

$$y_e = -\left(\frac{1}{2} + \frac{1}{2}\mathbf{u}\right) = \widehat{y}_e.$$

The computation of the correction $fl\,(\widehat{x} + y)$ for the two neighbors of $\widehat{y}_e$ yields

$$fl\,(\widehat{x} + \widehat{y}_e^-) = x^- \quad \text{and} \quad fl\,(\widehat{x} + \widehat{y}_e^+) = x^+.$$

The unique correcting term is $\widehat{y}_e$.

5 Conclusion

Additive symmetry is a basic operator in exact arithmetic with well known properties providing numerous applications. We have considered this operator in floating point arithmetic and discuss the elementary properties of existence, uniqueness and consistency of the additive symmetric in the non-negative case. 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 $\mathbf{R}_2$) 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 circumvent with 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 $\mathbf{R}_4$ validates the classic update strategy since the correcting factor is designed to tend to zero. Hence the difficulty remains the choice of a good initial value for these iterative processes that ensures such convergence.

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 in simple algebraic derivation. Alas, the most cases need long and tedious computation to cover all the possible cases. 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 [12, 13, 14] and [15]. Of course, validated results are the general theorems that promote future direct proofs.

When no existence is proved, the natural following question is to explicit the best approximate additive symmetric — or similarly the best correcting term. When the existence is proved, the exact additive symmetric — or correcting term — is (one of) the solution(s) of the problem. Let us consider now for example the following slight generalization of our correction problem 3. Let $\widehat{x}$ be a floating point number and x 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 $x = 8 + u(4) + u(2) - u(1) - \xi$, with $\xi > 0$ small enough, we have

$$fl\,(x) = 8^+.$$

Choosing $\widehat{x} = 1 - u(1)$, we verify that $\widehat{y}_e = 7^+ = 7 + u(4)$ and

$$fl\,(\widehat{x} + \widehat{y}_e) = 8 = fl\,(x)^-,$$

whereas

$$fl\,(\widehat{x} + \widehat{y}_e^+) = 8^+ = fl\,(x).$$

The non exact $\widehat{y}_e^+$ yields the expected corrected value. "Floating point arithmetic is by nature inexact".

References

[1] D. E. Knuth, The Art of Computer Programming, Volume 2, Seminumerical Algorithms, 3rd Edition, Addison-Wesley, Reading, MA, USA, 1998.

[2] P. H. Sterbenz, Floating-Point Computation, Prentice-Hall, Englewood Cliffs, NJ, USA, 1974.

[3] U. W. Kulisch, Rounding near zero, in: J.-C. Bajard, et al. (Eds.), Proceedings of RNC-4, Real Numbers and Computer Conference, Dagsthul, 2000, pp. 23–29.

[4] J.-M. Muller, Some algebraic properties of floating-point arithmetic, in: J.-C. Bajard, et al. (Eds.), Proceedings of RNC-4, Real Numbers and Computer Conference, Dagsthul, 2000, pp. 21–38.

[5] P. Langlois, F. Nativel, Automatic reduction of round-off errors in floating point arithmetic, in: J.-C. Bajard, other (Eds.), Proceedings of Second Real Numbers and Computer Conference, 1996, pp. 199–214, (in french).

[6] P. Langlois, Automatic linear correction of rounding errors, BIT 41 (3), (To appear, also available as INRIA Research Report RR-4025, October 2000).

[7] N. J. Higham, Accuracy and Stability of Numerical Algorithms, Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 1996.

[8] G. Hanrot, J. Rivat, G. Tenenbaum, P. Zimmermann, Some properties of floating point numbers, submitted (Jan. 2001).

[9] A. Edelman, When is $x * (1/x) \neq 1$?, manuscript available at URL = `http://www-math.mit.edu/~edelman` (Dec. 1994).

[10] J. Markoff, Circuit flaw causes Pentium chip to miscalculate, Intel admits, New York Times 24 November.

[11] C. B. Moler, A tale of two numbers, SIAM News 28 (1995) 1,16, also in MATLAB News and Notes, Winter 1995, 10–12.

[12] J. Harrison, Verifying the accuracy of polynomial approximations in HOL, in: Proceedings of the 10th International Conference on Theorem Proving in Higher Order Logics, Murray Hill, New Jersey, 1997, pp. 137–152.

[13] J. S. Moore, T. Lynch, M. Kaufmann, A mechanically checked proof of the AMD5K86 floating point division, IEEE Transactions on Computers 47 (9) (1998) 913–926.

[14] J. Harrison, A machine-checked theory of floating point arithmetic, in: Y. Bertot, G. Dowek, A. Hirschowitz, C. Paulin, L. Théry (Eds.), 12th International Conference on Theorem Proving in Higher Order Logics, Nice, France, 1999, pp. 113–130.

[15] M. Daumas, C. Moreau-Finot, L. Thery, Computer validated proofs of a toolset for adaptable arithmetic, Research report 4095, Institut National de Recherche en Informatique et en Automatique, Le Chesnay, France (2001).

Unit´e de recherche INRIA Lorraine, Technopôle de Nancy-Brabois, Campus scientifique,
615 rue du Jardin Botanique, BP 101, 54600 VILLERS LÈS NANCY
Unit´e de recherche INRIA Rennes, Irisa, Campus universitaire de Beaulieu, 35042 RENNES Cedex
Unit´e de recherche INRIA Rhône-Alpes, 655, avenue de l'Europe, 38330 MONTBONNOT ST MARTIN
Unit´e de recherche INRIA Rocquencourt, Domaine de Voluceau, Rocquencourt, BP 105, 78153 LE CHESNAY Cedex
Unit´e de recherche INRIA Sophia-Antipolis, 2004 route des Lucioles, BP 93, 06902 SOPHIA-ANTIPOLIS Cedex

Éditeur
INRIA, Domaine de Voluceau, Rocquencourt, BP 105, 78153 LE CHESNAY Cedex (France)
http://www.inria.fr
ISSN 0249-6399